Contributions à la modélisation du sens par approches formelles, linguistiques et statistiques

Jeanne Villaneau

Université de Bretagne-Sud

Habilitation à diriger des recherches (HDR)

Contributions à la modélisation du sens par approches formelles, linguistiques et statistiques

Jeanne Villaneau

Membres du jury :

Ioannis Kanellos (examinateur, président de jury)

Jean-Yves Antoine (rapporteur)

Olivier Ridoux (rapporteur)

Sophie Rosset (rapporteur)

Pierre-François Marteau (examinateur)

9 mars 2016

IRISA Expression

Université de Bretagne Sud

“”

Sommaire

Passe encore d'enseigner
Mais chercher à cet âge...

(d'après La Fontaine)

Introduction

L'habilitation à diriger des recherches est l'occasion de faire un bilan et de chercher le fil directeur qui relie des activités de recherche qui peuvent apparaître éparses et multiformes.

Dans les activités professionnelles qui furent les miennes, la recherche est arrivée très tard. J'ai enseigné les mathématiques en lycée pendant 17 années avec un plaisir certain et, alors même que la routine risquait de s'installer, la création de l'UBS m'a donné l'occasion de partir vers de nouvelles expériences. Ici, point de routine : la vie de professeur agrégé dans une université qui allait se créer puis, qui venait de l'être, n'a été ni ennuyeuse, ni de tout repos. Le travail de doctorat en informatique, mené en parallèle avec le service d'enseignement d'un prag joint à diverses directions d'étude, a été un chemin long et difficile : merci encore à ceux qui m'ont très patiemment aidée et soutenue ; ainsi qu'à l'UBS qui a finalement adopté le principe de pouvoir accorder des décharges de service aux prag pour leur permettre de terminer un doctorat. Enfin, alors que je venais d'être nommée sur un poste de Maître de Conférences (merci aux collègues pour ce changement de statut qui, à bien y réfléchir, ne profitait guère à l'institution), la création de l'ENSIBS a été l'occasion de franchir un pas de plus dans les responsabilités administratives, avec la direction des études d'une école encore à monter...

Comme il était d'emblée exclu de faire de la recherche un objectif de carrière, j'ai surtout profité des occasions qui se présentaient pour travailler sur des sujets qui m'intéressaient. Si la rencontre avec le Traitement Automatique des Langues (TAL) est un peu dû au hasard (et à Jean-Yves Antoine), il se trouve que c'est un domaine qui correspond très bien à mes propres centres d'intérêt, du fait de sa pluridisciplinarité et de son caractère inépuisable. Le travail sur la langue permet de comprendre toute sa complexité ; il permet également de réaliser à quel point la communication est un processus complexe, basé sur des implicites et des non-dits et dans lequel intervient une multitude d'interprétations. Par ailleurs, la langue naturelle est elle-même multiple et en perpétuelle évolution, en même temps que les outils qui tentent de la traiter. Tout cela fait du TAL un sujet de recherche, certes difficile et parfois ingrat, où l'on ne maîtrise pas tout, voire pas grand

chose, mais en même temps passionnant. Enfin, les outils utilisés ont parfois des liens inattendus avec diverses branches des mathématiques, un point positif supplémentaire pour moi.

Le travail de doctorat (réalisé sous la direction conjointe d'Olivier Ridoux et de Jean-Yves Antoine que je ne saurais assez remercier) a été consacré au dialogue oral Homme-Machine (DOHM) et à la conception d'un système dit de « compréhension », le système Logus. En l'occurrence, cette recherche du sens consistait à transformer une liste de mots en une formule logique susceptible de représenter le sens de l'énoncé dans un contexte applicatif déterminé. Par la suite, le thème général des recherches s'est étendu en direction des corpus de textes et de la langue écrite d'une part et de la détection des opinions et des émotions d'autre part. En même temps, s'y est introduit l'utilisation d'outils statistiques ; ceux-ci se sont en effet imposés dans une très grande part des recherches en TAL.

Outre l'intérêt personnel que l'on peut porter à comprendre comment la langue naturelle participe à la transmission du sens, les travaux présentés ne sont pas purement spéculatifs. Certes, les résultats des recherches dans le domaine du DOHM ont pu décevoir ceux qui rêvaient de systèmes vocaux en lieu et place d'interlocuteurs humains. Mais, lorsqu'il s'applique aux textes, le problème du sens est également au cœur de la conception des systèmes susceptibles d'aider à extraire des informations pertinentes de la masse des documents actuellement disponibles sur le Web. Quant à la détection automatique de l'émotion ou de l'opinion, tout particulièrement sur les réseaux sociaux, elle intéresse fortement le monde économique, consommateurs comme fournisseurs, sans oublier bien sûr le monde politique.

Le chapitre 1 décrit des travaux qui ont prolongé le développement du système Logus conçu lors du doctorat. Le laboratoire Valoria qui regroupait l'équipe des informaticiens de l'UBS pendant ces années-là était en effet engagé dans le challenge Evalda-Media du programme Technolangue, destiné à évaluer les systèmes de compréhension en dialogue Homme-Machine.

Les travaux qui ont été développés à l'occasion du projet ANR en robotique Emotirob font l'objet des chapitres 2 et 3. L'objectif d'Emotirob était de concevoir un robot-compagnon en peluche, destiné à des enfants en hospitalisation longue. Ce robot devait être doté de la capacité à exprimer des émotions à l'aide de mouvements faciaux. Dans les travaux décrits dans le chapitre 2, la tâche consistait à détecter l'émotion perceptible chez l'enfant au travers de ses propos. Ce travail a donné lieu au développement d'un système de détection linguistique des émotions, Emologus et fait l'objet d'une thèse

de doctorat, sous la direction conjointe de Dominique Duhaut et de Jean-Yves Antoine, et que j'ai co-encadrée.

Le chapitre suivant (chapitre 3) est consacré à un projet très exploratoire. Son objectif était d'explorer comment on pouvait simuler les capacités cognitives d'un robot, pour le doter de réactions moins stéréotypées et dans le but d'une extension de ses capacités expressives. Il a également donné lieu à une thèse de doctorat, sous la direction de Dominique Duhaut, que Farida Saïd-Hocine et moi-même avons co-encadrée.

Les travaux décrits dans le chapitre 4 combinent outils statistiques et linguistique de corpus. En l'occurrence, le but était d'explorer des approches pour l'extraction automatique de relations sémantiques : patrons de verbes, grammaires de segments basées, entre autres, sur la ponctuation, étude particulière des parenthétiques. Il a fait l'objet d'une thèse de doctorat que j'ai (co)-encadrée.

Dans le chapitre 5, les méthodes statistiques dominent, pour une étude des similarités entre phrases avec le résumé multi-textes comme objectif final. Ces travaux ont fait l'objet d'une thèse de doctorat qui vient juste de se terminer, sous la direction de Pierre-François Marteau et co-encadrée par Farida Saïd-Hocine et moi-même.

Enfin, le chapitre 6 décrit des travaux toujours en cours, menés avec Jean-Yves Antoine et Anaïs Lefeuvre. Leur objectif est d'éclaircir les indications fournies par les accords-interannotateurs pour ce qui concerne la qualité des corpus annotés ; ces corpus représentent en effet une ressource essentielle du TAL.

Le chapitre final (page 94) est une brève conclusion avec quelques perspectives de recherche qui expliquent les raisons de cette HDR imprévue et hyper-tardive...

Chapitre 1

Logus et le challenge Evalda-Media

Les travaux de doctorat qui se sont terminés fin 2003 ont abouti à la réalisation de Logus, un système de compréhension de la langue orale spontanée dans le cadre d'un dialogue Homme-Machine. Logus a donné lieu à des travaux de développement pendant les années suivantes dont les plus importants sont décrits dans ce document.

– Ce chapitre est consacré à la participation de Logus au challenge du projet Technolangue Evalda-Media et à la compréhension en contexte de dialogue développée sur le corpus élaboré pour ce projet.
– La première partie du chapitre suivant traite du développement d'une extension de Logus implémentée dans le cadre du projet ANR Emotirob ; elle consistait à mesurer la valeur émotive contenue dans les propos tenus par un jeune enfant.

1.1 Le système de compréhension Logus et la compréhension en contexte de dialogue

1.1.1 Introduction

Dans un système de dialogue oral Homme-Machine (DOHM), le module de « compréhension » de la langue orale spontanée remplit une tâche essentielle. Comme il est indiqué dans le schéma très simplifié de la figure 1.1, le module de reconnaissance reçoit le signal vocal émis par l'utilisateur du système et il rend la liste ou le graphe

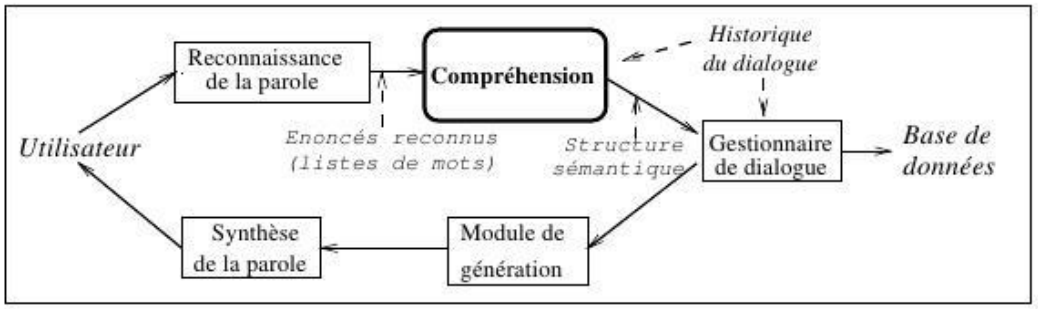


FIGURE 1.1: Architecture (simplifiée) d'un système de Dialogue Oral Homme-Machine.

de mots le plus probable correspondant. Le rôle du module de « compréhension » est de construire une structure sémantique qui puisse rendre compte du sens de ces mots en conciliant précision et simplicité ; le résultat rendu doit en effet garder toutes les informations utiles données par le locuteur tout en restant utilisable par le module de dialogue.

Les principales difficultés auxquelles se heurte l'analyse de la liste de mots reconnus sont de deux ordres : d'une part les spécificités de l'oral spontané, caractérisé par les hésitations et les reprises du locuteur qui cherche ses mots et d'autre part, les erreurs, souvent nombreuses, de la reconnaissance vocale. Ces problèmes rendent illusoire la possibilité d'une analyse syntaxique complète des énoncés traités et obligent à recourir à une connaissance du domaine et de la tâche.

Lorsque le système de DOHM est conçu pour une tâche très restreinte, horaires de train ou d'avion par exemple, l'interprétation du message peut se limiter à la détection d'une séquence de concepts, sur la base de structures sémantiques prédéfinies. Mais, lorsque le domaine d'application s'élargit, cette prédéfinition des requêtes devient plus complexe et la compréhension requiert d'autres approches [Van Noord et al. (1999)].

LOGUS est un système de compréhension de la parole spontanée dans le cadre d'un DOHM conçu pour des domaines restreints, mais néanmoins plus étendus que les domaines où opèrent la plupart des systèmes encore actuellement opérationnels et qui réalise les approches non-contextuelle et contextuelle dans un même module. L'approche logique, décrite dans la section suivante, utilise différents formalismes pour combiner des outils syntaxiques et sémantiques. La résolution des références s'appuie également sur une approche symbolique et logique et vient ainsi en complément de celle utilisée dans la conception générale du système.

Ce chapitre présente essentiellement les travaux d'implémentation de la compréhension en contexte de dialogue réalisés sur le système LOGUS à partir du corpus élaboré

pour le projet Technolangue MEDIA (Méthodologie d'Evaluation automatique de la compréhension hors et en contexte du DIAlogue). Après une brève exposition dans la section 1.1.2 des principes qui ont présidé à la conception du système LOGUS, la section 1.2.1 décrit le cadre du projet MEDIA et fait une brève analyse de son corpus, afin de dégager l'intérêt de son utilisation pour une telle expérimentation. Les principes de la compréhension en contexte et, plus précisément, de la résolution des références mises en œuvre dans LOGUS sont présentés dans la section 1.3.2. La section 1.3.3 présente une analyse quantitative et qualitative des résultats. Le chapitre se termine par la section 1.3.4 où sont présentées quelques conclusions que l'on tirer de cette expérience.

1.1.2 LOGUS : une utilisation de formalismes logiques pour la compréhension

LOGUS a été conçu pour fonctionner dans un domaine sensiblement plus large que ceux habituellement considérés pour ce type d'applications [1], où une représentation sémantique de l'énoncé par listes préconstruites d'attributs-valeurs s'avère suffisante. Néanmoins, l'analyse s'appuie sur une connaissance sémantique du domaine qui doit donc rester bien délimité et relativement étroit.

À partir d'une liste de mots issue d'un module de reconnaissance de la parole, LOGUS produit une formule logique qui représente le sens de l'énoncé. Le formalisme utilisé est adapté de la logique illocutoire de Vanderveken (2001) ; la formule logique s'obtient par application d'un acte de langage (sa *force propositionnelle*) à une structure construite à partir des « objets » de l'énoncé connus du système (son *contenu propositionnel*). La représentation sémantique peut également être mise sous la forme d'un graphe conceptuel à la Sowa (2001). La figure 1.2 donne un exemple de la structure sémantique obtenue à partir d'un énoncé du corpus MEDIA (cf. 1.2.1) et du graphe conceptuel correspondant.

L'analyse de l'énoncé est incrémentale et progressive ; elle se fait par étapes qui utilisent successivement différents formalismes logiques.

- Un lexique permet d'attacher à chaque mot « connu » une ou plusieurs « définitions ».
- Une première analyse partielle rattache les mots grammaticaux à leur tête lexicale. Cette étape peut être considérée comme un « *chunking minimaliste* ». Elle utilise des règles adaptées de celles des grammaires catégorielles de type AB et les termes simplement typés du λ-calcul [Villaneau et al. (2004); Villaneau and Antoine (2004)].
- Les étapes suivantes s'appuient sur une connaissance du domaine (ontologie) qui décrit le type des concepts du domaine d'application et les liens sémantiques qui peuvent les

1. Pour une description plus détaillée du système, on peut consulter les références suivantes : Villaneau et al. (2004); Villaneau (2003).

L'énoncé :

« je souhaiterais réserver dans un hôtel Mercure trois étoiles à Belfort pour les quatre derniers jours de juin »

La formule logique construite par LOGUS :

```
(demande (de    (reservation [
                             (date (num_mois (derniers (entier 4)) (nom "juin"))))])
                (hotel [     (marque_hotel (nom "Mercure")),
                             (etoiles (entier 3)),
                             (lieu (ville [(identification (nom "Belfort"))])))])
```

La sortie LOGUS figurée sous la forme d'un graphe conceptuel.

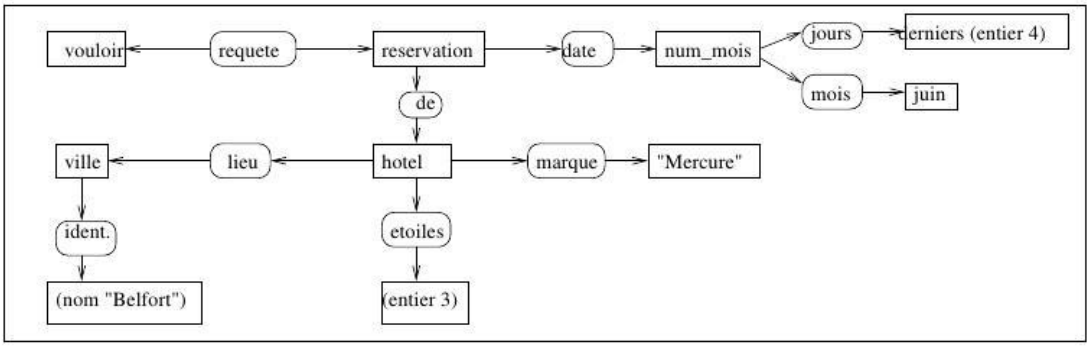


FIGURE 1.2: Un énoncé du corpus MEDIA et la sortie LOGUS correspondante

réunir. L'analyse de ces liens dans l'énoncé utilise les règles génériques d'une grammaire qui combinent les indices syntaxiques et sémantiques des différents composants. Ces règles sont appliquées en cascade, avec un assouplissement progressif des contraintes syntaxiques.

– La dernière étape prend en compte le contexte du dialogue pour compléter et préciser la compréhension de l'énoncé (cf. section 1.3.2). C'est essentiellement à l'expérimentation sur le corpus MEDIA de cette contextualisation qu'est consacré ce chapitre.

1.2 EVALDA-MEDIA : compréhension hors contexte de dialogue

1.2.1 Le projet EVALDA-MEDIA et son corpus

Financé par le Ministère français délégué à la Recherche et aux Nouvelles Technologies (MRNT), le projet EVALDA-MEDIA du programme Technolangue, avait pour objectif l'évaluation de différents systèmes de compréhension en dialogue Homme-Machine, hors et en contexte de dialogue. Les partenaires du projet avaient choisi d'enregistrer un corpus à partir d'un serveur de réservation hôtelière. Le corpus enregistré par ELDA (appellé corpus MEDIA par la suite) pour la campagne d'évaluation comporte 1250 dialogues : les 250 utilisateurs du système ont interrogé le système suivant différents scenarii de réservation d'hôtels, élaborés par les partenaires du projet. L'enregistrement s'est fait suivant le principe du Magicien d'Oz : les locuteurs ont dialogué avec un système simulé à leur insu par un opérateur humain. La figure 1.3 donne un extrait de dialogue du corpus. Dans cet exemple, *Ut* désigne l'utilisateur du système et *Co* le compère qui simule le système. Les expressions soulignées indiquent les marques linguistiques qui renvoient à une résolution des références en contexte de dialogue d'après les conventions MEDIA.

Le corpus ainsi enregistré a ensuite fait l'objet d'une transcription manuelle par ELDA, puis d'une annotation sémantique suivant les règles d'un manuel d'annotation mis au point par les partenaires du projet [2] : dans l'annotation sémantique hors-contexte, chaque énoncé est divisé en segments conceptuels « porteurs de sens ». À chacun d'entre eux est attribué un triplet *(mode, attribut, valeur)* ; des spécifieurs sont attachés aux attributs, afin de préciser les liens entre les différents concepts [3]. La figure 1.4 donne un énoncé extrait du corpus MEDIA avec son annotation sémantique.

Dans l'annotation sémantique en contexte de dialogue, les expressions référentielles portent les numéros des segments conceptuels auxquels elles renvoient.

2. La mesure d'accord entre annotateurs (kappa) se situe au-dessus de 80%.
3. Pour plus de détails, on peut consulter Devillers et al. (2004).

...	...
Ut	**"réserver une chambre simple du premier au six décembre dans trois endroits différents"**
Co	*"dans quelle ville souhaitez vous vous rendre"*
Ut	**"alors du un au trois décembre je veux être à Chalon Sur Saône du trois au cinq décembre à Bourg en Bresse et du cinq au six à Chaumont"**
Co	*"veuillez patienter nous recherchons cette information à Chalon Sur Saône du premier au trois décembre deux hôtels correspondent à votre demande le Saint Régis et le Ibis centre ville voulez vous des détails ou réserver dans l un de ces hôtels"*
Ut	**"hum des détails"**
Co	*"quelle information voulez vous"*
Ut	**"euh je voudrais le prix de la chambre"**
Co	*"'a l hôtel le Saint Régis la chambre individuelle est à cinquante euros à l hôtel Ibis centre ville la chambre individuelle est à cinquante euros souhaitez vous faire une réservation dans l un de ces hôtels"*
Ut	**"euh est ce que l un de ces hôtels accueille les animaux et est ce qu il y a un tennis"**
...	...
Co	*"... souhaitez vous réserver dans l un de ces hôtels"*
Ut	**"oui"**
Co	*"si oui lequel"*
Ut	**"euh le premier"**
...	...

FIGURE 1.3: Extrait de dialogue du corpus MEDIA

je souhaiterais réserver	+ :command-tache :reservation
dans un hôtel Mercure	+ :hotel-marque-reservation :mercure
trois étoiles	+ :hotel-etoile :3etoile
à Belfort	+ :localisation-ville-hotel :belfort
pour les quatre	+ :nombre-temps-reservation :4
derniers	+ :temps-axetps-reservation :dernier
jours	+ :temps-unite-reservation :jour
de juin	+ :temps-mois-reservation :6

FIGURE 1.4: L'énoncé de la figure 1.2 et son annotation MEDIA

1.2.2 LOGUS et l'Évaluation hors-contexte

Le système LOGUS a participé à la campagne d'évaluation hors-contexte [Bonneau-Maynard et al. (2006)]. Il y a obtenu des résultats honorables mais assez peu significatifs [4]. En effet, la principale difficulté rencontrée pour la participation de LOGUS à cette campagne n'a pas été l'adaptation du système à la tâche, mais bien la transformation de la formule logique obtenue en la suite ordonnée de triplets *(mode, attribut, valeur)* demandée pour l'évaluation. En effet, les sorties LOGUS sont globales : il y a « oubli » de l'ordre des mots et de la forme linguistique attachée à l'expression des requêtes. Comme on pouvait le craindre, l'annotation « collée au texte » de MEDIA s'est révélée souvent difficile voire impossible [5] à reconstituer à partir de la représentation sémantique finale : ainsi les deux tiers des erreurs relevées pour LOGUS ont été imputables à la transformation des sorties du système en la liste ordonnée des triplets demandée.

1.3 MEDIA : compréhension en contexte de dialogue

1.3.1 Les références dans MEDIA

L'un des principes retenus par les partenaires MEDIA était que seules les expressions se rapportant à des références hors énoncé devaient être prises en considération.

Étant donné le domaine d'application retenu pour le projet, la résolution des références ne portait que sur quatre types d'objets : les hôtels, les chambres, les tarifs et les dates. Par ailleurs, les dialogues MEDIA sont généralement assez simples. Dans un dialogue standard du corpus, les énoncés les plus complexes sont ceux où l'utilisateur expose ses exigences. Ensuite, le compère pose des questions et, le plus souvent, l'utilisateur lui donne des réponses courtes et elliptiques. Malgré tout, les expressions anaphoriques y sont très diverses et représentatives de l'ensemble des difficultés classiquement rencontrées lors de la résolution des références.

Le corpus contient par exemple un grand nombre d'expressions définies et toutes les classes d'anaphores qui leur correspondent. Par exemples, suivant la classification proposée par Gardent and Manuelian (2005), « *les* » dans l'expression « *les animaux* » de

4. Parmi les sept systèmes classés lors de ce challenge, LOGUS a été classé quatrième derrière les 2 systèmes du LIMSI et le système du LORIA (approche symbolique) et devant le système du LIA (approche stochastique).

5. Sauf à remodeler profondément le système, solution qui, faute de temps, avait été a priori exclue.

l'extrait de dialogue donné figure 1.3 est une description autonome : elle ne donne pas lieu à une résolution. Dans ce même extrait le « *le* » de l'expression « *le prix de la chambre* » est une description associative ; comme son référent se trouve dans l'énoncé, il n'y a pas de résolution suivant les conventions MEDIA. En revanche, « *la* » dans cette même expression peut être considérée comme une description contextuelle, liée au référent des hôtels précités. Mais ce « *la* » est également coréférentiel dans la mesure où « *la chambre* » s'identifie avec la chambre demandée par l'utilisateur. Le choix MEDIA retient d'ailleurs les deux typologies puisque les segments référentiels de l'annotation contextuelle contiennent les caractéristiques de la chambre demandées par l'utilisateur (*chambre simple*) et les propriétés des hôtels proposés par le compère. Des expressions anaphoriques similaires sont introduites par des adjectifs démonstratifs : *cet hôtel, ces deux chambres, etc.*.

Le corpus contient également un très grand nombre d'expressions anaphoriques incluant une notion d'ordre : *le premier, le dernier, le deuxième, etc.* ou une notion d'exclusion : *l'autre, les autres, les deux autres, un autre, d'autres.*

On trouve également des pronoms qui désignent des référents au sens MEDIA tels que le *la* dans *je la réserve* et le *ils* dans *est-ce qu'ils acceptent les chiens*, etc. alors que, par convention MEDIA, le *y* dans l'expression « *il y a* » n'est jamais coréférentielle.

1.3.2 LOGUS : compréhension en contexte

1.3.2.1 Les principes généraux de la compréhension en contexte

Le principe général adopté pour la résolution des références dans LOGUS reste le même que celui qui prévaut à la compréhension hors contexte : combiner les critères syntaxiques et les critères sémantiques, ceux-ci prévalant sur ceux-là. En effet, si, dans les textes, les critères syntaxiques sont généralement plutôt bien respectés [Boudreau and Kittredge (2005)], il est loin d'en être de même à l'oral où, généralement, l'implicite domine. Le corpus MEDIA permet d'illustrer ces affirmations : par exemple, l'une des formulations les plus fréquentes dans le corpus MEDIA pour demander si un hôtel accepte les animaux est : « *est-ce qu'ils acceptent les chiens* ». Cette utilisation du pluriel est si fréquente qu'il est difficile de penser qu'il s'agit là d'une faute de syntaxe. Elle correspond ici plutôt à une ellipse pour les « *gens de l'hôtel* ». On trouve également des expressions telles que « *celle à cinquante euros* » alors même que le référent logique d'un point de vue strictement textuel est un hôtel. Il s'agit bien là encore d'une ellipse pour « *la chambre de l'hôtel* ».

L'une des relations sémantiques fondamentales utilisée pour la construction de la représentation du sens de l'énoncé dans LOGUS indique une dépendance entre deux objets. Cette relation conceptuelle générique, désignée par « *de* », inclut par exemple les relations *partie-tout* ; elle permet de construire les « chaînes d'objets », un concept utilisé dans la représentation sémantique de LOGUS. Ainsi, « *le prix d'une chambre à l'hôtel Ibis* » correspond à la chaîne :

(tarif []) de (chambre []) de (hotel [(marque "Ibis")])

où l'objet « terminal » est *(hotel [(marque "Ibis")])*, en l'occurrence une entité nommée. La notion qui prévaut à la compréhension en contexte de dialogue pour le système LOGUS est celle de complétion des chaînes d'objets : dans un énoncé, une propriété ou un sous-objet peuvent être complétés par des chaînes de sur-objet du contexte, si cette complétion a un sens, donc si l'ontologie du domaine le permet. Par exemple, pour un énoncé « *quels sont les tarifs* » l'objet *tarif* serait automatiquement complété par la chaîne *(chambre []) de (hotel [(marque "Ibis")])* si cette chaîne est l'objet contextuel le plus proche.

1.3.2.2 LOGUS : mise en œuvre de la résolution des références pour le corpus MEDIA

Les dialogues du corpus MEDIA correspondent à un jeu de rôles relativement simple. Le locuteur énonce des contraintes ; le système propose des noms d'hôtels qui sont censés les satisfaire. Au cours du dialogue, l'utilisateur fait évoluer ses exigences et il y a succès si un accord finit par être trouvé entre celles-ci et les propositions du compère. Dans la pratique, les références liées au dialogue portent essentiellement sur les hôtels proposés par le compère et les sous-objets ou propriétés de ces hôtels liés aux exigences du demandeur.

L'objectif étant de pouvoir mesurer objectivement les performances du système à partir des exigences MEDIA, il convenait de respecter les principes généraux de l'annotation et de se limiter aux références qui correspondaient à des indices linguistiques explicites. Les principes généraux exposés dans le paragraphe précédent ont donc dû être largement amendés. Plus précisément, les références renvoyant à des chambres ou à des tarifs ont été traitées conformément à ces principes, avec un recopiage de la chaîne d'objets correspondante jusqu'au sur-objet de cette chaîne : l'hôtel concerné. Afin de faire un choix parmi les chaînes d'objets contextuels désignées comme sémantiquement possibles par l'ontologie, un traitement particulier a dû être appliqué pour chaque forme linguistique de la référence.

Chacun de ces traitements comporte généralement plusieurs niveaux : une première recherche où critères syntaxiques et sémantiques sont respectés, suivie d'un relâchement progressif des contraintes. Par exemple, « *le premier hôtel* » est d'abord recherché comme le premier élément de la dernière liste d'hôtels énoncés par le compère. Cependant, rien n'est moins sûr que cette liste existe. Les alea du dialogue, tours de parole interrompus par exemple, font que le compère n'a pas forcément proposé les différents hôtels dans un seul tour de parole : si donc cette première recherche ne rend pas de résultat, « *le premier hôtel* » sera alors recherché comme le premier hôtel proposé par le compère. De la même manière, chacune des différentes expressions contenant le mot « *autre* » : *les deux autres, un autre, l'autre,* etc. donne lieu à une stratégie de recherche particulière.

La résolution doit également prendre en compte les erreurs potentielles des locuteurs : erreurs de genre ou de nombre. Des exemples en ont déjà été donnés dans le paragraphe précédent (cf. 1.3.2.1) sous la forme d'ellipses ; il peut aussi s'agir parfois de véritables erreurs de la part du locuteur « *ces deux hôtels* » alors qu'il y a trois hôtels par exemple. Dans la résolution, ce type de contraintes sur les quantités exprimées ne sont relâchées qu'en tout dernier lieu. Il n'est d'ailleurs pas certain qu'elles devraient l'être dans un véritable système : il serait en effet sans doute plus pertinent que le système signifie à son interlocuteur qu'il ne l'a pas compris.

1.3.3 Analyse des résultats

Les tests ont été faits sur 100 dialogues pris au hasard dans le corpus annoté parmi ceux qui n'avaient pas servi au développement du système. Le tableau 1.1 donne les résultats chiffrés ainsi obtenus, et ce, de deux façons différentes. Les premiers sont calculés à partir des segments conceptuels définis dans l'annotation du corpus. Les seconds sont obtenus à partir des objets MEDIA eux-mêmes, un objet étant en général défini par plusieurs segments. Par exemple, un hôtel est en général référencé par deux segments conceptuels : son nom et sa ville. Une erreur sur l'un d'entre eux correspond en fait à une erreur sur l'objet lui-même. En revanche, on peut considérer que l'identification de la taille, de la date et de l'hôtel suffisent à référencer une chambre, même si le (ou les) segments conceptuels qui précisent son prix a été oublié. Lorsque ces nombres ont été collectés, la différence entre les deux méthodes de calcul semblait flagrante. Or, si les résultats obtenus peuvent être très différents lorsqu'ils se rapportent à un ou deux dialogues, il est étonnant de constater que sur l'ensemble des dialogues testés, ils sont finalement globalement très comparables.

Qualitativement, on peut classer les fautes faites par le système en quatre catégories.

Nb de segments MEDIA (A)	Segments corrects (C)	Segments incorrects (I)	Rappel $\frac{C}{A}$	Précision $\frac{C}{C+I}$	F $\frac{2RP}{R+P}$
572	405	42	0,71	0,91	0,80
Nb d'objets MEDIA (A')	Objets corrects (C')	Objets incorrects (I')	Rappel $\frac{C'}{A'}$	Précision $\frac{C'}{C'+I'}$	F' $\frac{2R'P'}{R'+P'}$
212	155	19	0,73	0,89	0,80

TABLE 1.1: Résolution des références dans le corpus MEDIA : résultats chiffrés.

– Comme l'indique le taux relativement bas du Rappel, la première cause d'erreurs est l'absence de détection de certaines références. Les articles définis sont particulièrement redoutables à cet égard. Par exemple, il n'est pas évident de savoir si une condition sur « *les chambres* » se rapportent ou non aux hôtels proposés précédemment. Par ailleurs, à l'instar du « *it* » de la langue anglaise [Boyd et al. (2005)], le pronom personnel « *il* » a beaucoup d'occurrences non référentielles et mériterait un traitement spécifique qui n'est actuellement pas réalisé.
– Certaines erreurs sont dues aux difficultés de compréhension des... énoncés du compère (cf. la discussion de la section suivante). Dans un énoncé tel que
« *Astor Sofitel Novotel arc de triomphe Libertel Arc de triomphe* »,
une segmentation correcte pour détecter les trois hôtels proposés n'est pas si évidente.
– D'autres erreurs sont dues à une mauvaise compréhension de la référence elle-même. Ainsi, pour des expressions telles que « *la deuxième proposition* », « *celui qui reste* », « *celui qui est près de l'autoroute* », LOGUS donne une référence erronée.
– Il semble relativement facile de corriger une bonne partie des bugs précités. En revanche, pour résoudre certaines références, il faudrait munir le système d'une connaissance du contexte autrement plus complexe que celle dont il est actuellement pourvu. Par exemple, dans l'un des dialogues on a les tours de parole suivants :

  Compère : « *... il reste cinq cents chambres disponibles* »

  Compère/Utilisateur : tours de parole concernant la date

  Utilisateur : « *oui vous me la réservez* »

  Le « *la* » fait référence aux cinq chambres demandées par l'utilisateur en début de dialogue, dans des tours de parole relativement éloignés. En relâchant les contraintes de nombre, LOGUS choisit les cinq cents chambres...

En conclusion, les traitements utilisés n'étant pas très sophistiqués, il serait sans doute relativement facile d'augmenter le rappel sans nuire à la précision. En revanche, un certain nombre de références font appel au « sens commun », celui qui est si difficile à cerner et à implémenter...

1.3.4 Discussion et conclusion

Une première critique possible de cette expérience est le caractère quelque peu artificiel de l'exercice.

– La compréhension en contexte de dialogue sur ce corpus a demandé que soit implémentée une compréhension des énoncés compère. Cette tâche n'aurait pas lieu d'être dans le module de compréhension d'un véritable système. Cela dit, comprendre un énoncé-système est beaucoup plus simple que comprendre un énoncé-utilisateur puisque les formes linguistiques utilisées sont connues et stéréotypées. Mais aussi consciencieux et appliqués que soient les compères qui simulent un système dans un corpus élaboré par la technique du Magicien d'Oz, ils ne peuvent pas simuler parfaitement un véritable système. Les expressions qu'ils utilisent ne sont pas entièrement stéréotypées et laissent place à une assez large variabilité. Par ailleurs, il leur arrive également de se tromper, c'est à dire, en l'occurrence, de proposer des réponses non conformes à celles que pourrait proposer le système.
– On peut également discuter le bien-fondé du choix fait dans MEDIA d'avoir utilisé un corpus dont la retranscription « gomme » les erreurs dues à la reconnaissance de la parole car on sait qu'il s'agit là d'une des plus grandes difficultés rencontrées par les systèmes de compréhension de la langue orale. On peut aussi défendre la position selon laquelle les problèmes traités sont suffisamment complexes pour mériter d'être clairement séparés.

L'utilisation du corpus MEDIA pour tester le système LOGUS présente un autre type d'inconvénients, liés à la nature même du système testé.

– Le choix fait dans LOGUS de réaliser une partie de l'interprétation contextuelle dans le module de compréhension n'est pas celui qui prévaut dans les systèmes de DOHM classiques, où cette tâche revient au module de dialogue. Telle qu'elle est envisagée dans MEDIA, la compréhension hors-contexte se présente sous la forme d'une annotation du texte qui consiste à identifier ses différents segments et leur interprétation possible dans le cadre du sysème de dialogue. La compréhension contextuelle consiste alors à résoudre les références, qui correspondent à des segments particuliers essentiellement pronominaux. À l'inverse, dans LOGUS, la compréhension correspond à une interprétation de l'ensemble de l'énoncé, sur le principe que se sont les associations entre les différents concepts qui « font sens ». De ce fait, la compréhension en contexte de LOGUS n'est pas uniquement liée à la résolution des références sur la base d'indices linguistiques et elle s'inscrit naturellement dans le module de compréhension.

  Par exemple, dans MEDIA, l'expression « *est-ce qu'ils acceptent les chiens* » demande la résolution d'une référence à cause du pronom *ils* alors que la question posée sous la forme « *est-ce que les chiens sont acceptés* » n'est pas considérée comme référentielle.

Dans l'approche LOGUS, les deux expressions appellent une résolution identique. L'acceptation des animaux est une propriété relative à un hôtel : la compréhension en contexte demande que soit recherché le (ou les) hôtels éventuellement concerné(s) par cette interrogation. Tester les principes de la compréhension en contexte de LOGUS à partir du corpus MEDIA a donc demandé que soient mis de côté certains des principes fondamentaux de la compréhension contextuelle.

– Enfin, comme il a été dit précédemment, le système LOGUS a été conçu pour essayer d'élargir les domaines potentiels de la compréhension en dialogue homme-machine. Un domaine tel que la réservation hôtelière reste trop étroit pour valider complètement l'approche utilisée. La représentation sémantique des énoncés choisie par les partenaires MEDIA en triplets *(mode, attribut, valeur)* reste pertinente pour une telle application et les formules logiques construites par LOGUS peuvent apparaître comme inutilement complexes.

En même temps et malgré les réserves précédentes, le corpus MEDIA est composé de véritables dialogues dans lesquels, malgré leur relative simplicité, on retrouve la plupart des problèmes classiques liés à la résolution des références. L'annotation des références en fait un corpus de langue française parlée très intéressant pour la mise au point de systèmes de compréhension dans le domaine du DOHM. Il permet en particulier de mesurer à quel point la résolution automatique des références, déjà très complexe dans la langue écrite, devient particulièrement délicate dans la langue orale spontanée.

Les résultats obtenus par LOGUS au cours de cette expérience[6] semblent prouver que la notion de chaînes d'objets utilisée pour la représentation sémantique est un point de départ solide pour la résolution des références et la compréhension en contexte. L'approche demanderait cependant à être validée dans d'autres contextes du DOHM.

6. Deux systèmes seulement ont participé à l'évaluation officielle en contexte de dialogue. Leur évaluation a donné des F-mesures comprises entre 44 et 53%. Cependant, ils étaient soumis à des formats de sortie spécifiques contraignants qui rendent toute comparaison avec LOGUS impossible.

Chapitre 2

Le projet ANR Emotirob : détection linguistique des émotions

Les travaux présentés dans ce chapitre ont été developpés en lien avec des équipes de recherche qui travaillent à la réalisation de « robots compagnons ». Ces collaborations semblent naturelles si l'on considère que le dialogue oral est un mode de communication privilégié pour les robots interactifs, particulièrement lorsqu'ils sont conçus pour interagir avec des enfants ou des personnes âgées. Cependant, l'état de l'art du dialogue oral Homme-machine (DOHM) et de ses domaines connexes tels que par exemple la détection de l'état émotif du locuteur, fait que leur développement se heurte à de nombreuses difficultés.

2.1 Le projet Emotirob

Le développement de robots compagnons susceptibles d'exécuter des tâches complexes et doués d'un système d'interaction enrichi avec les êtres vivants est actuellement un champ d'étude important de la robotique. Dans le domaine de la RAT (*Robot Assisted Therapy*), les premières expérimentations, menées par T. Shibara avec son robot phoque Paro dans une maison de retraite, avaient donné lieu à des résultats prometteurs et prouvé que les robots compagnons pouvaient apporter un peu de réconfort à des personnes fragilisées. Par la suite, Paro a été expérimenté en France : à Kerpape [1] avec des enfants handicapés

1. Centre Mutualiste de Rééducation et de Réadaptation Fonctionnelles de Kerpape, BP 78, 56275 Ploemeur Cedex, France. http ://www.kerpape.mutualite56.fr

puis à l'IEA3 [2], avec des adolescents autistes. Tout en démontrant l'apport possible des robots dans ce type de situation, ces expériences avaient également mis en évidence qu'il était nécessaire d'augmenter les capacités interactives et expressives de ces campagnons artificiels.

Le projet EmotiRob est un projet soutenu par l'Agence Nationale de la Recherche qui s'est déroulé entre 2007 et 2010. Son but était de concevoir un robot compagnon en peluche autonome et réactif, capable d'interagir émotionnellement avec des enfants en longue hospitalisation, afin de leur apporter quelques distractions. Dans le projet ANR Emotirob, l'objectif était de concevoir un robot susceptible d'exprimer des émotions à l'aide de mouvements faciaux joints à l'émission de petits sons. Les expressions émotionnelles du robot devaient apparaître comme une réponse plausible de ce qui pouvait être perçu de l'état émotionnel de l'enfant : la détection de cet état est donc apparue comme un préalable indispensable à une réaction pertinente. Pour ce faire, il a été prévu de combiner des indices visuels (analyse de visage) avec l'analyse des productions orales : indices prosodiques et/ou contenu linguistique des propos de l'enfant.

Notre collaboration à ce projet a consisté à concevoir un module susceptible d'évaluer ce que l'on peut détecter de l'état émotionnel de l'enfant à partir du contenu linguistique de ses propos [Le Tallec et al. (2010)] ; elle a fait l'objet de la thèse de Marc Le Tallec, soutenue début 2012. Les résultats obtenus devaient être combinés avec ceux d'autres équipes qui devaient analyser les indices visuels et prosodiques.

2.2 EmoLogus : détection de l'émotion dans le message porté par les mots

Le premier problème qui s'est présenté était de préciser ce que l'on entend par état émotionnel et de choisir une modélisation, sachant qu'il n'existe pas de réel consensus sur le sujet. Tout le monde s'accorde sur le fait qu'une émotion est un état cognitif complexe influencé par le contexte à court-terme (contexte et historique de l'interaction) comme à long terme (vécu personnel et socioculturel) et dont la perception varie de manière sensible d'une personne à l'autre.

Deux grandes approches ont été utilisées pour caractériser les émotions. La première établit une catégorisation nominale des émotions en classes appelées modalités émotionnelles [Ekman (1999); Cowie and Cornelius (2003)]. Outre l'état émotionnel

2. Centre IEA : Institut d'Éducation Adaptée, Le Bondon, Association Renouveau, Vannes. 26-32 rue Georges Caldray, BP 278, 56007 Vannes.

neutre, on distingue en général la colère, la joie, le dégout, la peur, la surprise et la tristesse. La seconde approche réalise une catégorisation ordinale dans un espace multidimensionnel. Parmi les échelles de valeurs retenues, on trouve le degré d'excitation ou la valence émotionnelle (émotion positive vs. négative). Quelle que soit l'approche suivie, les travaux sur l'émotion dans les dialogues oraux aboutissent à deux conclusions [Devillers and Vasilescu (2005); Lee and Narayanan (2005); Forbes-Riley and Litman (2004); Callejas and Lopez-Cozar (2008)] :

1. en interaction réelle, les tours de parole ne portent majoritairement aucune émotion perceptible : plus de 80% des énoncés peuvent ainsi être qualifiés de neutres ;
2. toutes les expériences d'annotation en émotion présentent un faible accord inter-annotateurs, avec des valeurs de Kappa comprises entre 0,32 et 0,55 [Landis and Koch (1977)]. Par conséquent, une annotation de référence ne peut être obtenue que par vote majoritaire entre plusieurs experts.

La plupart des recherches concernant la détection des émotions en situation de dialogue s'appuient sur des indices prosodiques. Notre objectif était d'étudier ce que peut apporter une détection linguistique. Cet axe ayant été très peu exploré, nous avons fait le choix de caractériser les émotions en précisant leur valence (positive, nulle ou négative) et leur degré d'intensité.

2.2.1 Détection des émotions avec EmoLogus : principe de compositionnalité

Une première approche envisageable pour la détection linguistique des émotions est une approche dite « sac de mots ». Elle consiste à attribuer une valence émotionnelle à chaque mot et à calculer la valence émotionnelle globale d'un énoncé comme la somme (éventuellement normalisée) des valences lexicales de ses termes. Cette première approche, qui ne considère pas la structure de l'énoncé, a été utilisée comme baseline.

À l'inverse, le système EmoLogus que nous avons conçu repose sur le principe que l'émotion contenue dans un énoncé est compositionnelle [Le Tallec et al. (2011)] : elle dépend à la fois de l'ensemble des valeurs émotionnelles des mots et de la structure sémantique qui décrit précisément les relations des mots entre eux. En pratique, les mots non prédicatifs du vocabulaire se voient attribuer une valeur émotionnelle intrinsèque (valence), tandis que les mots prédicatifs, essentiellement verbaux ou adjectivaux, sont associés à des fonctions qui permettent de calculer une valeur émotionnelle à partir de celles des arguments du prédicat.

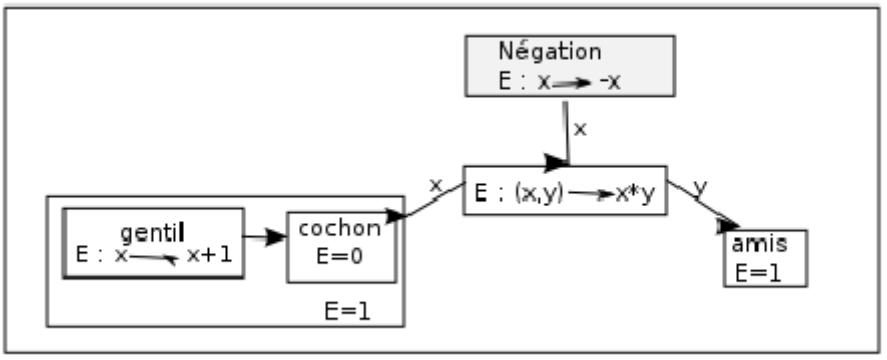


FIGURE 2.1: Exemple de calcul émotionnel d'un énoncé.

La figure 2.1 illustre le fonctionnement du système pour la phrase (extraite d'un conte imaginé par un enfant) : « *il était une fois un gentil cochon qui n'avait pas d'ami* ».

Le calcul commence par la prise en compte de la valence émotionnelle des mots *cochon* ($E = 0$) et *amis* ($E = 1$), qui ne sont que de simples arguments de la formule qui représente la traduction sémantique de l'énoncé. L'adjectif *gentil*, le verbe *avoir* et la négation *ne... pas* fonctionnent comme des prédicats. Le terme *gentil* ayant comme définition émotionnelle $E : x \mapsto x + 1$, le groupe nominal *gentil cochon* va avoir pour valeur $E = 1$. L'application successive de ces prédicats permet d'arriver au résultat final ; pour l'énoncé proposé en exemple, on obtient $E = -1$.

Le lexique que nous avons élaboré comporte une dizaine de fonctions émotionnelles, comme celles présentées en exemple ci dessous [3] :

Décalage positif	E : $x \mapsto x + 1$	exemple : *mignon*
Décalage négatif	E : $x \mapsto x - 1$	exemple : *énervé*
Emotion opposée à l'objet	E : $(x, y) \mapsto -y$	exemple : *casser*
Emotion dépendante des deux arguments	E : $(x, y) \mapsto x * y$	exemple : *perdre*

Le principe de compositionnalité sur lequel repose le système nécessite :

1. une compréhension robuste de la parole spontanée pour fournir la structure prédicative correcte,
2. une base lexicale émotionnelle propre sur laquelle se base le calcul des émotions.

Ces deux éléments font l'objet des deux paragraphes suivants.

3. Dans les faits, les fonctions utilisées sont un peu plus complexes car elles doivent maintenir les valeurs dans un intervalle fixé ([-2 ; +2]).

2.2.2 Adaptation de Logus au langage enfantin

Dans les systèmes de dialogues contraints, les énoncés de l'utilisateur sont censés correspondre à des intentions attendues du système : le « comprendre » consiste donc uniquement à reconnaître le sens de cet énoncé dans une liste de sens prédéfinis. Lorsque l'utilisateur peut prendre des initiatives, la tâche est plus difficile : la compréhension de ses intentions et le sens de son message doit se déduire des énoncés eux-mêmes, replacés dans leur contexte. Une analyse plus fine est donc nécessaire.

Dans le projet EmotiRob, l'objectif est que le robot réagisse aux propos de l'enfant. Il n'y a pas de tâche prédéfinie et on ne peut pas prévoir ce que l'enfant va raconter à son jouet en peluche. Par ailleurs, le pouvoir d'expression de ce dernier se limite exclusivement à de petits sons et des mouvements du visage ; il est probable que l'enfant va interpréter les réponses du robot avant de choisir quelle suite donner à ses propos et toute initiative lui est laissée dans le déroulement de l'interaction. Dans ce contexte, il n'est pas envisageable de chercher à modéliser des cadres sémantiques pouvant représenter à l'avance les intentions du locuteur et la construction d'une représentation sémantique ne peut que s'appuyer sur une analyse aussi précise que possible des propos émis, jointe à une connaissance du « monde de l'enfant ». Dans l'état actuel de l'état de l'art, une telle analyse n'est envisageable qu'en restreignant le vocabulaire et les concepts connus du système. Le choix a été fait de les restreindre à ceux que maîtrise un jeune enfant d'environ 5 ans.

2.2.2.1 Lexique et langage cible

L'une des difficultés rencontrées a été l'absence de corpus représentatif de la tâche. Nous sommes partis du principe que l'enfant aurait la possibilité de raconter des histoires à son compagnon en peluche et basé une partie du travail sur l'étude d'un corpus de conte de fées destiné aux jeunes enfants. Ce corpus a fait l'objet d'une étude linguistique et ses caractéristiques sont décrites dans le chapitre suivant (cf. page 47).

Le langage cible de Logus correspond aux constituants de la formule finale qui représente le sens des énoncés. En l'occurrence, il s'agit donc de l'ensemble des concepts qui définissent les briques de la connaissance sémantique donnée au système.

Pour le définir, nous avons utilisé une étude concernant l'acquisition du langage chez les enfants à partir de 3 ans [Bassano et al. (2005)], menée par une équipe de psychologie cognitive spécialisée dans le développement cognitif de la petite enfance. Basée sur des questionnaires remplis par les parents, elle a permis de faire une liste des concepts que maîtrise le public visé. La base ainsi établie comporte un peu plus de 800 termes : verbes, noms et adjectifs.

Le lexique correspond à l'ensemble des mots que le système est en mesure de traiter. Il s'est avéré très rapidement que la base des termes précédente n'était pas suffisante pour traiter le corpus que nous envisagions d'utiliser. Si donc la base lexicale pouvait servir de cadre pour définir l'ensemble des concepts connus du système, à savoir son langage cible, il convenait d'élargir le vocabulaire que pouvait traiter le système, c'est à dire son langage source. Pour ce faire, nous avons utilisé deux ressources librement disponibles, à savoir Novlex (`http://www2.mshs.univ-poitiers.fr/novlex/`), une base de données lexicales pour les élèves de primaire et Manulex (`http://www.manulex.org/fr/home.html`), une base de données lexicales qui fournit les fréquences d'occurrences de mots calculées à partir d'un corpus de 54 manuels scolaires (1,9 millions de mots). Plus précisément, nous avons fait un croisement de ces deux bases de données et conservé uniquement les termes ayant une fréquence suffisante, pour un lexique final d'environ 6000 mots. Les mots de ce lexique ont ensuite été projetés dans le langage cible défini précédemment.

2.2.2.2 L'organisation des connaissances

Même en restreignant le domaine couvert à celui que maîtrise un jeune enfant, le nombre d'objets et de concepts est plus important et surtout, le domaine qu'ils couvrent est beaucoup plus large que dans l'application pour laquelle Logus avait été conçu, à savoir le renseignement touristique. Il nous a donc semblé qu'une étude linguistique s'avérait nécessaire pour avoir une réelle connaissance des liens sémantiques effectivement utilisés dans le monde des enfants. Un premier regroupement des objets du lexique d'EmoLogus a été effectué suivant des catégories pragmatico-sémantiques. Ce tri a été réalisé manuellement pour les noms et les adjectifs et il a donné lieu à deux classifications, l'une concernant les concepts et l'autre les propriétés. Une étude sur le corpus a servi de base d'étude à cette organisation. Ensuite, le problème se posait des liens sémantiques qui reliaient les

verbes aux classes de concepts précédemment définies. L'étude sur corpus s'est basée sur les travaux de Hanks (2008b), qui reposent sur l'idée que le sens d'un mot se définit par ses usages. L'objectif est de vérifier dans des données réelles l'existence de liaisons sémantiques afin de faire évoluer la catégorisation et de trouver les compléments prototypiques des verbes. Cette étude a été réalisée lors des travaux de thèse d'Ismaïl El Maarouf et les détails en sont donnés page 47.

2.2.3 Norme émotionnelle et définition des prédicats

La première ressource nécessaire à EmoLogus est la valeur émotionnelle de chacun des lexèmes du vocabulaire. La psychologie expérimentale a depuis longtemps défini des normes lexicales émotionnelles associant à un mot sa valence. À notre connaissance, lorsque ce travail a été mené, il n'existait que deux études de ce type spécifiques aux enfants : Vasa et al. (2006) pour l'anglais et Syssau and Monnier (2009) pour le français. Ces études ont montré que si les normes émotionnelles sont stables chez l'adulte et l'adolescent, elles varient fortement au cours des premières étapes du développement cognitif de l'enfant. La norme émotionnelle de Syssau et Monnier repose sur une catégorisation ordinale en trois modalités (négatif, neutre, positif). Nous avons complété ces travaux par l'évaluation de 80 mots absents de la norme et présents dans notre lexique qui couvre globalement le vocabulaire d'un enfant de 7 ans. Cette nouvelle norme a été établie pour deux âges différents (5 et 7 ans) par expérimentation dans 4 écoles primaires.

La baseline repose entièrement sur la norme lexicale. Par contre, dans EmoLogus, la norme lexicale associée aux mots prédicatifs est remplacée par une fonction émotionnelle. Cette seconde norme a été obtenue suivant une procédure d'annotation réalisée par 5 experts adultes. Chacun d'entre eux a proposé au plus deux définitions pour chaque prédicat, avant qu'une procédure de consensus ne définisse la fonction retenue à chaque fois. Il est intéressant de noter qu'il a finalement été possible d'arriver à un accord total. La notion de fonction prédicative émotionnelle semble donc paradoxalement être plus stable que celle de valence émotionnelle : les sujets semblent arriver plus facilement à un accord sur l'émotion portée par un prédicat (comme par exemple *tuer*) que sur des mots simples (*banane, école*), où le vécu personnel de chacun entre plus facilement en jeu.

Il est important de préciser que dans le cas où la structure sémantique de l'énoncé ne peut pas être définie ou si elle ne peut l'être que partiellement, sa valeur

émotionnelle s'obtient en appliquant les principes de la baseline, c'est à dire en moyennant les valeurs émotionnelles des mots ou des groupes de mots reconnus et analysés.

2.3 Expérimentations et résultats

Dans le cadre du projet Emotirob, le système EmoLogus rend une mesure de l'émotion contenue dans un énoncé hors-contexte.

Le comportement hors contexte du système a été évalué sur le corpus Brassens [Le Tallec et al. (2009)] comportant 173 énoncés enfantins annotés en émotions. Pour réaliser une annotation hors-contexte des énoncés, ces derniers ont été présentés dans un ordre aléatoire à 5 annotateurs adultes qui devaient leur attribuer une modalité parmi 5, de -2 (très négatif) à +2 (très positif). Cette annotation combinait donc valence émotionnelle (positif/neutre/négatif) et intensité de l'émotion portée. La valeur référence de chaque phrase a été déterminée par un vote majoritaire sur les décisions des experts. Les résultats de cette annotation ont été présentés dans Le Tallec et al. (2009). La table 2.1 compare la précision d'annotation du système EmoLogus et de la baseline sur le corpus de test.

	EmoLogus	Baseline
Précision	90,00%	68,80%

Table 2.1: Précision d'annotation du système EmoLogus et de la baseline.

EmoLogus présente des résultats encourageants avec un taux de bonne réponse à 90%, bien supérieur à ce que donnerait une procédure basique de calcul des émotions.

La table 2.2 présente la matrice de confusion des erreurs pour les deux systèmes testés. Les valeurs en abscisse représentent les valeurs données par l'annotation humaine, alors que les valeurs en ordonnée représentent les valeurs données en sortie par le système. L'erreur la plus grave que peut commettre un système est de détecter une valence émotionnelle opposée à celle attendue. Ce type d'erreur n'est jamais observé avec EmoLogus, à la différence de ce que l'on peut observer pour la baseline. La majorité des erreurs (47%) consiste à attribuer une émotion neutre à un énoncé annoté comme positif ou négatif. On pourrait qualifier cette situation

EmoLogus

	-2	-1	0	1	2
-2	4	2	0	0	0
-1	2	18	0	0	0
0	1	5	116	2	0
1	0	0	3	16	1
2	0	0	0	1	2

Baseline

	-2	-1	0	1	2
-2	4	0	0	1	0
-1	3	12	7	0	0
0	0	6	90	4	0
1	0	4	18	11	1
2	0	0	7	3	2

Table 2.2: Matrices de confusion des erreurs du système EmoLogus (à gauche) et de la baseline (à droite).

de problème de rappel, si le neutre ne constituait pas un état émotif propre. Enfin, l'insertion d'une émotion inexistante aux yeux des experts ne concerne que 18% des erreurs d'EmoLogus. Une part significative des erreurs observées ne concerne par ailleurs que l'estimation de l'intensité émotionnelle, la valence étant correctement détectée (exemple : positif vs. très positif). Si l'on se limite à la détection de la valence émotionnelle, la précision du système EmoLogus atteint alors 94%.

L'analyse qualitative des erreurs montre qu'EmoLogus rencontre des difficultés à modéliser ce qui pourrait être qualifié d'isotopie émotionnelle. Prenons l'exemple du verbe *être enfermé*. En toute logique, les experts ont associé à ce prédicat la fonction E : $x \mapsto -x$. Ainsi, si le sujet est associé à une valence positive (exemple : *princesse*), son enfermement est perçu négativement. Cependant, si l'on adopte cette fonction, un sujet non porteur d'émotion, ou mal identifié ($E = 0$) conduira à considérer que le couple *(sujet verbe)* est porteur d'une émotion neutre. Pourtant, il semble que les annotateurs ressentent une légère émotion négative dans ces situations. Il semble donc nécessaire d'attribuer une valeur négative par défaut à ces prédicats lorsque la valeur émotionnelle de leur argument est inconnue ou neutre. C'est l'émotion portée par les arguments qui modifie le comportement émotionnel du prédicat, un peu comme, en sémantique, les sèmes isotopiques sont activés en contexte.

Un autre problème, circonscrit principalement à certains adjectifs, résulte de ce qu'on pourrait qualifier de *polysémie émotionnelle*. Prenons le cas de l'adjectif *petit*. Globalement cet adjectif a tendance à changer l'argument qu'il qualifie pour le rendre plus positif (exemple : *le petit loup*) : $x \mapsto x + 1$. D'autres comportement émotionnels doivent toutefois être considérés. Par exemple, les annotateurs ne perçoivent pas de décalage vers le positif sur un exemple tel que *la petite maison*. Des expressions telles que *petit salaire* ou *petit boulot* sont également des exemples

où *petit* est généralement employé dans un sens péjoratif. Il semble donc que certains prédicats doivent se voir attribuer plusieurs comportements émotionnels qui sont à désambiguïser dans le contexte local de l'énoncé.

Il reste enfin à aborder la question de l'influence du contexte général du discours sur la perception des émotions, qui n'a pas été évaluée dans le cadre de ces premières expérimentations.

2.4 Conclusion et perspectives

Si l'on s'en tient à l'attribution d'une valeur émotionnelle aux phrases d'un conte pour enfants, les premières évaluations du système EmoLogus sont encourageantes. Les résultats obtenus semblent en effet montrer qu'il est possible de détecter la bonne émotion dans un énoncé dans 90% des cas. Un point positif important est que le système n'annote jamais un énoncé avec une émotion inverse, une erreur qui conduirait à une mauvaise réaction du robot. Un problème essentiel, et qui n'a été que partiellement traité, est celui de la prise en compte du contexte. Outre les difficultés classiques liées à la résolution des anaphores, cette détection soulève deux problèmes.

1. La dynamique des émotions dans un récit, qui définit comment la valeur émotionnelle d'un énoncé dépend de celle des énoncés précédents.
2. Le suivi de la valeur émotionnelle des « personnages » dans un récit.

Quelques travaux en ce sens (qui ont fait l'objet d'un stage de master 2) ont montré la complexité de ces deux points.

Du point de vue de la tâche initialement envisagée, EmoLogus n'a malheureusement pas fait l'objet d'évaluations en situation. D'une part, certains maillons du projet n'ont été que partiellement réalisés. D'autre part, la mise en place d'évaluations d'un robot compagnon avec de jeunes enfants hospitaliés pose des problèmes déontologiques et juridiques qui n'ont pas été entièrement résolus avant la fin du projet.

Les travaux décrits dans ce chapitre ne sont pas sans lien avec les recherches concernant la détection d'opinion. Cette tâche fait actuellement l'objet de nombreux travaux, à cause de son intérêt applicatif évident et des nombreux avis que déposent les internautes sur le web. Si les textes à analyser partagent avec la langue orale

leur variabilité et leur agrammaticalité, l'utilisation d'un système tel que LOGUS serait difficile. En effet, parce qu'il vérifie que les associations de mots « font sens », LOGUS ne peut être efficace que si l'on construit une base des concepts manipulés, ce qui peut constituer un travail très lourd, voire irréaliste. Néanmoins, l'approche prédicative initiée dans ces travaux mériterait d'être testée, quitte à la restreindre à certains verbes ou modifieurs.

Chapitre 3

Emotirob : interaction langagière et modélisation des connaissances enfantines

Comme ceux du chapitre précédent, les travaux décrits dans ce chapitre sont liés à l'amélioration des capacités réactives des robots-compagnons. En l'occurrence, l'objectif était de doter le robot de capacités cognitives et langagières pour enrichir son interaction avec un jeune enfant. Ces expérimentations ont été développées dans le cadre du projet MAPH [Achour et al. (2008a)], un projet satellite du projet ANR Emotirob, et elles ont fait l'objet des travaux de thèse d'Amel Achour, soutenue fin 2010 [Achour (2010)].

3.1 Principes et approches

Concevoir un module d'interaction cognitive robot-enfant est un sujet de recherche vaste et ambitieux et de nombreuses pistes de recherche étaient possibles. Les travaux présentés ici ne sont qu'une phase exploratoire et il était difficile d'aller très loin alors même qu'on ne disposait pas de la première version - non langagière - du robot.

3.1.1 Interactions langagières élémentaires

L'objectif initial a été de doter le robot de la possibilité de répondre par des mots à ceux prononcés par l'enfant [Achour et al. (2008a)]. La première partie des travaux a permis de concevoir quelques interactions langagières simples qui pourraient avoir lieu entre un jeune enfant (d'environ 5 ans) et un robot compagnon [Achour et al. (2008b)]. La liste des mots que maîtrise un enfant d'environ 5 ans issu des travaux de Bassano et al. (2005) a déjà été déjà citée dans le chapitre précédent (cf. page 22). Ce corpus a servi de base à la construction d'une taxonomie des connaissances qui mêle des critères syntaxiques, sémantiques et affectifs. Jointe à un ensemble de propriétés définies pour certains mots du corpus, celle-ci a permis de définir des coefficients de rapprochement entre les différents concepts.

Nous avons ensuite implémenté une première version d'un module de génération de phrases et quelques jeux interactifs entre l'enfant et le robot sur la base de phrases simples, affirmatives ou interrogatives, formées à partir des mots du corpus. Cette génération des phrases reposait essentiellement sur le calcul des coefficients de rapprochement entre les composants de la phrase d'entrée et ceux de la phrase de sortie ; en même temps, elle imposait une certaine cohérence dans la formation de la phrase dans un contexte réaliste. Les jeux proposaient un ensemble d'interactions susceptibles de distraire l'enfant et de tester ses connaissances. Ils reposaient sur les relations sémantiques définies entre les concepts. Un jeu de devinettes permettait par exemple de tester les connaissances de l'enfant sur les propriétés des animaux du corpus.

Cependant, si la taxonomie réalisée correspondait à une organisation claire des connaissances, elle présentait l'inconvénient majeur d'être rigide et figée. De ce fait, les capacités de réaction du robot apparaissaient pauvres et pire encore, stéréotypées. Par ailleurs, cette organisation des connaissances avait été pensée par des adultes telle que ceux-ci imaginaient le monde enfantin, sans qu'aucune vérification n'ait été faite de son adéquation avec la façon dont les enfants organisent leurs connaissances.

Nous avons alors décidé de réaliser une modélisation vraisemblable des connaissances d'un jeune enfant, ou tout au moins d'une partie d'entre elles, et de « simuler automatiquement » son processus cognitif, tant pour ce qui est de l'organisation de connaissances déjà acquises que pour l'insertion de nouveaux concepts dans une organisation existante.

3.1.2 Modélisation cognitive du vocabulaire

Les conditions que nous avons imposées à la modélisation des connaissances relatives au vocabulaire d'un jeune enfant ont été les suivantes :

– être fidèle à une perception du monde des enfants plausible en s'appuyant sur des données provenant des enfants eux-mêmes ;
– avoir une structure claire et souple permettant une éventuelle mise à jour ultérieure et son enrichissement.
– offrir la possibilité de validation aux différentes étapes de son développement.

Dans son Histoire Naturelle de 1749, Georges de Buffon affirme déjà que « *le seul moyen de faire une méthode instructive et naturelle, est de mettre ensemble des choses qui se ressemblent et de séparer celles qui diffèrent les unes des autres* ». Plus récemment, Lakoff (1987) déclare que « *la classification semble être chez l'être humain un processus mental naturel et spontané qui lui permet de représenter le monde et les connaissances* ». Selon [Sloutsky (2003)] et [Mareschal and Quinn (2001)], non seulement la faculté de catégorisation permettrait à l'individu d'organiser ses connaissances et de mieux exploiter ses ressources cognitives mais également, elle lui faciliterait l'apprentissage et l'acquisition de nouvelles connaissances.

Étant donné l'intérêt de la catégorisation dans la représentation des connaissances et des informations, d'abondantes recherches lui ont été consacrées. Ainsi, de nombreuses théories ont vu le jour, qui visent à modéliser et simuler automatiquement le processus de catégorisation. Par ailleurs, le développement de l'informatique a permis le développemement de techniques et d'outils qui peuvent être mis en œuvre très rapidement et sur de grandes quantités de données.

3.2 Modélisation des connaissances et catégorisation

Si donc la catégorisation est une piste intéressante pour la modélisation et la simulation du processus cognitif, le choix des méthodes à adopter dépend étroitement de l'objectif de l'application ainsi que de la nature des données. Dans le but de construire une modélisation plausible d'une partie du monde cognitif d'un jeune enfant, plusieurs d'entre elles ont été testées.

Deux domaines sémantiques ont été choisis pour faire cette étude : le domaine des animaux et le domaine des aliments. En effet, dans ces deux domaines, les noms communs du vocabulaire enfantin présents dans le corpus sont nombreux et variés. On y dénombre en effet 81 noms d'animaux et 112 noms d'aliments, ce qui correspond à une base de connaissances bien adaptée aux travaux envisagés.

3.2.1 Sélection des variables de classification

Quelle que soit l'approche choisie, l'étape de la sélection des paramètres de classification est indispensable et particulièrement délicate car c'est d'elle que dépend la qualité des résultats ultérieurs.

La méthode retenue pour sélectionner les variables de classification s'est déroulée en deux étapes.

1. La première étape a consisté à collecter des données. Des enfants d'âge préscolaire [1] ont été amenés à réaliser des classifications et à décrire les motivations de leurs choix. Cette collecte a abouti à une classification de référence qui représente la classification « experte » des enfants.
2. La seconde étape a été basée sur les critères de choix désignés par les enfants eux-mêmes ; elle a consisté à sélectionner un ensemble de variables qui permette d'obtenir une classification aussi proche que possible des classifications observées lors de la collecte.

En l'occurrence, la catégorisation obtenue à partir d'une classe maternelle particulière ne peut évidemment pas être considérée comme représentative d'une classe d'âge d'enfants dans un contexte culturel donné. Cependant, l'objectif final étant de doter un robot d'une connaissance plausible, la représentativité du groupe d'enfants n'était pas un problème dans ce cadre applicatif particulier.

Concernant les animaux, les variables qui ont été ainsi retenues sont les suivantes :
– le nombre de pattes (0, 2, 4 ou plus) ;
– la taille (petit, moyen, grand, très grand) ;
– le moyen de locomotion (voler ou non, marcher ou non, nager ou non) ;
– le mode de vie : l'animal vit dans la ferme, est sauvage ou autre ;
– la nuisance : l'animal pique ou non, griffe ou non, etc. ;

1. Les expérimentations ont été menées dans une classe maternelle de la ville de Lorient.

– l'effet produit : l'animal fait peur ou non.

3.2.2 Les cartes auto-organisatrices de Kohonen

L'objectif principal des méthodes de classification automatique est de répartir les éléments d'un ensemble en groupes, c'est-à-dire d'établir une partition de cet ensemble. La détermination des groupes repose sur le principe qui consiste à minimiser la distance entre les individus d'un même groupe, tout en maximisant celle qui sépare les individus de groupes différents.

Les méthodes de classification existantes se répartissent en méthodes de classification hiérarchique et méthodes de classification à plat. Dans les méthodes de classification hiérarchique, on ne se contente pas d'une partition des données, on établit également une hiérarchie des parties. Les enfants interrogés lors de la collecte de données ont réalisé des classifications à plat des individus sans faire des agrégations de groupes comme le fait la classification hiérarchique. Pour cette raison, une simple classification à plat nous a semblé mieux adaptée à la modélisation de leur processus cognitif.

Parmi les méthodes de classification à plat, la méthode des cartes auto-organisatrices introduite par Kohonen (1982) ou SOM (Self Organizing Maps) se distingue par le fait que les cartes obtenues respectent la topologie de l'espace des données. Plus précisément, les SOM répartissent les individus dans des classes entre lesquelles existe une notion de voisinage. Cette proximité entre les classes est porteuse d'informations : elle permet d'évaluer la proximité topographique entre les individus dans la carte et donne une vision globale sur l'organisation des connaissances. C'est cette méthode qui a été expérimentée pour modéliser le processus de classification enfantin.

Dans son principe, l'algorithme de Kohonen (2001) s'inscrit dans le cadre des réseaux de neurones à apprentissage non supervisé. Il représente une généralisation de la méthode des « centres mobiles » ou « nuées dynamiques » en introduisant une notion de voisinage entre les neurones. Ceux-ci sont organisés selon une structure choisie a priori que l'on appelle aussi carte de Kohonen et qui peut être de dimension un (ficelle) ou deux (grille). En partant d'un ensemble de données $D = \{X_1, X_2, ..., X_N\}$ contenant N observations, l'objectif de l'algorithme de Kohonen est de faire correspondre à chaque unité de la carte ou neurone u un vecteur

code C_u et un sous-ensemble G_u de D. Chaque sous-ensemble G_u représente une *classe* formée par les observations $X_i \in D$ les plus proches de C_u au sens d'une distance définie sur $\mathbb{R}^p$ (p étant la dimension de l'ensemble D). Les classes associées aux différents neurones de la carte forment une partition de l'ensemble de données D.

Plus précisément, le déroulement de l'algorithme est indiqué dans la figure 3.1 ci-dessous.

Étapes de l'algorithme :

1. Choix des dimensions de la grille neuronale G.
2. Initialisation aléatoire des vecteurs w_r, $r \in G$, vecteurs référents de chacun des neurones.
3. Présentation du corpus d'apprentissage D un certain nombre de fois :

tant que ($t < tmax$) et (variation des vecteurs codes $\geq a$ fixée)
pour chaque vecteur v **de** D **(ordre aléatoire)**
* déterminer le neurone s dont le vecteur référent approche au mieux v :
$s = \text{argmin}_{r \in A} \parallel v - w_r \parallel$
* rapprocher de v les vecteurs référents de s et ceux de ses "voisins" :
$w_r^{t+1} = w_r^t + \Delta w_r^t$ avec
$\Delta w_r^t = \epsilon(t).h(r,s,t).(v - w_r^t)$ (1)
$\epsilon(t)$ est le coefficient d'apprentissage et
$h(r,s,t)$ est la fonction de voisinage.
fin pour
fin tant que

FIGURE 3.1: L'algorithme de Kohonen.

– Le choix des dimensions de la carte est une étape particulièrement importante qui conditionne la qualité de la classification obtenue : la méthode que nous avons choisie est détaillée dans la section suivante (cf. 3.3.1).
– Le nombre d'itérations maximal ($tmax$) est choisi empiriquement. Kohonen (1990) suggère qu'il soit au moins égal à $500 \times U$, U étant le nombre de neurones.
– Le neurone gagnant est celui dont le vecteur référent est le plus proche du vecteur présenté au sens de la distance choisie : en l'occurrence, la distance entre vecteurs que nous avons utilisée est celle du χ^2.
– La fonction de voisinage h pondère la correction à apporter aux vecteurs référents des neurones ; elle détermine la déformation subie par la carte après

qu'a été choisi le neurone s qui approche au mieux le vecteur v. Plus un neurone est proche de s, plus son vecteur est rapproché du vecteur v.
La fonction est en général définie par $h(r, s, t) = \exp\Big(-\frac{\| \vec{r} - \vec{s} \|^2}{2\sigma^2(t)}\Big)$ où

* $\| \vec{r} - \vec{s} \|$ représente la distance entre neurones dans la grille. Dans notre cas, la distance entre deux neurones (i, j) et (i', j') de la grille a été définie par $\max(|i - i'|, |j - j'|)$. Avec cette définition, les cellules situées par exemple à une distance 1 d'une cellule donnée sont les 8 cellules qui lui sont adjacentes.
* $\sigma(t)$ est un coefficient de voisinage qui décroît en fonction du temps. La fonction permet de règler la distance à partir de laquelle la déformation n'est plus sensible.
* Le coefficient d'apprentissage $\epsilon(t)$ décroît en fonction du temps : on peut par exemple choisir $\epsilon(t) = \epsilon_0 \times \exp(-t/t_{max})$.

La règle (1) de la figure 3.1 permet ainsi, à chaque étape, d'accentuer la ressemblance entre une donnée et le vecteur référent du neurone dont elle est la plus proche. De plus, la donnée modifie également les vecteurs référents des neurones voisins de son neurone gagnant.

– Ainsi, l'algorithme de Kohonen minimise la somme des carrés des écarts de chaque observation non seulement à son vecteur code, mais aussi aux vecteurs codes voisins. La convergence de l'algorithme de Kohonen n'a été prouvée que pour une structure de carte en ficelle. Toutefois, il a été montré empiriquement que dans la majorité des cas, l'algorithme converge vers un minimum local. La carte résultante dépend très largement des données initiales et de l'ordre de présentation des observations.

3.3 Cartes de Kohonen : méthodologie et résultats

3.3.1 Mise en œuvre - Résultats

En classification non supervisée, la détermination du nombre de classes est un problème ouvert et difficile. Or, ce choix conditionne très fortement la qualité même de la classification. Dans le cas des cartes de Kohonen, le problème est de choisir les deux dimensions de la carte, sachant que leur produit correspondra au nombre maximal de classes possibles. Le principe de la méthode du Gap consiste à

comparer la performance de l'algorithme de classification (mesurée en terme d'un critère d'évaluation de la qualité du clustering comme la dispersion intra-classe) obtenue pour le jeu de données initial à celle obtenue pour un jeu de données aléatoire (ne présentant pas de classes), et cela en fonction du nombre de clusters. Le « bon » nombre de clusters correspond à celui où le « gap » entre les deux performances est le plus important. L'expérimentation de cette méthode sur les données relatives aux animaux nous a conduit au choix d'une carte de Kohonen 3×5.

Par ailleurs, si le choix de la dimension de la carte de Kohonen a une grande importance sur le résultat obtenu, d'autres paramètres tels que l'initialisation des neurones et l'ordre de présentation des observations influent également sur le résultat obtenu. Ainsi, avec la base de données des animaux, les différentes exécutions de l'algorithme ont fait apparaître des classes stables, telles que celle des « animaux sans pattes » ou des « animaux de ferme à quatre pattes » ; et des animaux peu stables, qui changent de clsse d'une classification à l'autre tels que le *canard* ou le *lapin*, des animaux dont on ne sait pas trop s'ils sont sauvages ou non, le *hérisson* à cause de ses piquants et le *pingouin* en tant qu'oiseau atypique.

En l'occurrence, une technique de ré-échantillonage a été appliquée sur les données : elle a permis de déterminer une « table moyenne de voisinage » calculée à partir des tables de Kohonen obtenues sur les différents échantillons. Ensuite, l'algorithme de Kohonen a été exécuté plusieurs fois sur l'ensemble des données et la table retenue a été la plus proche (au sens de la distance de Froebenius entre matrices) de la table moyenne.

La figure 3.2 donne la carte de Kohonen qui a été ainsi obtenue avec les animaux du lexique. Cette carte comporte 12 classes. Elle fait apparaître par exemple la proximité entre la classe des oiseaux de basse-cour et celle des autres animaux « de ferme » ; elle montre également la singularité du *pingouin*, isolé dans une classe mais malgré tout voisin des autres oiseaux ; par ailleurs, le *dragon* et le *dinosaure* se trouvent réunis dans la même classe en tant qu'animaux fantasmatiques ou disparus. Cette classification, peu conforme à ce que l'on pourrait obtenir à partir de critères scientifiques, nous a semblé plausible dans la perspective d'une représentation du monde cognitif d'un jeune enfant.

canard cygne moineau chouette pigeon pivert rouge-gorge chauve-souris perroquet tourterelle aigle	pingouin	escargot poisson-rouge requin serpent ver de terre baleine dauphin	lapin tortue	poule coq poussin paon autruche
		crapaud grenouille kangourou		agneau âne chèvre cochon mouton poney taureau vache chat hamster chien veau cheval
coccinelle mouche papillon abeille moustique libellule	araignée pou fourmi cafard crabe	écureuil marmotte rat singe souris hérisson taupe raton laveur koala castor	chameau crocodile éléphant girafe léopard lion loup ours panda renard tigre zèbre rhinocéros hippopotame chamois biche	dragon dinosaure

FIGURE 3.2: Carte de Kohonen 3×5 des animaux du lexique.

3.3.2 Classification mixte

La classification « mixte » est une approche qui consiste à coupler l'algorithme de Kohonen et la Classification Ascendante Hiérarchique (CAH) dans la réalisation d'une répartition de l'ensemble des données. Le but de cette méthode est d'avoir plusieurs niveaux de classification en prenant initialement un grand nombre de clusters pour la carte de Kohonen et en affinant la classification obtenue à l'aide d'une CAH sur les vecteurs codes des différentes classes. La méthode de Ward a été utilisée pour la CAH ; elle consiste à réunir les deux clusters dont le regroupement fera le moins baisser l'inertie interclasse. La distance entre deux classes est celle de leurs barycentres au carré, pondérée par les effectifs des deux clusters.

Selon le nombre de clusters choisi pour la CAH, on obtient un certain nombre de méta-classes ou *super-classes* que l'on peut visualiser sur la carte. Comme l'algorithme de Kohonen respecte la topologie, les super-classes regroupent forcément des classes adjacentes. De ce fait, la classification présente un double avantage :

– elle permet d'avoir plusieurs niveaux de granulométries en allant de classifications « grossières » vers des classifications de plus en plus fines.
– elle permet également de vérifier la proximité entre unités voisines sur la carte.

La figure 3.3 page 38 présente une classification mixte des animaux avec une grille de taille 6×6 et 11 méta-classes. Cet exemple illustre comment les classes ont été fusionnées en allant de classes plus spécifiques vers des méta-classes plus génériques : la méta-classe des animaux sauvages à quatre pattes par exemple (en haut à droite dans la classification donnée) se partage en plusieurs classes plus spécifiques : la classe des animaux grands et qui mordent (*crocodile, léopard, etc.*, ceux qui sont très grands (*chameau, etc.*), les grands qui sont considérés comme inoffensifs, etc. Par ailleurs, on peut remarquer également que des méta-classes caractérisées essentiellement par un critère dominant comme le nombre de pattes se divisent en plusieurs classes suivant un critère moins influent dans la classification. Par exemple, la méta-classe des animaux sans pattes se partage en animaux aquatiques et terrestres. De même, celle des animaux ayant plus que quatre pattes (en haut à gauche) comporte la classe des animaux qui volent (*coccinelle*, etc.) et ceux qui ne volent pas (*araignée, fourmi*, etc.).

coccinelle mouche papillon abeille moustique libellule	araignée fourmi cafard crabe pou		dinosaure dragon	chameau elephant girafe	crocodile léopard lion loup tigre
					renard rhinocéros ours
poisson-rouge requin baleine dauphin			crapaud grenouille kangourou	singe	chamois panda zèbre biche hippopotame
	escargot serpent ver-de-terre		écureuil marmotte hérisson souris taupe koala rat	castor raton-laveur	
aigle					hamster tortue lapin chat
moineau chouette pigeon pivert rouge-gorge chauve-souris perroquet tourterelle	canard cygne	pingouin	autruche paon poussin poule coq		agneau chèvre cochon mouton poney taureau vache âne

FIGURE 3.3: Classification mixte des animaux : grille 6×6 et 11 méta-classes

3.4 Cartes de Kohonen : acquisition de nouvelles connaissances

La capacité d'acquérir de nouvelles informations et surtout de pouvoir les organiser à côté des données déjà acquises traduit une certaine flexibilité du processus cognitif et une capacité d'évolution qui donne un côté « intelligent » au robot. Sur ce point, les cartes auto-organisatrices possèdent des propriétés qui permettent de répondre à l'objectif d'enrichissement et d'évolution de la représentation du monde par l'apport de nouvelles connaissances. En effet, les SOM sont bien adaptées à

l'ajout de nouveaux individus dans une classification existante, que ce soit par l'affectation du nouvel individu à la plus proche classe ou par la réalisation d'une nouvelle classification. Les cartes de Kohonen permettent donc de suivre l'évolution de la représentation du monde au fur et à mesure que de nouvelles connaissances apparaissent.

Pour simuler le processus d'acquisition de nouvelles connaissances, on ajoute un nouvel individu à la classification déjà établie. Deux cas sont alors possibles.

– Dans le premier cas, on considère que l'effet apporté par l'ajout d'un seul individu reste négligeable devant le nombre total des individus déjà classés et que, de ce fait il ne doit pas modifier la classification initiale. On peut alors affecter le nouvel individu à la classe qui lui est la plus proche dans la carte sans autres transformations des classes existantes. Pour ce faire, on calcule sa distance aux différents neurones de la carte et on l'affecte à la classe correspondant au neurone qui lui est le plus proche.
– Dans le second cas, on réalise une nouvelle classification en ajoutant le nouvel individu à l'ensemble des individus initial. On aboutit alors à une nouvelle carte traduisant une réorganisation qui correspond à une adaptation des connaissances due à l'acquisition de nouvelles informations.

Le première de ces solutions suppose qu'un léger apport de données ne doit pas trop impacter l'organisation des connaissances alors que la deuxième traduit une plus grande flexibilité du processus cognitif. Dans le cas où c'est la première qui est retenue, il faut penser à déterminer un seuil correspondant à la quantité maximale de nouvelles informations à partir duquel on doit entamer une totale réorganisation des connaissances.

3.4.1 Classement d'un individu à valeurs manquantes

Une donnée manquante peut correspondre à une propriété que l'on ignore ou que l'on ne veut pas fournir pour une quelconque raison. Le vecteur représentatif d'un individu à propriétés manquantes est donc incomplet dans les champs correspondants, ce qui pose a priori un problème pour l'exécution de l'algorithme de classification. Cependant, il a été montré que l'algorithme de Kohonen présente une grande robustesse face aux données manquantes [Cottrell et al. (2003); Ibbou (1998)]. Le principe consiste à ignorer les composantes manquantes de l'individu et à faire les calculs de distance le concernant uniquement à partir des composantes

disponibles. Si donc un nouvel individu que l'on veut placer dans une classification existante a des propriétés manquantes, on peut l'affecter à la classe qui correspond au neurone le plus proche avec ce calcul de distance tronqué. Ainsi, aucune interpolation ni estimation des valeurs manquantes n'est effectuée a priori, ce qui rend la procédure d'ajout plus fiable en évitant la répercussion d'erreurs commises lors de l'approximation.

Cette approche a été expérimentée en insérant un nouvel animal dans la carte de Kohonen des animaux déjà constituée. L'expérience a consisté à classer un nouvel individu à partir des attributs que l'on peut deviner à partir de son image. Ainsi, le *mulet* a été défini par les propriétés suivantes :
– il a quatre pattes,
– il est grand,
– il ne vole pas,
– il marche,
– il ne nage pas,
– il vit dans la ferme,
– il ne pique, ne griffe ni ne mord, etc.
la propriété manquante étant celle qui consistait à savoir s'*il fait peur*.

En appliquant les principes précédemment décrits, le *mulet* a été classé avec les autres grands « animaux de ferme », avec l'agneau, l'âne, la chèvre, etc. comme on peut le voir sur la figure 3.4.

3.4.2 Estimation de propriétés manquantes

Une fois que l'individu à valeurs manquantes a été affecté à sa nouvelle classe, il reste à estimer ses propriétés manquantes. On utilise pour cela un processus d'*inférence* qui consiste à attribuer les propriétés d'une catégorie donnée à ses différents membres.

Comme l'algorithme de Kohonen finit avec un apprentissage à rayon nul, les vecteurs codes à la fin de l'apprentissage peuvent être considérés comme une approximation des vecteurs moyennes des classes correspondantes. Il est donc naturel de penser à utiliser les vecteurs référents des neurones (les vecteurs codes) pour estimer les valeurs manquantes du nouvel individu. Toutefois, cette méthode d'estimation n'est précise que quand les classes obtenues sont homogènes et bien

séparées et que les variables sont corrélées entre elles. Supposons par exemple que, dans la classification obtenue pour les animaux, l'on prenne un animal fictif ayant les mêmes caractéristiques physiques que le *lion* dont on veut estimer si c'est un animal qui peut « faire peur ». En l'ajoutant à la classification de Kohonen, ce nouvel animal serait très probablement affecté à la famille des animaux « sauvages à quatre pattes », dans laquelle sont classés des animaux aussi différents que l'*ours* ou la *biche*. Or, cette classe est très hétérogène vis à vis de la propriété « faire peur » et une approximation du critère « peur » par la valeur du vecteur code ou par la moyenne est donc complètement inappropriée.

Pour résoudre le problème des mauvaises approximations dues à l'éventuelle hétérogénéité des classes, nous avons envisagé une autre méthode d'estimation qui consiste à attribuer aux propriétés manquantes de l'individu ajouté celles de l'individu le plus proche dans la classe d'affectation. Bien que cette méthode semble être théoriquement convaincante, un autre problème lié à la présentation des individus a fait que les résultats obtenus ne sont pas tout à fait satisfaisants. En effet, si on prend un individu quelconque de l'ensemble de données, il est représenté par un vecteur de la forme $(v_1, v_2, ..., v_p)$ où p est le nombre de modalités et les composantes $v_1, v_2, ..., v_p$ sont comprises entre 0 et 1. Chaque valeur v_i représente en effet la fréquence des enfants qui ont choisi la modalité i pour l'individu.

Les valeurs des composantes présentes dans le vecteur représentatif d'un nouvel individu à classer sont égales à 0 ou à 1 étant donné qu'elles correspondent à un choix donné des propriétés de l'individu : la composante v_i du nouvel individu est égale à 1 s'il vérifie la propriété i, et à 0 sinon. La présentation d'un nouvel individu à classer sous la forme d'un vecteur binaire pose en fait le problème de l'exploration de l'espace des solutions qui devient très limitée. En effet, chaque nouvel individu présenté représente le sommet d'un hypercube ; les individus se situant aux centres des classes ne peuvent donc pas être retrouvés par une stratégie du plus proche voisin. Il est donc nécessaire d'explorer plus efficacement l'espace des données pour contourner le problème.

Pour ce faire, nous avons testé la stratégie suivante : au lieu d'estimer les propriétés manquantes du nouvel individu par celles de l'individu le plus proche dans la classe d'affectation, on sélectionne les individus de la classe les plus proches et on calcule ensuite leur vecteur moyenne. Ce dernier servira alors à estimer les propriétés manquantes du nouvel individu. La difficulté consiste à déterminer le

« bon » nombre des individus les plus proches à considérer : faut-il prendre la moitié de la classe, le tiers, le quart, etc. et existe-t-il un nombre pertinent ?

Pour tenter de répondre à cette question, nous avons opté pour la Classification Ascendante Hiérarchique afin d'effectuer une répartition de la classe d'affectation en plusieurs sous-classes homogènes à faible variance. Le nombre de sous-groupes formés est déterminé automatiquement de façon à respecter un *seuil* de densité de regroupement fixé a priori. Parmi les différentes sous-classes obtenues, on garde celle qui contient les individus les plus proches de l'individu ajouté.

L'application de la CAH nous permet d'avoir une zone de forte densité proche du nouvel individu : on calcule alors le vecteur centre de gravité de cette sous-classe et on l'utilise dans l'approximation des propriétés manquantes. Nous pouvons également déterminer l'individu le plus « proche »[2] du nouvel individu classé : il est l'individu de la sous-classe qui est le plus proche du centre de gravité.

Dans l'expérimentation menée avec le *mulet*, il s'agit de savoir si les propriétés que l'on a perçues à partir de son image permettent de savoir si l'animal « fait peur ». L'utilisation de la CAH comme décrit plus-haut nous permet de déterminer la sous-classe la plus proche de l'animal ajouté. Le dendrogramme de la figure 3.5 page 44 donne l'arbre hiérarchique résultant de la classification. La sous-classe la plus proche du *mulet* est alors celle qui est constituée par *âne, poney, vache, cheval*. On estime donc la propriété manquante du nouvel animal par celle du centre de gravité de cette sous-classe. Cette stratégie conduit à estimer que le *mulet* a $16,67\%$ de chance de faire peur aux enfants et donc 83.33% de ne pas leur faire peur.

Pour valider l'approche, des expérimentations analogues ont été menées sur les aliments, un autre domaine important du lexique Bassano et al. (2005) des enfants de 5 ans. La figure 3.7 donne une carte 4×4 obtenue sur les mêmes principes que celle des animaux. L'aliment *emmental* a été ajouté à la carte avec la propriété manquante : « se consomme-t-il en apéritif ? ». Il a été affecté à une classe de produits laitiers où la CAH (cf. figure 3.6) l'a placé dans une sous-classe contenant le *gruyère*, le *beurre* et le *fromage*, ce qui a permis de lui attribuer 80% de chances d'être consommé en apéritif.

2. la notion de proximité considérée ici n'est plus la proximité usuelle au sens de la distance de χ^2.

canard cygne moineau chouette pigeon pivert rouge-gorge chauve-souris perroquet tourterelle aigle	pingouin	escargot poisson-rouge requin serpent ver de terre baleine dauphin	lapin tortue	poule coq poussin paon autruche
		crapaud grenouille kangourou		agneau âne chèvre cochon mouton poney taureau vache chat hamster chien veau cheval **MULET**
coccinelle mouche papillon abeille moustique libellule	araignée pou fourmi cafard crabe	écureuil marmotte rat singe souris hérisson taupe raton laveur koala castor	chameau crocodile éléphant girafe léopard lion loup ours panda renard tigre zèbre rhinocéros hippopotame chamois biche	dragon dinosaure

FIGURE 3.4: Classement du nouvel animal dans la carte de Kohonen

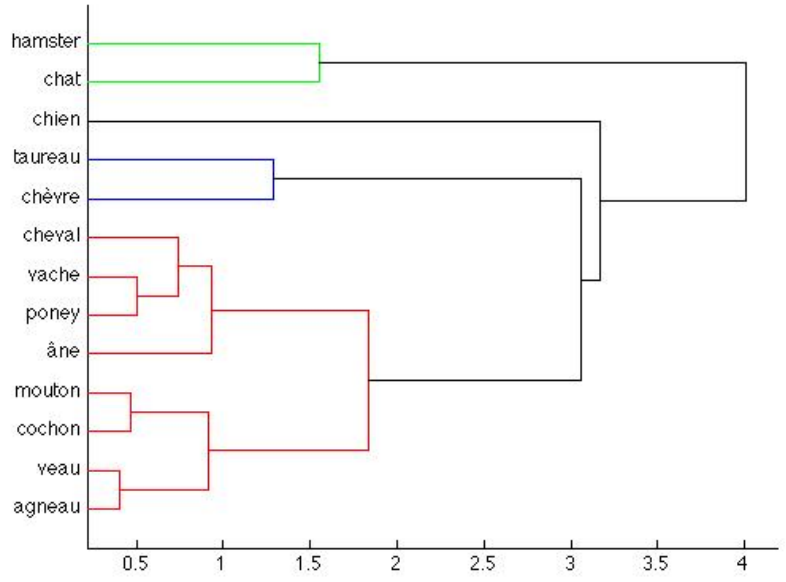


FIGURE 3.5: Résultat de la CAH appliquée à la classe d'affectation du nouvel animal

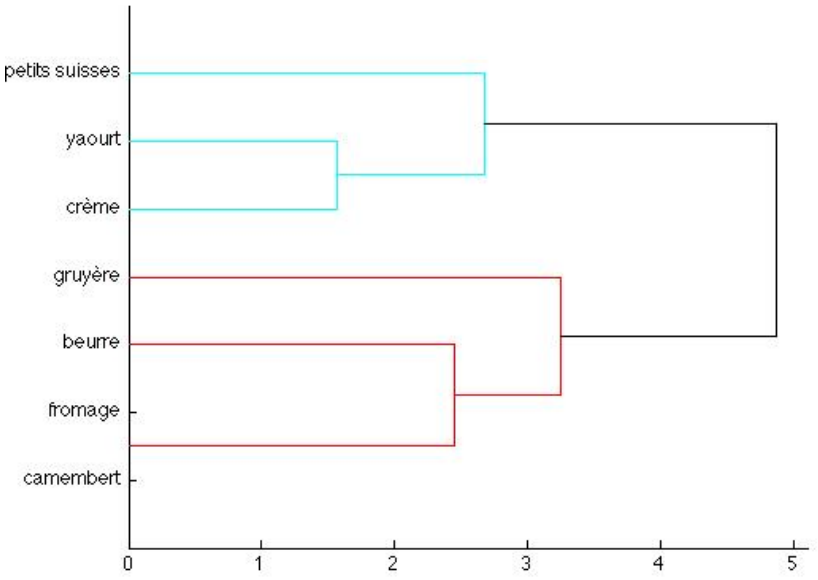


FIGURE 3.6: Résultat de la CAH appliquée à la classe d'affectation du nouvel aliment.

compote confiture miel nutella sucre vanille	biscuit bonbon brioche caramel céréales chewing-gum chocolat corn-flakes crêpe croissant flan, glace gâteau galette madeleine sucette	crevette moule poisson saumon	café coca-cola eau jus d'orange jus de fruit lait sirop thé vin
abricot banane cerise fraise framboise grenade groseille kiwi, melon mandarine myrtille orange pêche, poire pomme raisin			sauce soupe thon
carotte chou-fleur maïs oignon pomme de terre potiron radis	citron tomate	mayonnaise moutarde poivre sel	baguette cacahuète noisette noix pain tartine
avocat concombre cornichon cresson épinards haricot laitue menthe olive persil petits-pois salade	boudin jambon lardon oeuf poulet saucisse saucisson steak haché viande	chips frites hamburger omelette pâtes pizza purée riz sandwich semoule spaghettis	beurre camembert crème fromage gruyère petits-suisses yaourt EMMENTAL

FIGURE 3.7: Carte de Kohonen 4×4 des aliments du lexique.

3.5 Bilan et conclusion

La modélisation des connaissances enfantines est un problème complexe et ouvert. La proposition que nous avons développée repose sur l'utilisation des cartes auto-organisatrices de Kohonen ; la possibilité de visualiser les données, la facilité à les réorganiser et à traiter les données manquantes offrent des avantages qui en font un outil pratique et pertinent, conforme aux contraintes de clarté, de souplesse et de capacité de remise à jour que nous avions imposées.

À ce propos, Farida et moi avons la quasi-certitude que le temps a manqué pour mener cette recherche à son terme. Notre intention est de reprendre ce travail pour conforter ses résultats et le mener plus avant.

Chapitre 4

Étude de méthodologies d'extraction automatique de relations sémantiques

Ce chapitre décrit l'expérimentation de quelques approches destinées à l'extraction automatique de relations sémantiques en corpus : patrons sémantiques ontologiques, grammaires de segments, classification et détection du lien sémantique qui relie un fragment de texte entre parenthèses avec son contexte. Ces travaux ont été menés lors de l'encadrement de la thèse d'Ismaïl El Maarouf, un étudiant issu d'un master de linguistique (soutenue fin 2011) [El Maarouf (2011)].

4.1 Patrons sémantiques ontologiques : corpus de contes de fées et EmoLogus

Aux milieux des années 50, les linguistes tels que Harris (1954) et Firth (1957) ont avancé l'idée que *"You shall know a word by the company it keeps"* (« on peut connaître un mot à partir de ses fréquentations »). L'analyse « collocationnelle » repose sur ce principe : elle consiste à étudier l'environnement textuel d'un mot pour caractériser son sens. Le même principe servira de fondement à la similarité distributionnelle qui fera l'objet des expérimentations décrites dans le chapitre 5.

L'analyse collocationnelle d'un mot se fait en plusieurs étapes :

1. définition d'une fenêtre de taille arbitraire autour d'un mot-cible,
2. extraction des collocats,
3. tri des collocats en fonction d'un indice de pertinence (Z-score ou information mutuelle par exemple).

Les associations obtenues sont ensuite évaluées du point de vue de leur statut sémantique [Church and Hanks (1996)]. Telle quelle, la méthode se heurte à plusieurs difficultés majeures : outre le fait que la proximité entre mots peut n'être que fortuite, la multiplicité des collocations rend difficile leur interprétation et l'émergence de patrons.

4.1.1 Approche CPA et patrons de verbes

L'approche CPA (Corpus Pattern Analysis) a été proposée par Patrick Hanks [Hanks (2008a)]. Inspirée par le Lexique Génératif de Pustejovsky [Pustejovsky (1998)] et la sémantique des préférences de Wilks Wilks (1975), elle a pour objectif de construire un dictionnaire de patrons des principaux verbes de la langue anglaise « *all the normal patterns for all the normal verbs in English* », en fonction des catégories sémantiques de leurs arguments syntaxiques. La méthode repose sur le principe suivant lequel chaque patron est associé à un sens particulier du verbe et que l'étude de ces patrons permet en outre d'ordonner les sens des patrons en fonction de leur fréquence observée en corpus. Les catégories sémantiques sont définies par une ontologie (Brandeis Semantic Ontology) surfacique de *types sémantiques*. Celle-ci est structurée et hiérarchisée suivant les observations faites en corpus plutôt que sur une organisation aristotélicienne [1]. La figure 4.1 en montre quelques éléments.

L'exemple des patrons du verbe *to irritate* est donné par Hanks (2008b) :

irritate

PATTERN 1 (90%) :	[[Anything]] irritate [[Human]]
IMPLICATURE :	[[Anything]] causes [[Human]] *to feel mildly annoyed.*
PATTERN 2 (8%) :	[[Stuff]] irritate [[Body Part]]
IMPLICATURE :	[[Stuff]] causes [[Body Part]] *to become inflamed and somewhat painful.*

1. http ://www.pdev.org.uk

* **Entity**
  + Abstract_Entity
    - *Concept*
    - *Information_Source*
    - *Numerical_Value*
    - *Psych*
    - *Time_Period*
    - *...*
  + Energy
  + Physical_Object
    - *Animate*
    - *Inanimate*
      × *Artifact*
      × *Light_Source*
      × *...*
    - *Plant*
  + Particle
  + Self
* **Eventuality**
  + Event
  + State_of_Affairs
* **Group**
  + Human_Group
  + Vehicle_Group
  + Animal_Group
  + Physical_Object_Group
* **Part**
  + Language_Part
  + Music_Part
  + Physical_Object_Part
  + Speech_Act_Part
  + ...
* **Property**
  + Cognitive_State
  + Role
  + Visible_Feature
  + ...

FIGURE 4.1: Quelques éléments de l'ontologie BSO.

La méthode a été appliquée pour aider au développement d'EMOLOGUS dans le projet ANR Emotirob (cf. page 22).

Faute de disposer d'un corpus adapté à la tâche, nous avons utilisé un corpus de 139 contes de fées en langue française extraits du Web ; certains écrits par des enfants et d'autres par des adultes. La table 4.1 précise les caratéristiques du corpus concernant ses auteurs. On peut remarquer que si ce sont les enfants qui ont été les auteurs de la plus de la moitié des contes, leurs textes sont manifestement

Type auteur	Nb de mots	en %	Nb de contes	en%
Conteur moderne adulte	63 217	39%	24	17%
Enfant	53 109	34%	70	51%
Inconnu	34 314	21%	37	27%
Conteur classique	9 900	6%	7	5%
Total	**160 540**		**138**	

TABLE 4.1: Données concernant les auteurs du corpus de contes de fées : taille du corpus et nombre de textes.

relativement courts puisqu'ils ne constituent que 34% du corpus. Par ailleurs, les corpus de textes écrits respectivement par les enfants et les conteurs modernes adultes sont de tailles comparables, ce qui rend possible la comparaison qui est développée ci-après (cf. 4.1.2).

Les verbes dont les patrons ont été étudiés sont 90 verbes courants présents dans le corpus du lexique enfantin développé par Bassano et al. (2005). Les patrons sémantiques ainsi définis ont été modélisés dans la connaissance sémantique du système EMOLOGUS, afin de contrôler l'association entre les concepts connus du système. Ces travaux ont nécessité un très gros travail d'annotation : un grand nombre d'expressions référentielles et 24688 occurrences ont été annotées [El Maarouf and Villaneau (2012a)].

4.1.2 Étude comparée des patrons de deux corpus

Ces travaux ont été prolongés par une étude comparée des patrons obtenus dans ce corpus, entre contes écrits par les adultes et par les enfants. Une deuxième étude a permis de comparer les patrons détectés dans le corpus de contes de fées avec les patrons de ces mêmes verbes dans un corpus de presse [El Maarouf et al. (2009)].

Ces études montrent clairement que les patrons trouvés dépendent très fortement du type du corpus analysé. Plus précisément, elles ne font pas apparaître de différences significatives entre les patrons des contes écrits par les adultes et ceux des contes écrits par les enfants. Par contre, les patrons d'un verbe donné ne sont pas les mêmes, suivant que ce verbe est utilisé dans un conte de fées ou dans un article de presse.

Pour faire apparaître les différences entre les corpus, nous avons défini une mesure de similarité entre les catégories sémantiques, sur la base de leur utilisation dans

les patrons verbaux. Ainsi, la similarité entre les deux catégories c_i et c_j est définie par :

$$s_{ij} = \frac{n_{ij}}{n}$$

où n_{ij} est le nombre de contextes verbaux partagés par c_i et c_j et où n est le nombre total de contextes verbaux pris en considération.

La classification hiérarchique de Ward a abouti à la construction des dendrogrammes présentés dans la figure 4.2. Ils font apparaître les différences entre les classifications obtenues pour les catégories sémantiques à partir de la mesure de similarité précédemment définie.

– Dans le corpus de contes de fées, les animaux, les plantes, les êtres imaginaires et les humains partagent une très stricte similarité. De fait, un processus qui peut être qualifié d'« humanisation » est appliqué aux êtres vivants et imaginaires qui se voient doter de capacités normalement attribuées aux humains.
– Dans le corpus de presse, des extensions sémantiques de même type sont appliquées aux organisations, concepts abstraits, etc.

Ainsi par exemple, dans les contes de fées, tous les êtres vivants *disent* et *décident*. Dans la presse les sujets des verbes *décider* ou *parler* sont essentiellement des êtres humains ou des organisations.

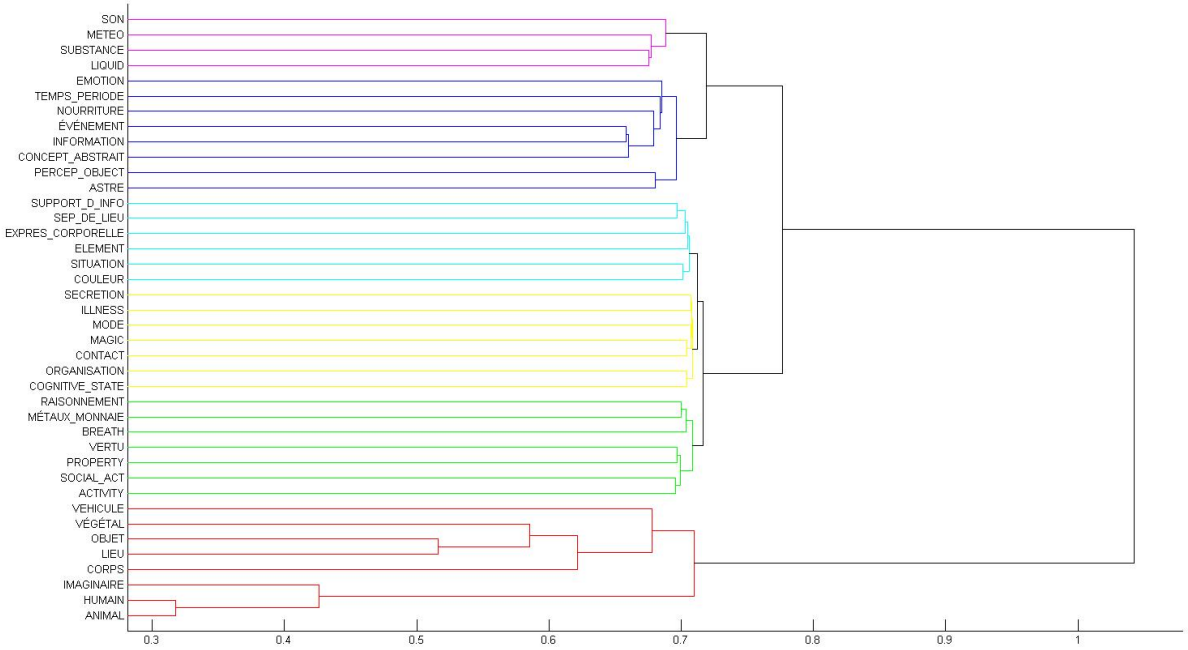
En revanche, dans le corpus de contes de fées, peu de différences ont été observées entre les productions des enfants et des adultes. Cependant, des « violations » de catégories sémantiques sont plus fréquemment observées dans les contes de fées écrits par les adultes, du fait que ces derniers utilisent des expressions idiomatiques telles que :

Les blessures qui déchirent vos coeurs.

Ces observations suggèrent que les enfants maîtrisent moins bien le style métaphorique ou idiomatique que ne le font les adultes.

Si l'intérêt des patrons sémantiques des verbes est évident, l'inconvénient de cette approche est le très gros travail d'annotations qu'elle requiert, sans compter le problème de la mise à jour des patrons ainsi collectés.

Corpus de contes de fées.



Corpus de presse.

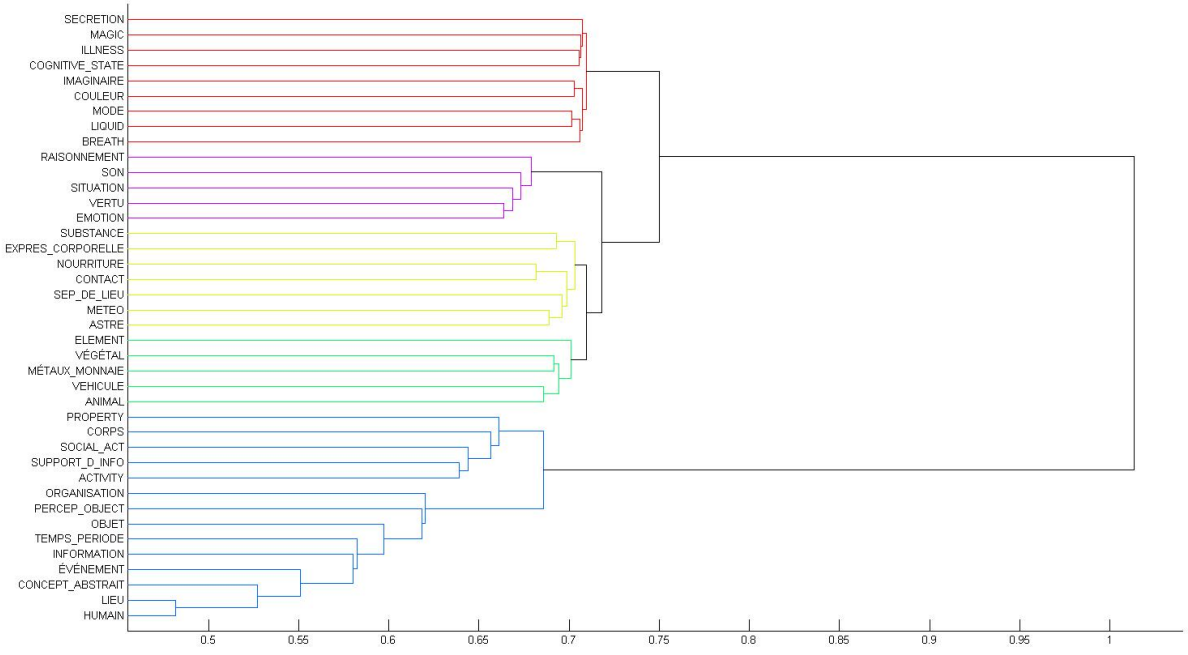


FIGURE 4.2: Les dendrogrammes des catégories sémantiques.

4.2 Grammaires de segments

L'une des pistes explorées dans la recherche de patrons sémantiques s'appuie très fortement sur une spécificité de la langue écrite : la ponctuation. En dehors de quelques travaux tels que ceux de Marcu (2000) concernant les relations de dicours ou ceux de Morin (1998) ou de Hearst (1992) concernant la détection de patrons lexico-syntaxiques, la ponctuation avait été relativement peu étudiée jusqu'alors, en linguistique comme en TAL. Pourtant, la virgule est la forme la plus fréquente dans le corpus que constituent les articles du journal leMonde de 2003 (6,81%) devant le mot *de* et le point qui, quant à eux, réalisent respectivement 4,37% et 4,04% des signes du corpus.

4.2.1 Segments et frontières

L'approche proposée consiste à proposer une structuration du texte plus fine que celle qui consiste à distinguer paragraphes et phrases, en s'appuyant sur des phénomènes discursifs de surface. Les frontières entre segments sont classées en trois types :

– les frontières dites *fortes* : point, points d'exclamation et d'interrogation, point virgule, deux points et points de suspension ;
– les frontières *semi-faibles* ou *englobantes* : parenthèses, guillemets et crochets ;
– les frontières dites *faibles* : virgules, conjonctions de subordination, etc.

Les relations entre chunks ont ensuite été étudiées suivant qu'elles concernaient deux chunks situées dans un même segment (relations intra-segments) ou dans deux segments différents (relations extra-segments).

4.2.2 Relations intra-segments et reconnaissance des entités nommées

La définition des segments et l'étude des relations inter-segments a été expérimentée pour aider à la reconnaissance des entités nommées (EN) [El Maarouf et al. (2011)]. Les travaux ont été menés dans la perspective d'adaptation d'un analyseur linguistique intégrant la détection d'EN (Ritel-nca) utilisé par le système de questions-réponses Ritel développé par le LIMSI [Rosset et al. (2008)]. Le système était destiné à corriger les résultats de l'analyseur à partir de patrons

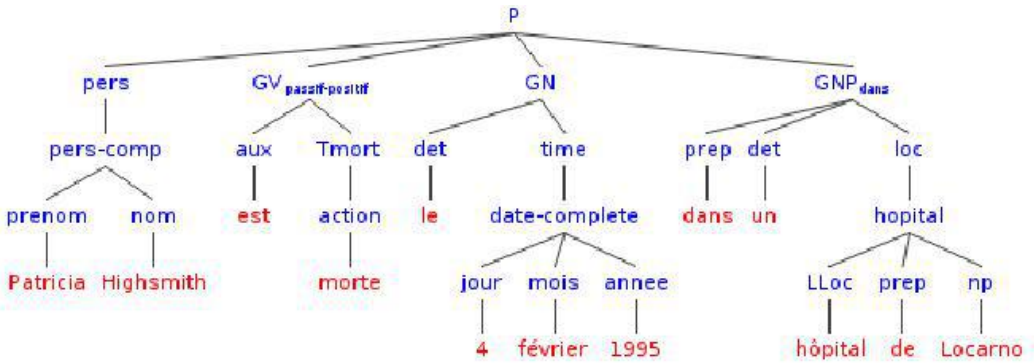


FIGURE 4.3: Exemple de sortie de Ritel-nca, après la phase complémentaire de chunking.

sémantiques extraits de corpus écrits ; les tâches de correction sont en effet considérées comme importantes. En effet, la reconnaissance des EN fait partie des tâches relativement bien maîtrisées si l'on en juge par les scores de réussite dans les campagnes d'évaluation. Cependant, la dégradation des scores des systèmes face à des types de textes ou des EN inconnus justifie la conception de modules de correction [Grishman (2010)].

4.2.2.1 Les principes de l'approche

Ritel-nca est un analyseur linguistique à base de règles dont les sorties sont présentées en structure arborescente. Le système a fait l'objet d'un développement particulièrement approfondi : au moment des travaux décrits dans ce chapitre, il permettait l'accès à des lexiques catégorisant plus d'un million de mots, dont une grande partie de noms propres, et près de 2000 règles y étaient implémentées. La taxonomie utilisée comprenait plus de 300 types, dont les EN classiques (Personne, Organisation et Lieu), affinés et structurés en sous-types et en composants. La F-mesure associée à la classification d'entités classiques était de 0,8 sur l'écrit et à hauteur de l'état de l'art pour les corpus oraux [Rosset et al. (2008)].

Pour faciliter l'analyse, une phase de chunking grammatical complémentaire a été ajoutée aux sorties de Ritel, qui intègre aux chunks les mots grammaticaux et précise la nature grammatical du chunk : un exemple est donné dans la figure 4.3 avec les trois noeuds GV, GN et GNP_{dans}.

Une phase de segmentation de surface isole des séquences de chunks en fonction d'indices de surface tels que la ponctuation, les segments ainsi constitués se

FIGURE 4.4: Niveaux de représentation des chunks.

Chunk	-	-	GNP_au	GNP_de
Entité	<_pers>	<_action>	<_subs>	<_loc>
Forme	Jacques Monod	rappelait	colloque	Caen

FIGURE 4.5: Exemples de patrons extraits en fonction du modèle choisi.

	E1	*E2*	*E3*	*E4*
CF	Jacques Monod	rappelait	GNP_au/colloque	GNP_de/Caen
CE	pers	action	GNP_au/subs	GNP_de/loc
CEM	pers	rappeler	GNP_au_colloque	GNP_de/loc

FIGURE 4.6: Un exemple : le segment « *Jacques Monod rappelait au colloque de Caen* ».

définissant par leurs frontières gauche et droite et le nombre de chunks qu'ils contiennent. L'étude des relations sémantiques entre chunks a été réduite aux chunks situés à l'intérieur d'un même segment, excluant de ce fait les 30% des segments qui sont composés d'un unique chunk.

Les patrons extraits au sein des segments peuvent s'appuyer sur les chunks, les entités ou les formes. Le premier tableau de la figure 4.6 indique les éléments correspondants pour chacun des quatre chunks du segment « *Jacques Monod rappelait au colloque de Caen* ». À partir de ces informations, les modèles testés sont : la combinaison des niveaux Chunk et Forme (modèle CF), des niveaux Chunk et Entité (modèle CE) et un modèle mixte (modèle CEM) combinant les niveaux Chunk et Entité, en substituant les entités « substantif », « action » et « adjectif » par les formes correspondantes, les verbes étant lemmatisés. L'existence de ce dernier modèle est motivée par l'hypothèse que ces classes contiennent régulièrement des informations sémantiques pertinentes qui seraient autrement masquées. Le second tableau de la figure 4.6 fait figurer les éléments extraits dans le segment donné en l'exemple, en fonction de chacun de ces trois modèles.

Alors que le modèle CEM cherche à optimiser les informations détenues par chaque élément, le modèle CE est le plus générique. Quant au modèle CF, il peut être plus précis en cas d'erreurs d'analyse des entités.

Le système s'appuie sur les patrons intra-segment observés dans le corpus de

développement. Pour chaque modèle, nous avons sélectionné les segments contenant une des entités classiques (« personne », « organisation » ou « lieu ») fournies par l'analyseur Ritel-nca, en excluant les segments de taille 1. Un patron correspond à un chunk identifié dans un segment contenant une EN, modélisé selon un niveau de représentation. À partir de ces données, le système calcule deux scores d'association d'un patron pour chaque classe : la probabilité de cooccurrence entre un chunk et une classe d'EN donnée (PROBA) et l'information mutuelle (IM). Ces scores permettent de prédire la classe d'EN la plus probable vis-à-vis d'un patron donné.

4.2.2.2 Évaluation et résultats

Nous avons utilisé un corpus qui correspond à l'année 2003 du journal LeMonde. Une partie de ce corpus a servi de corpus de développement et un quart du corpus a été réservé à l'évaluation. Pour cette dernière, 200 articles de presse ont été annotés afin d'obtenir plus de mille instances d'entités de chaque classe (plus exactement 1426 organisations, 1004 lieux et 1377 personnes). Un certain nombre d'EN n'ayant pas été détectées par Ritel-nca, l'évaluation n'a été réalisée que sur les EN détectées.

Les résultats montrent qu'aucun des différents modèles testés ne rivalise avec le modèle de référence Ritel-nca. En revanche, les patrons peuvent être utilisés en correction. Pour cela, nous avons sélectionné les patrons dont le score est sans appel (100%), il est alors possible de corriger les erreurs du modèle de référence, et, ce faisant, de juger de la pertinence linguistique des patrons correcteurs. On note globalement que la prise en compte des corrections permet d'améliorer les F-mesures de 5% pour les Personnes, et de 10% pour les Lieux et les Organisations.

Le problème majeur de l'approche réside dans la sélection des patrons de correction parmi la totalité des patrons générés par chaque modèle. L'intervention humaine semble indispensable pour permettre d'y remédier mais l'utilisation de méthodes de filtrage automatique ou partiellement automatique n'est pas exclue.

4.3 Étude d'une relation inter-segment particulière : les parenthétiques

À proprement parler, les travaux concernant les éléments de textes entre parenthèses sont à inscrire dans l'étude des relations inter-segments puisque, par définition, celles-ci sont des bornes de segments. Cependant, cette étude concerne également les relations intra-segments puisqu'on s'y intéresse également aux relations entre chunks du segment ainsi constitué [El Maarouf and Villaneau (2012b)]. Elle fait l'objet d'une section à part entière dans la mesure où nous avons tenté d'en faire une étude complète : constitution des corpus, proposition de classification et détection automatique des catégories ainsi définies.

Définies dans cette étude comme du texte entre parenthèses, les parenthétiques avaient été auparavant peu étudiées en TALN. Pourtant, elles sont omniprésentes et très fréquentes ; ainsi Bretonnel Cohen et al. (2010) rapporte avoir identifié 17 000 parenthétiques dans un corpus de 97 articles scientifiques d'environ 600 000 mots. Comparativement, les 136 000 articles de presse que nous avons étudiés en contenaient en moyenne quatre.

Si les parenthétiques avaient fait l'objet d'études particulières telles que l'extraction de paires de traduction [Cao et al. (2007)] ou l'étude des abbréviations [Okazaki et al. (2008)], il manquait une approche globale des relations syntaxiques et sémantiques qui les rattachent à leur contexte. Cette étude se proposait un double objectif.

– Le premier but était de proposer un nouveau schéma de classification des relations du texte entre les parenthétiques avec leur contexte, à savoir le texte dans lequel elles sont insérées.
– Le second but était de tester un système de reconnaissance automatique des relations précédemment définies.

4.3.1 La classification proposée

La classification des parenthétiques a été abordée sous l'angle de l'extraction de relations et divisée en trois sous-tâches : classification syntaxique, classification sémantique et reconnaissance des têtes interne et externe. L'étude d'un corpus de la presse française (Le Monde) a servi de support à cette première étude.

4.3.1.1 Reconnaissance des têtes

La tête interne d'une parenthétique est son élément informationnel majeur. Sa tête externe est l'élément du contexte auquel l'information entre parenthèses doit préférentiellement être rattachée. Une particularité de ces têtes est de couvrir à peu près toutes les classes grammaticales : texte, phrase, Entités Nommées, noms, verbes, adjectifs, etc. Trois catégories spécifiques ont dû être définies pour les têtes externes : *a, p* renvoient respectivement aux cas où la tête externe est le texte entier et la phrase dans sa globalité alors que *n* renvoie au cas où il est impossible de spécifier un rattachement particulier. Dans les exemples de la Table 4.2, les têtes sont en caractères gras. La Table 4.3 (en bas à gauche) donne des précisions statistiques sur la nature des têtes dans le corpus étudié.

4.3.1.2 Classification syntaxique

L'étude du corpus a permis d'identifier 11 classes syntaxiques, organisées suivant différents critères. Les exemples indiqués renvoient à la Table 4.2.

- parenthétique de nature propositionnelle (inter-clause : exemples 5 à 8) ou non (intra-clause, exemples 1 à 4) ,
- apposition/adjonction (exemples {1, 4, 5, 8}/ {2, 3, 6, 7}),
- présence ou absence de mots introductifs (soulignés dans les exemples),
- la parenthétique est (ou non) en coordination avec sa tête externe (exemples (3b), (7b)).

Ces critères permettent de définir 10 classes principales ; la Table 4.2 donne un exemple de chacune d'entre elles. Une onzième classe est définie comme une classe *autres*.

La classification sémantique ne traite que d'une classe syntaxique particulièrement fréquente puisque sa fréquence est de 82% dans le corpus étudié : les parenthétiques appositives non introduites et intra-propositionnelles (exemple (1) Table 4.2), à savoir celles pour lesquelles le lien sémantique avec le reste du texte est totalement implicite.

	Inter-clause
(1)	le **produit intérieur brut** (**PIB**).
(2)	il est (**très**) **réussi**.
(3a)	son **taux** directeur (à 2,5%).
(3b)	elle a connu la **liberté** (et les **pressions**).
(4)	La cérémonie a **lieu** mercredi (cf. **page** 15)
	Intra-clause
(5)	elle est **partie** (Gustave **avait** 6 ans).
(6)	elle est **partie** ce jour-là (Gustave **ayant** 6 ans).
(7a)	elle est **partie** (alors que Gustave **avait** 6 ans).
(7b)	elle est **partie** (et Gustave **avait** 6 ans).
(8)	je ne **suis** pas (ici elle **baissa** la voix qui tremblait) de l'avis de sa majesté !

TABLE 4.2: Exemples pour la classification syntaxique.

4.3.1.3 Classification sémantique

Pour faire la classification des parenthétiques appositives, non introduites et non propositionnelles, dix-huit classes sémantiques ont été définies. Elles ont été organisées en quatre groupes différents.

Les têtes des parenthétiques sont soulignées dans les exemples indiqués.

1. (a) Le premier groupe *Co-reference (CoRef)*, correspond aux cas où les têtes externe et interne désignent la même entité en utilisant des noms différents. On y distingue les classes suivantes :
   i. **Abbreviation** : la parenthétique contient une abbréviation de sa tête externe qui correspond à sa forme pleine (cf. exemple (1) de la Table 4.2).
   ii. **Explicitation** : la parenthétique définit un acronyme (on est dans le cas inverse du précédent).
   iii. **Traduction** : la parenthétique contient une traduction de sa tête externe dans un autre langue. Exemple :

   et A Chjama naziunale (l'Appel national)
   iv. **Reformulation d'une entité (RefEnt)** : toute autre relation coréférentielle non couverte par les relations précédentes telle que le rôle tenu par un acteur dans un film. Exemple :

   le bouquin de Z (Le Grand Voyage).
   v. **Reformulation d'une valeur (RefVal)** : la parenthétique est la traduction de la valeur que désigne sa tête externe suivant une autre échelle de valeurs (par exemple dans une autre unité). Exemple :

L'opération a rapporté environ 2,3 milliards de dollars (50 pourcent du PIB afghan)

(b) Le second groupement de classes, *Categorisation (Cat)*, correspond a une relation asymétrique des têtes du type entité-catégorie.

i. **Type** : la parenthétique précise la catégorie de l'entité désignée par sa tête externe (qui en est un hyponyme). Exemple :

l' obligation faite aux sociétés des segments NextPrime (secteurs technologiques)

ii. **Instanciation** : la relation entre les têtes est inverse de la celle qui est définie dans la classe précédente. La parenthétique correspond à un ou plusieurs hyponymes de la tête externe. Par exemple :

*les **pays** du golfe (Iran, Irak, etc.).*

iii. **Précision de valeur (ValPrec)** : la parenthétique précise la valeur de sa tête externe, qui désigne un élément quantifiable (chute, croissance, etc.) Par exemple :

le vote masculin reste prépondérant (16 pourcent des hommes, contre 12 pourcent de femmes).

(c) Le troisième groupement de classes regroupe les *Relations circonstancielles (Circ)*. La plupart correspondent à des relations sémantiques classiquement définies en extraction d'informations [ACE (2008)].

i. **Source de production (PS)** : la parenthétique et sa tête externe sont dans une relation œuvre/info production ou œuvre/auteur. Exemples :

dans Beautés du diable (Ed . Arthaud , 198 p. , 45 euros)

Le Lit de la vierge (Philippe Garrel) , La Dernière Femme (Marco Ferreri)

ii. **Affiliation** : la parenthétique précise l'organisation à laquelle appartient sa tête externe. Par exemple :

*(10) le **maire** (PS) de Toulouse.*

Cette classe a fait l'objet d'une définition spécifique à cause de sa fréquence.

iii. **Ancrage spatial (SA)** : la parenthétique précise la localisation géographique de sa tête.

iv. **Ancrage temporel (TA)** : la parenthétique donne des indications temporelle (de n'importe quelle entité) ; la classe se subdivise en *Date* et *Période*.

v. **Valeur d'un argument (ArgVal)** : la parenthétique précise la valeur de sa tête externe ; un exemple typique est celui de l'âge.

(d) La quatrième classe correspond à une *Référence (Ref)*. La parenthétique donne des références ou une indexation.

i. **Référence Intertextuelle (IR)** : il s'agit d'une référence à un journal, un média, etc. de sa tête externe qui en est une citation.

ii. **Référence paratextuelle (PR)** : la parenthétique donne une référence dans le texte lui-même (figure, note de bas de page, etc.)

iii. **Coordonnées** : la parenthétique donne des coordonnées de l'entité qui correspond à sa tête externe (numéro de téléphone, adresse postale, etc.).

iv. **Indexation** : la parenthétique donne des références qui correspondent à des indexations du document.

Des conventions d'annotation ont été élaborées pour permettre l'annotation complète d'un corpus de 1000 parenthétiques. La Table 4.3 donne une description du corpus en termes de classes et la Table 4.4 précise l'accord inter-annotateurs.

Syntactic Class	Frequency
Intra App NI	801
Intra Adj IN Not-Coord	60
Inter App NI	27
Truncation	25
Intra Adj NI	21
Inter Adj IN NotCoord	22
Intra Adj IN Coord	21
Inter Adj NI	1
Total	978

Semantic Class	Frequency
NULL	177
CoRef-Abbreviation	150
Sit-SA	87
Cat-Instantiation	78
Sit-ArgVal	72
Sit-Affiliation	72
Ref-IR	55
CoRef-EntRef	49
Other	43
Sit-PS	43
Cat-ValPrec	28
CoRef-ValRef	27
Cat-Type	25
CoRef-Translation	22
Sit-TA-Date	21
Ref-PR	9
CoRef-Explanation	9
Sit-TA-Period	7
Ref-Coordinates	4
Total	978

Head Class	Frequency
ID	869
p	62
n	25
a	22
Total	978

TABLE 4.3: Fréquences des classes dans le corpus.

Bien que, dans la classe des parenthétiques sémantiquement classées, la relation sémantique soit totalement implicite, le bon accord inter-annotateur montre, s'il

en était besoin, que le lecteur décode en général sans ambiguïté la nature de l'information qui lui est donnée entre parenthèses.

La robustesse de la classification proposée a été testée avec succès sur d'autres types de corpus : corpus encyclopédiques, littéraires, juridiques et scientifiques.

4.3.2 Classification automatique des parenthétiques

Le système proposé comme baseline combine les CRF (pour la détection des candidats) et les SVM (pour la classification), pour chaque tâche indépendamment et toutes tâches confondues. L'évaluation de ce système (Table 4.5) a permis tout d'abord d'observer que les ensembles de variables (formes, étiquettes morpho-syntaxiques, Entités Nommées, etc.) avaient un impact qui variait en fonction de la tâche : les étiquettes morpho-syntaxiques (T) sont par exemple les plus utiles à la classification syntaxique. De plus, la détection des candidats est une tâche cruciale, étant donné que le nombre de couples candidats aux frontières correctement délimitées est responsable d'une chute de la F-mesure globale du système (0,674, indépendamment, 0,518 toutes tâches confondues). Ces résultats sont conformes à ceux obtenus par Zhou et al. (2005) en Extraction de Relation à grand nombre de classes.

Si les scores du système automatique de détection sont corrects pour ce qui est de la nature syntaxique des parenthétiques, il n'en va pas de même pour ce qui est de la nature sémantique de leur lien avec leur contexte. Même si les travaux qui ont été menés n'ont pas exploré cette piste, il semble évident qu'une prise en compte du domaine auquel appartient le texte aurait permis d'améliorer assez largement les résultats.

Tâche	# accord.	# désaccord	Total	Kappa
Syntaxe	109	5	114	0.89
Semantique	88	13	101	0.79
Têtes	103	11	114	/

TABLE 4.4: Accords inter-annotateurs.

Feature	Pre-detection	Exact-Rec.	Soft-Rec	Syntax	Semantics
F	**0,965**	0,426	0,680	0,861	0,512
C	0,914	**0,499**	0,705	0,859	**0,637**
T	0,955	0,470	0,714	**0,908**	0,582
Ab	0,888	0,318	0,642	0,818	0,312
Pre-detection	-	0,349	**0,719**	-	-
Size	0,886	-	-	0,796	0,286
All	0,963	0,674	0,774	0,902	0,716
Baseline	0,888	0,3	0,649	0,818	0,182

TABLE 4.5: Résultats obtenus par le système en fonction des ensembles de variables utilisés. *(Independent task results on each feature set.)*

4.4 Conclusion

Les différents travaux présentés explorent quelques pistes de détection des relations sémantiques et font apparaître la multiplicité et la complexité de la tâche. La dernière expérimentation concernant les parenthétiques est révélatrice à bien des égards de la façon dont le lecteur humain interprète les relations sémantiques qui lient les éléments d'un texte.

– Pour comprendre la nature sémantique du lien entre deux groupes de mots, il est très souvent nécessaire de connaître ce dont le texte parle. Dans le cas des parenthétiques, la majorité d'entre elles sont appositives et, par conséquent, liées à leur contexte par un lien implicite. Le lecteur doit souvent faire des interprétations en fonction de connaissances « externes ».

  En ce sens, la pratique de l'annotation se révèle très instructive : lors de celle que nous avons faite des parenthétiques, nous avons souvent dû avoir recours au Web pour déterminer la catégorie sémantique des parenthétiques appositives lorsque le texte traitait d'un sujet peu connu.

– Ces études ont également montré combien leurs résultats dépendent directement du type des textes considérés. Par exemple, les natures syntaxique et sémantique des parenthétiques sont très dépendantes des corpus et des domaines auxquels ils appartiennent : scores dans les articles consacrés au sport, affiliation à un parti dans les domaines politiques, etc. De la même façon, les études de patrons comparés des verbes ont montré de notables différences entre les corpus de contes de fées et les articles de presse.

Ces constatations concernant la diversité des textes recouvrent des lieux communs : face à un texte qui traite d'un domaine un peu spécialisé qu'il ne connaît pas, le lecteur se sent souvent perdu ; on peut également citer la grande distance qui

séparent la langue des tweets de celle des textes de loi, même en faisant abstraction des sujets traités : vocabulaire, usage de la syntaxe, longueur des phrases, usage de la ponctuation, etc.

Parallèlement à l'utilisation de modèles statistiques qui permettent, sur de grands corpus, de trouver les usages les plus fréquents, une approche linguistique, et particulièrement celle de la linguistique de corpus, permet de mieux comprendre les phénomènes qui rendent la détection des liens sémantiques aussi complexe et aussi riche. Elle fait également douter de l'efficacité des approches génériques pour des tâches un peu spécifiques liées à des domaines particuliers.

Chapitre 5

Modèles vectoriels : similarité entre phrases et résumé multi-documents

Les expérimentations présentées dans ce chapitre sont liées aux travaux de thèse d'Hai Hieu Vu (soutenance janvier 2016). Leur objectif était la réalisation d'un système capable d'extraire les éléments les plus importants d'un domaine spécifique à partir d'un ensemble de documents. Le système devait être robuste et générique, utilisable pour des documents en langue française et en langue anglaise.

Concrètement, la plus grande part des expérimentations a été consacrée à l'évaluation de la similarité entre phrases, qui est apparue comme un prérequis indispensable à la méthode de résumé automatique que nous avions choisie [Vu et al. (2014, 2015)].

Au sens de l'application visée, évaluer la similarité entre deux phrases consiste à mesurer jusqu'à quel point ces phrases « parlent de la même chose » et relatent les mêmes faits ou actes. Par ailleurs, l'approche choisie repose sur la similarité distributionnelle selon laquelle des termes sémantiquement proches tendent à apparaître dans des contextes similaires qui suppose elle-même l'hypothèse distributionnelle déjà citée page 47 : *"You shall know a word by the company it keeps"* (on peut connaître un mot à partir de ses fréquentations).

Pour construire un système global de résumé automatique de textes robuste, générique et aisément portable, le choix a été fait de faire reposer le module qui calcule la similarité entre phrases sur le modèle vectoriel du Generalized Vector Space Model (GVSM) [Wong et al. (1985)] et la sémantique statistique qui suppose que les modèles statistiques de l'usage d'un mot peuvent être utilisés pour comprendre leur sens. Nous avons fait le choix d'utiliser l'encyclopédie Wikipédia comme ressource linguistique à cause de sa disponibilité dans de nombreuses langues et du grand nombre de domaines qu'elle couvre.

Par rapport aux travaux décrits dans les chapitres précédents, l'approche est donc résolument statistique. De plus, le choix même de Wikipédia correspond à l'objectif d'une représentation sémantique universaliste et neutre.

5.1 Vecteurs de termes - Similarité entre phrases

5.1.1 Construction des vecteurs de termes

En analyse distributionnelle, le modèle initial consiste à construire des matrices *termes*×*contextes* dont les éléments sont une mesure de co-occurrence. Cette représentation correspond à la représentation de chaque terme comme un point dans un espace de très grande dimension et la similarité entre deux termes y est mesurée comme une distance entre les deux points qui les représentent.

La définition de *contexte* d'un terme est extrêmement variable : elle peut se réduire aux quelques mots qui entourent le terme en question, ou bien s'étendre au document entier. Sahlgren (2006) consacre un chapitre de sa thèse à cette question et ses conclusions nous ont amenés à choisir comme contexte les concepts définis par chacun des documents présents dans Wikipédia.

Quel que soit le choix du contexte, les matrices *termes*×*contextes* sont creuses et de très grande dimension. Différentes techniques sont couramment utilisées pour réduire leur taille et obtenir une représentation des termes plus dense. La décomposition en valeurs singulières (SVD) est l'une des plus courantes. Une solution qui évite cette étape de réduction est mise en œuvre dans les réseaux de neurones ou le Random Indexing : elle consiste à construire directement une représentation des termes dans un espace de faible dimension. Un autre avantage

de ces méthodes est de permettre plus facilement l'ajout de nouveaux mots ou documents.

Notre choix s'est porté sur le Random Indexing [Sahlgren (2005)] : il repose sur une projection de l'espace de départ dans un espace de vecteurs index de dimension réduite presque orthogonaux, qui préserve approximativement les distances entre points. Les coefficients de la matrice correspondent au *tf-icf* introduit par Reed et al. (2006). Ce coefficient est une approximation du très courant *tf-idf*; par rapport à ce dernier, il offre l'avantage de coûts de calculs réduits, particulièrement en cas d'ajouts de documents. Il est défini par :

$$tf\text{-}icf_{ij} = \log(1 + f_{ij}) \times \log\left(\frac{N+1}{n_i+1}\right)$$

où f_{ij} est le nombre d'occurrences du i-ième terme dans le j-ième document, N le nombre total de documents (du corpus entier ou d'un sous-corpus statique) et n_i le nombre de documents où apparaît le terme d'indice i.

Le vecteur qui représente le terme d'indice i se définit comme : $v_i = \Sigma_{i=1}^{n} tf\text{-}icf_{ij}.c_j$ où c_j est le vecteur du j-ième contexte. Enfin, la similarité entre termes est définie comme le cosinus de leurs vecteurs respectifs.

Le calcul de la similarité entre phrases à partir des vecteurs de termes a donné lieu à plusieurs expérimentations. Nous avons d'abord mis en œuvre une méthode « classique », qui consiste à représenter une phrase par sommation des vecteurs des termes qui la composent [Chatterjee and Mohan (2007)]. Nous avons ensuite implémenté une deuxième approche, déclinée sous plusieurs formes, qui évite cette sommation de vecteurs et qui surclasse la première sur les corpus liés à la tâche visée (résumé multi-documents dans un domaine donné), comme le montrent les résultats présentés dans la section suivante (cf. section 5.1.4).

5.1.2 Similarité entre phrases par définition d'un vecteur sémantique de phrase

Dans cette approche, pour calculer la similarité entre deux phrases, chacune d'elles est d'abord représentée comme un vecteur sémantique.

On suppose que Wikipédia a une couverture des concepts et des mots suffisamment large pour contenir la plupart des termes sémantiquement significatifs utilisés dans

les phrases en question. Le vecteur sémantique d'une phrase se calcule en faisant la somme des vecteurs sémantiques des termes qui la composent, suivant la formule (5.1).

$$\vec{S} = \sum_{i=1}^{n} \overrightarrow{terme_i}. \tag{5.1}$$

Toutefois, cette mesure ne prend pas en considération le poids interne des mots dans le texte ou dans l'ensemble de textes d'où la phrase est extraite. L'hypothèse est que, si un mot est très fréquent dans les documents concernés, il convient de minimiser son importance au niveau de la phrase. Pour cela et conformément aux travaux de Neto *et al.* (2000 et 2002), nous utilisons la pondération par le *tf-isf* (term frequency × inverse sentence frequency). Le *tf* est ici le nombre d'occurrences du terme dans la phrase et l'*isf* est calculé d'après la proportion de phrases dans l'ensemble des documents qui contiennent le terme :

$$tf\text{-}isf_{is} = tf_{is} \times \log(\frac{|S|}{SF_i}) \tag{5.2}$$

où $|S|$ est le nombre de phrases et SF_i le nombre de phrases qui contiennent le terme d'indice i. Ainsi, l'importance d'un terme qui apparaît dans un grand nombre de phrases de l'ensemble des documents s'en trouve réduite.

Par ailleurs, les vecteurs sémantiques des termes peu fréquents sont essentiellement des vecteurs creux : en d'autres termes, ils contiennent principalement des coordonnées nulles. Conformément à Higgins and Burstein (2007), les vecteurs des mots rares peuvent être enrichis en utilisant le vecteur centroïde du texte défini suivant la formule suivante.

$$\overrightarrow{centroid} = \frac{1}{n} \sum_{i=1}^{n} \overrightarrow{terme_i}, \tag{5.3}$$

où n est le nombre de termes distincts dans le texte à calculer.

Introduire dans le calcul du vecteur sémantique d'une phrase son vecteur centroïde, augmente le poids des coordonnées des vecteurs des termes rares et réduit le biais introduit par la fréquence des termes généraux. Le vecteur sémantique d'une phrase est finalement calculé avec la formule (5.4).

$$\overrightarrow{S_i} = \sum_{j=1}^{n} tf\text{-}isf_{ij} * (\overrightarrow{terme_j} - \overrightarrow{centroid}), \tag{5.4}$$

où $\overrightarrow{term_j}$ est le vecteur du terme d'indice j et n le nombre de termes distincts dans la phrase d'indice i.

Autre normalisation pour le calcul du vecteur de phrase

Lors des premières expérimentations, nous avons analysé finement les mesures de similarité obtenues entre certains termes et groupements de termes pour mieux comprendre les particularités de la méthode. Dans la similarité entre groupement de termes, des dysfonctionnements s'observent lorsque se trouvent associés des termes qui diffèrent de par leur fréquence. Après avoir décrit le phénomène, nous proposons une modification dans le calcul des coordonnées des vecteurs de termes.

Wikipédia est une encyclopédie qui couvre un très grand nombre de domaines. De ce fait, les termes spécifiques à un domaine particulier n'apparaissent que dans un nombre restreint d'articles et leur coefficient *cf* est très faible. Or, leur rôle est essentiel pour évaluer la similarité entre deux phrases. À l'inverse, certains termes, que nous appelerons *mots généraux* y sont très fréquents. La table 5.1 donne quelques exemples de termes généraux et spécifiques pour la langue française, avec leur nombre d'occurrences dans Wikipédia, le pourcentage des articles dans lesquels ils apparaissent et la valeur de leur coefficient *icf*. Les poids *cf* des termes généraux sont très supérieurs à ceux des termes rares ou spécifiques. En effet, les mots généraux ont tendance à se trouver dans une grande proportion des articles de Wikipédia.

Terme	**cf**	**Couvert.**	**icf**	**Terme**	**cf**	**Couvert.**	**icf**
naître	298 963	29,60%	0,52	joli	7 331	0,72%	2,14
pouvoir	293 035	29,01%	0,53	NASA	3 528	0,35%	2,45
grand	263 987	24,14%	0,58	peste	4 917	0,49%	2,31
nouveau	235 462	23,31%	0,63	sida	1 524	0,15%	2,82

TABLE 5.1: Exemples de l'importance comparée des termes dans le Wikipédia français.

Il est par ailleurs intéressant de noter que les scores comparés de ces termes sont assez différents de ceux donnés par la base Lexique (`http://www.lexique.org/`). Par exemple, le rapport de fréquence entre *grand* et *joli* est d'environ 20 dans la base Lexique et il est beaucoup plus élevé dans Wikipédia. On remarque également la valeur du *cf* du lemme *naître*, sur-représenté dans Wikipédia. Dans le cas du lemme *joli*, on peut avancer l'hypothèse du style neutre de Wikipédia, les

consignes données aux auteurs étant d'éviter les jugements de valeurs. La sur-représentation du mot *naître* peut être expliquée par le contenu de l'encyclopédie et plus précisément, par ses très nombreuses biographies.

Lorsque l'on évalue la similarité entre groupements de termes où sont associés un terme très fréquent avec un terme spécifique, l'influence du terme le plus fréquent écrase celui du terme spécifique. Par exemple, les lemmes *robot* et *infection* ont respectivement des *cf* relativement faibles, respectivement égaux à 5930 et 3593. À ce titre, ils peuvent être considérés comme des mots spécifiques. Par ailleurs, leur score de similarité (calculé comme le cosinus de leurs vecteurs de terme) est très faible (peu différent de 0,007). Or, les groupements de termes *petit robot/petite infection* obtiennent, avec le calcul de similarité défini précédemment, un score peu différent de 0,89, une valeur très élevée, due à la prééminence du vecteur de termes *petit* sur les deux autres vecteurs de termes. On en conclut donc que, bien que l'*icf* ait considérablement réduit le poids des termes généraux, la réduction qu'il opère n'est pas suffisante : les poids des coordonnées des vecteurs termes généraux et des termes plus rares sont déséquilibrés, ceux associés aux termes fréquents étant plus importants que ceux associés aux termes moins fréquents.

L'objectif est donc de rééquilibrer le poids des termes très fréquents (mots généraux) par rapport à celui des termes plus rares, souvent spécifiques à un domaine donné, comparativement aux valeurs obtenues par le calcul classique du *tf-icf*. Pour ce, on introduit un paramètre $\alpha > 1$, destiné à renforcer le poids de l'*icf*, selon la formule (5.5).

$$tf\text{-}icf_{\alpha} = tf * icf^{\alpha}, \tag{5.5}$$

Un développement mathématique simple permet de se rendre compte que cette opération augmente le poids des termes rares dont la fréquence relative dans le corpus $\frac{n_i}{N}$ vérifie $\frac{n_i}{N} < 0,1 - \frac{0,9}{N} \simeq 0,1$, et cette pondération est d'autant plus importante que α est grand. Ce rééquilibrage entre termes rares et fréquents améliore les résultats des calculs de similarité entre groupements de termes ou entre phrases, comme le montrent les évaluations de la section 5.1.4.

5.1.3 Similarité entre phrases par optimisation des similarités entre termes

La somme de vecteurs de termes est une méthode éprouvée qui donne des résultats acceptables : les résultats décrits dans la section 5.1.4 en donnent une illustration. Cependant, il n'est pas simple de comprendre ce que représente une somme de plusieurs vecteurs sémantiques de termes et ce caractère de « boîte noire » laisse assez peu de place aux tentatives d'amélioration. Plusieurs expérimentations ont été menées sur le principe d'une similarité calculée en maximisant la somme des similarités entre les termes des deux énoncés suivant une formule proche de celle donnée par Mihalcea et al. (2006).

$$Sim(P1, P2) = \frac{1}{2}\left(\frac{\sum_{t_i \in P1} max_{t_j \in P2} sim(ti, tj)}{n1} + \frac{\sum_{t_j \in P2} max_{t_i \in P1} sim(ti, tj)}{n2}\right)$$

où n_1 est le nombre de termes de l'énoncé P_1 et n_2 est le nombre de termes de l'énoncé P_2.

Ainsi, chaque terme considéré comme sémantiquement signifiant dans chacun des énoncés est associé au terme de l'autre énoncé qui lui est sémantiquement le plus proche au sens de la similarité entre vecteurs de termes. La formule permet d'attribuer un score compris entre 0 (similarité nulle) et 1 (similarité totale). Cette mesure de similarité entre énoncés par optimisation des similarités entre les termes qui les composent a été testée sous diverses formes :

1. en tenant compte ou non de la nature syntaxique des termes (noms, verbes, adjectifs ou adverbes) comme le font Mihalcea *et al.* ;
2. en optimisant un alignement des termes entre les deux énoncés ;
3. en utilisant diverses formules de similarité entre vecteurs de termes ;
4. en prenant en compte l'ordre des mots (bigrammes) ou leur rattachement syntaxique (chunking).

Concernant le point (1.) et malgré ce que pouvaient laisser présager les tests de similarité entre termes en fonction de leurs natures syntaxiques, les différents essais concernant la prise en compte de la nature syntaxique des termes ont tous conduit à une détérioration très nette des résultats. Ces discriminations entre termes induisent par exemple une mauvaise similarité entre des expressions telles que « *le président japonais...* » et « *au Japon, le président...* ».

La motivation qui justifie les expérimentations du (2.) est que certains termes, par exemple les verbes courants, obtiennent des similarités relativement proches avec un grand nombre d'autres termes, toutes catégories syntaxiques confondues. Ainsi, la présence d'un verbe général dans un énoncé peut artificiellement augmenter son score de similarité avec les énoncés auxquels il est comparé. La prise en compte des catégories syntaxiques n'ayant abouti à un aucun résultat intéressant, nous avons tenté de modifier l'algorithme en cherchant un appariement entre termes des deux énoncés qui optimise la somme de leurs similarités, interdisant ainsi qu'un même terme puisse être utilisé plusieurs fois dans un tel alignement.

Dans les expérimentations menées sur les formules de calcul de similarités entre vecteurs (point 3.), le *cosinus* classique et le *jaccard* ont donné des résultats très similaires et nettement supérieurs aux autres formules de similarités testées. Sur l'ensemble des tests d'évaluation, ces deux mesures obtiennent des résultats qui ne diffèrent qu'à partir de la quatrième décimale : les résultats donnés section 5.1.4 sont donc à rapporter indifféremment à l'une ou l'autre.

Concernant le dernier point (point 4.), la prise en compte de similarités entre bigrammes de termes, calculées sous diverses formes, a également nettement dégradé les résultats initiaux. Pour prendre en compte d'autres éléments que ceux considérés dans une approche *sac de mots* (ordre des mots ou relations syntaxiques), le *chunking* est la seule piste qui a permis une légère amélioration des résultats. L'opération qui s'est révélée la plus efficace consiste à constituer des groupes nominaux et verbaux, qui permettent essentiellement de rattacher adjectifs et adverbes aux termes auxquels ils se rapportent. La similarité entre phrases repose ensuite sur l'optimisation des similarités entre ces groupements. Comme dans le cas des termes, la prise en compte de la nature du chunk (verbal ou nominal par exemple) ou un alignement optimal des chunks en fonction de leurs similarités respectives entraînent une très nette détérioration des performances.

Les résultats donnés dans la section suivante ont été obtenus en utilisant la formule brute donnée précédemment (WikiRI$_2$), WikiRI$_{ch}$ qui désigne la version qui introduit le chunking et WikiRI$_{2ch}$ obtenu en combinant les résultats des deux versions précédentes. Les informations syntaxiques utilisées pour ces expérimentations ont été fournies par le Stanford POS Tagger (`http://nlp.stanford.edu/software/tagger.shtml`) et le Stanford Parser (`http://nlp.stanford.edu/software/`

lex-parser.shtml) pour la langue anglaise et le Melt POS Tagger et le MaltParser pour la langue française (http://alpage.inria.fr/statgram/frdep/fr_stat_dep_malt.html).

5.1.4 Similarité entre phrases : évaluations

Les différentes versions de WikiRI ont été évaluées sur des corpus de langue anglaise du défi SemEval et sur des ressources en langue française que nous avons construites.

5.1.4.1 Évaluation pour l'anglais

Depuis 2012, la tâche STS de SemEval confronte les résultats de différents systèmes concernant la similarité entre paires de phrases, presque tous consacrés à la langue anglaise. La version 2014 de SemEval a cependant proposé une évaluation des systèmes sur des phrases en espagnol, à laquelle 9 équipes ont participé [Agirre et al. (2014)].

L'évaluation de WikiRI a été réalisée sur les données de la tâche 10 de **SemEval-2014** qui contient 6 corpus différents à évaluer pour l'anglais.

1. **Discussion de forum** (deft-forum) : 450 paires d'énoncés, très agrammaticaux, avec des mots souvent mal orthographiés.
2. **Discussion de l'actualité** (deft-news) : 300 paires de phrases, en général bien formées, sans majuscules.
3. **Titres de l'actualité** (headlines) : 750 paires de phrases souvent incomplètes.
4. **Descriptions d'images** (image) : 750 paires d'énoncés, bien orthographiés et sous la forme de phrases en général incomplètes.
5. **Définitions extraites de OntoNotes et de WordNet** (OnWN) : 750 paires d'énoncés, composé de phrases presque toujours incomplètes et utilisant des formes spécifiques (*the act of, the state of...*).
6. **Titres et commentaires de nouvelles sur tweeter** (tweet-news) : 750 paires de phrases, très souvent agrammaticales.

Corrélations	deft-for.	deft-news	hdln	images	OnWN	tw.-news
max Sem.	0,483	0,766	0,765	0,821	0,859	0,764
moy Sem.	0,368	0,637	0,604	0,694	0,697	0,616
WikiRI_1	**0,470**	0,638	**0,566**	**0,759**	0,740	0,689
WikiRI_2	0,430	**0,736**	0,562	0,752	**0,789**	0,720
WikiRI_{ch}	0,369	0,657	0,563	0,716	0,767	0,698
WikiRI_{2ch}	0,434	0,732	**0,567**	**0,758**	**0,788**	**0,722**

TABLE 5.2: Tableaux des résultats obtenus sur les données de Semeval 2014 : corrélations.

Les deux corpus les plus intéressants pour la tâche finale pour laquelle est conçu WikiRi sont *deft-news* et *tweet-news*.

En 2014, 15 équipes ont participé à cette évaluation et les résultats de 38 systèmes ont été comparés sur la base des coefficients de corrélation de Pearson avec les gold standard des corpus.

Le premier tableau 5.2 donne, en fonction des 6 corpus, les résultats du meilleur système, de la moyenne des systèmes, de WikiRI_1 (sommation des vecteurs de termes et utilisation d'une valeur de $\alpha = 3$ déterminée avec les corpus de SemEval-2012), de WikiRI_2 (similarité des phrases calculée par optimisation des similarités entre termes). L'avant-dernière ligne de ce tableau concerne les résultats obtenus avec la version de WikiRI qui utilise le chunking, et la dernière les résultats obtenus en combinant le chunking avec les résultats de WikiRI_2.

Dans l'ensemble les résultats obtenus par WikiRI sont encourageants. En effet, les tests ont été réalisés sans utilisation de règles particulières en rapport avec les types de corpus évalués, le but n'étant pas d'optimiser nos performances sur les différents corpus de SemEval, mais de valider une méthode et de pouvoir comparer les différentes versions de WikiRI. Or, être compétitif dans un challenge tel que SemEval demande que l'on effectue des traitements ou prétraitements particuliers en fonction des corpus considérés. Par exemple, dans le corpus *OnWN*, la simple élimination du chunk *"the act of"* dans le calcul de similarité fait grimper le score de plusieurs points, tant la proportion d'énoncés commençant par ce groupe de mots est importante.

Par ailleurs, on peut tirer plusieurs conclusions de la comparaison entre les différentes versions de WikiRI. Si WikiRI_1 obtient le meilleur score dans le corpus *deft-forum*, il est largement distancé par WikiRI_2 dans les deux corpus les plus

proches de la tâche visée que sont *deft-news* et *tweet-news*. De plus, si la version de WikiRI basée sur les similarités entre chunks n'obtient jamais les meilleurs scores, elle permet,combinée avec WikiRI_2, d'améliorer les résultats de cette version dans quatre des six corpus et elle obtient de meilleurs résultats que WikiRI_1 dans deux d'entre eux (*OnWN* et *deft-news*). Certes, une utilisation efficace d'informations syntaxiques se heurte à de nombreuses difficultés : énoncés incomplets ou agrammaticaux, mots inconnus, erreurs des parseurs, etc. En même temps, le fait de constater une amélioration des résultats lorsque l'on insère un certain nombre d'informations syntaxiques dans une approche par sacs de mots, prouve l'intérêt de l'approche. Par ailleurs, il va de soi que la méthode se doit d'être adaptée à la nature et au niveau de langue des corpus que l'on désire analyser : on ne peut pas espérer une analyse syntaxique un tant soit peu fiable d'un corpus tel que *deft-forum*.

5.1.4.2 Évaluation pour le français

Si SemEval2014 contient des données pour l'anglais et pour l'espagnol, il n'existe pas de corpus annoté en français actuellement pour la tâche qui nous intéresse. Créer un tel corpus est un travail long et difficile : tester toutes les paires d'un ensemble de n phrases devient rapidement impraticable de par la croissance quadratique du nombre de paires en fonction de n. Nous avons extrait du Web deux corpus de textes français dans deux domaines différents définis respectivement par les mots-clefs « Épidémies » et « Conquête spatiale ». Dans chacun de ces deux corpus, nous avons sélectionné un ensemble de soixante-dix phrases, dont la longueur varie de 10 à 65 mots. Dix d'entre elles ont été choisies comme phrases de référence : elles contiennent diverses informations importantes concernant les domaines testés. Chacune de ces dix phrases a été associée à six autres phrases choisies de façon à échantillonner les différentes configurations de similarité. La table 5.3 permet de comparer ces corpus à ceux de SemEval. Les pricipales différences sont la longueur moyenne des phrases, nettement plus longues que dans les corpus SemEval, et le nombre moyen de mots communs entre les phrases des paires testées qui est lui, nettement moins élevé.

Sept volontaires humains, âgés de 18 à 60 ans, ont été impliqués dans la tâche d'annotation dont trois experts et quatre candides. Ils ont évalué la similitude des

	Mots/Ph	ADV	ADJ	NC	**NP**	V	%Com.
Epid.	22,9	18,6%	10,6%	24,9%	**3,1%**	12%	9,7%
Conq. sp.	26,1	23%	6%	22,4%	**8,9%**	14%	6,8%

TABLE 5.3: Comparaison des corpus de tests *épidémies* et *conquête spatiale*.

paires de phrases sur une échelle de 0,0 à 4,0, selon les consignes indiquées dans la Table 5.4 et suivant la procédure d'annotation décrite dans Li et al. (2006).

4.0 : *Les phrases sont complètement équivalentes ;*
3.0 : *Les phrases sont globalement équivalentes, mais elles diffèrent par quelques détails ;*
2.0 : *Les phrases ne sont pas équivalentes, mais elles partagent certaines parties de l'information ;*
1.0 : *Les phrases ne sont pas équivalentes, mais elles traitent du même sujet ;*
0.0 : *Les phrases ne sont pas liées.*

TABLE 5.4: Les instructions d'annotation pour le choix du score de similarité entre phrases

Les participants ont travaillé indépendamment et sans contrainte de temps sur une application Web. Pour chaque phrase de référence choisie au hasard, ses phrases associées ont été aléatoirement et successivement présentées à l'annotateur. Ce dernier disposait d'un historique des scores de similarité qu'il avait déjà attribués et il était libre de les modifier à tout moment. Pour estimer l'accord inter-annotateurs, nous avons comparé les scores de chaque annotateur à la moyenne des scores calculée sur le reste du groupe. Les coefficients de corrélation ainsi obtenus sont présentés dans la table 5.5. Compris entre 0,8 et 0,941, ils indiquent que les évaluateurs humains sont largement d'accord sur les définitions utilisées dans l'échelle, même s'ils ont trouvé la tâche d'annotation particulièrement difficile.

Annotateurs	1	2	3	4	5	6	7
Corr. (c. spatiale)	0,872	0,869	0,844	**0,941**	0,886	0,815	0,855
σ (c. spatiale)	0,586	0,640	0,714	0,364	0,624	0,671	0,568
Corr. (épidémies)	0,862	0,904	0,903	0,931	0,846	0,846	**0,800**
σ (épidémies)	0,544	0,514	0,622	0,367	0,651	0,580	0,617

TABLE 5.5: Coefficients de corrélation et écarts-types entre les scores de chaque annotateur et la moyenne des scores des six autres.

Pour chacun des deux corpus, la version WikiRI_1 du système a été testée avec différentes valeurs du paramètres α. Les résultats sont donnés dans le premier tableau de la table 5.6. Alors que la valeur optimale du paramètre α reste stable entre les différents corpus en langue anglaise de SemEval, il n'en est pas de même entre les deux corpus de domaine en langue française, puisque le meilleur résultat est obtenu avec avec $\alpha = 2,25$ pour le corpus *épidémies* et $\alpha = 4,75$ pour le corpus *conquêtes spatiales*. Par ailleurs, l'introduction de ce paramètre s'avère très efficace voire nécessaire à l'obtention de résultats acceptables : les résultats obtenus pour $\alpha = 1$, qui correspondent à l'utilisation du *tf-icf* classique, sont largement inférieurs à ceux obtenus pour les valeurs optimales (0,648 contre 0,800 et 0,648 contre 0,849) et à ceux obtenus par toutes les versions de WikiRI_2.

WikiRI_1 α	1	2	2,25	2,5	3	4,5	4,75	5
Epidémies	0,648	0,794	**0,800**	0,796	0,775	0,701	0,687	0,672
Conq. spat.	0,648	0,750	0,761	0,771	0,792	0,848	**0,849**	0,847

	WikiRI_1	WikiRI_2	WikiRI_{ch}	WikiRI_{2ch}
épid.	0,800	**0,855**	0,776	0,848
conq. spatiale	0,849	0,854	0,749	**0,855**

TABLE 5.6: Tableaux des résultats pour les corpus français : WikiRI1 pour les deux corpus en langue française suivant différentes valeurs du paramètre α et résultats comparés des différentes versions de WikiRI.

Le second tableau permet de comparer les résultats de WikiRI_1 (obtenus avec le α optimal), WikiRI_2, WikiRI_{ch} et WikiRI_{2ch}. Comme pour la plupart des corpus de Semeval, les mesures de similarités faites en optimisant les similarités terme à terme obtiennent de meilleurs résultats que WikiRI_1, même en choisissant le meilleur α pour chacun des deux corpus. En revanche, l'utilisation du chunking n'améliore pas les résultats, déjà très élevés si on les compare à ceux obtenus avec les corpus de Semeval. Ces bons résultats sont peut-être attribuables à la moyenne nettement moins élevée du nombre de mots communs entre les phrases des paires testées, dont l'existence éventuelle permet de détecter plus facilement la similarité entre paires d'énoncés.

5.2 Résumé multi-documents

La tâche de résumé multi-documents étant celle pour laquelle WikiRI a été conçu, compléter les travaux précédents par quelques tests dans le domaine s'imposait

comme une évidence, malgré le manque de temps.

5.2.1 Principes généraux et approche choisie

Le résumé automatique de textes est devenu un champ de recherche important du Traitement Automatique des Langues ; il fait l'objet de plusieurs champs de recherche et des livres entiers lui ont été consacrés, y compris en français [Inderjeet (2001); Torres-Moreno (2011)]. Le résumé multi-documents est un champ particulier de la tâche du résumé automatique.

La redondance est l'un des problèmes majeurs de ce type de résumés et diverses méthodes ont été expérimentées pour pallier le problème ; l'une des plus connues est celle de la pertinence marginale maximale (MMR) [Goldstein and Carbonell (1998)]. Un deuxième problème connexe spécifique au résumé multi-documents est la gestion des informations contradictoires : il s'agit là d'une problème complexe qui n'a pas été abordé.

Les conférences NIST/DUC ont exploré de 2001 à 2007 les problèmes liés à la tâche du résumé multi-documents et proposé différents challenges liés à cette tâche spécifique. Par exemple, en 2006 et 2007, les résumés demandés devaient permettre de répondre à une question ou à un ensemble de questions posées sur le thème de chaque ensemble de documents à résumer [Dang (2006)]. Un autre challenge proposé était celui de la capacité des mises à jour d'un résumé, grâce à un second ensemble de textes contenant de nouvelles informations.

Par ailleurs, l'évaluation des systèmes dans DUC est très élaborée ; elle combine des évaluations manuelles avec des évaluations semi-automatiques telles que Pyramid [Nenkova and Passonneau (2004)], BE (Basic Elements) [Hovy et al. (2006)] et Rouge [Lin (2004)]. Ces expérimentations permettent entre autres de préciser la corrélation entre ces différentes techniques d'évaluation. Il semble que Rouge ait été beaucoup utilisé par les participants pour mettre au point leurs systèmes. Malgré les faiblesses de cette mesure [Sjöbergh (2007)], les expérimentations DUC confirment des observations de [Lin (2004)] selon lesquelles Rouge est plutôt bien corrélée avec les évaluations manuelles.

Les systèmes les mieux placés de DUC 2006 et de DUC 2007 utilisent généralement l'extraction de phrases, accompagnée de pré et post-traitements, par exemple en éliminant des portions de phrases jugées inutiles [Toutanova et al. (2007);

Pingali et al. (2007)]. La technique de résumé multi-documents que nous avons expérimentée repose sur cette méthode de l'extraction de phrases ; l'approche consiste à construire un graphe qui représente les textes à résumer : les sommets en sont les phrases et les arcs y sont étiquetés par l'indice de similarité entre phrases calculé par WikiRI. Le score des phrases est issu du calcul de l'algorithme DivRank, une variation de PageRank proposé par Mei et al. (2010). L'objectif est d'améliorer PageRank en permettant que le prestige laisse néanmoins place à la diversité ; au sens du résumé multi-documents, DivRank devrait donc permettre de choisir des phrases dont l'information peut être considérée comme importante car souvent répétée, tout en privilégiant une certaine diversité des informations. Ainsi, sur la base des données de la tâche 2 de DUC2004, Mei et al. (2010) rapporte pour DivRank des résultats supérieurs à ceux de PageRank, MMR (marginal maximum relevance), GH (Grasshopper), etc.

5.2.2 Expérimentations en langue française

Les expérimentations que nous avons menées en langue française utilisent le corpus qui a été élaboré lors du projet ANR RPM2 [de Loupy et al. (2010)]. Nous utilisons l'algorithme DivRank sur le graphe des phrases dont les arcs sont pondérés avec les similarités rendues par WikiRI. Nous comparons les résultats obtenus en utilisant les deux versions principales de WikiRI : WikiRI_1 et WikiRI_2.

La portion du corpus que nous avons utilisée comporte 200 documents, qui sont des articles de plusieurs journaux de la presse française, publiés entre janvier et septembre 2009. Plus précisément, elle est composée de 10 articles de presse dans chacun des 20 sujets retenus dans l'actualité du moment. Chacun de ces ensembles de documents contient en moyenne environ 5000 mots et 200 phrases. Les sujets choisis et la composition du corpus en termes de catégories grammaticales sont précisés dans de Loupy et al. (2010).

Nous avons évalué les résumés obtenus par notre système en utilisant pour le graphe initial les similarités calculées à partir de WikiRI_1 et WikiRI_2. Très classiquement, nous avons utilisé une version de ROUGE (Rouge-SU2) pour évaluer la qualité des résumés obtenus. Nous avons utilisé cette même mesure de ROUGE pour mesurer l'accord entre le résumé de chacun des quatre annotateurs avec les

	Sujet	$WikiRI_1$	$WikiRI_2$	Moy_annot
01	Ingrid Bétancourt	0,165	0,132	0,246
02	Caisse d'Epargne	0,190	0,167	0,267
03	Crise bancaire	0,127	0,122	0,212
04	Dalaï Lama	0,136	0,188	0,211
05	Fichier Edvige	0,230	0,382	0,277
06	JO de Pékin	0,102	0,156	0,192
07	Jérôme Kerviel	0,164	0,251	0,256
08	Lance Armstrong	0,209	0,180	0,298
09	La loi Leonetti	0,114	0,202	0,209
10	Le petit Mohamed	0,161	0,218	0,300
11	Obama président	0,136	0,153	0,179
12	Licenciement de PPDA	0,262	0,201	0,295
13	Le temple de Preah Vihear	0,172	0,254	0,248
14	Election au PS	0,151	0,191	0,226
15	Grossesse Rachida Dati	0,242	0,239	0,238
16	Rachida Dati et les magistrats	0,149	0,162	0,227
17	Réforme du lycée	0,126	0,211	0,203
18	Réforme de l'audiovisuel public	0,114	0,162	0,229
19	Relance de l'économie	0,111	0,193	0,235
20	Crise au Tibet	0,131	0,172	0,208
	Moyenne	0,1596	0,1968	0,2378

TABLE 5.7: Scores rendus par ROUGE-SU2 pour les résumés du corpus RPM2 à partir des similarités rendues par $WikiRI_1$ et $WikiRI_2$ et en utilisant DivRank.

résumés des trois autres. Les résultats sont donnés dans la table 5.7 pour chacun des vingt thèmes. La moyenne des accords entre annotateurs par thème est indiquée dans la dernière colonne.

Comme c'est souvent le cas pour ce type de tâches, on constate que les accords inter-annotateurs donnés par ROUGE-SU2 sont faibles : la moyenne générale des accords est de 0,2378. Cependant, il convient de remarquer que les résumés produits par les annotateurs ne sont pas de simples extractions du corpus mais des reformulations ; ce qui peut peut-être expliquer les faibles scores rendus par ROUGE qui calcule un score de similarité en comptabilisant les mots et bigrammes communs.

Les résultats obtenus en utilisant les similarités rendues par $WikiRI_2$ sont supérieurs à ceux obtenus en utilisant celles données par $WikiRI_1$ dans 14 des 20 thèmes du corpus. Par ailleurs, leur moyenne, calculée sur l'ensemble des 20 thèmes, montre également une nette supériorité de $WikiRI_2$ sur $WikiRI_1$. De fait, $WikiRI_2$ permet d'obtenir un score qui se situe à distance quasi égale de celui obtenu avec $WikiRI_1$ et de la moyenne de ceux des annotateurs.

Ces constations sont en accord avec celles observées lors des évaluations directes de WikiRI_1 et WikiRI_2 avec des corpus destinés à évaluer la similarité. Elles démontrent également l'importance que revêt la qualité des résultats de similarité pour mettre en œuvre une approche de résumé par extraction de phrases telle que celle que nous avons choisie.

5.2.3 Expérimentations en langue anglaise

Nous avons expérimenté l'algorithme DivRank avec les similarités rendues par WikiRI_1 sur les données de Duc 2007.

Les documents à résumer dans DUC 2007 sont des articles de journaux extraits des *Associated Press*, du *New York Times* (1998-2000) et de la *Xinhua News Agency* (1996-2000). Il y a 25 documents pour chacun des 45 thèmes retenus. La tâche consiste à faire un résumé d'au plus 250 mots par thème. Au-delà, le résumé proposé est automatiquement tronqué. Quatre résumés de référence ont été élaborés pour chacun des 45 thèmes mais, contrairement aux données du corpus RPM2, ces résumés ne sont pas disponibles. En revanche, DUC met à disposition un ROUGE-*BE package* pour une évaluation automatique de résumés produits.

Les expérimentations que nous avons effectuées ont utilisé WikiRI_1 avec les deux paramètres α et λ de l'algorithme du DivRank empiriquement fixés respectivement à 0,5 et 0,7.

Les résultats donnés par le paquet ROUGE fourni par DUC figurent dans la table 5.8 ; ils n'ont qu'un caractère indicatif car, comme expliqué auparavant, les résultats officiels du challenge DUC sont calculés par diverses méthodes qui associent des évaluations manuelles à des évaluations semi-automatiques. La table donne les scores respectifs du meilleur et du moins bon système, ainsi que le score médian. Le rang correspond à l'interclassement de notre système parmi les 32 systèmes participants.

Comme dans le challenge SemEval, le score du système est au-dessus de la médiane des systèmes participants. Aucun pré ou post-traitement n'a été implémenté malgré l'importance qu'ils peuvent présenter pour améliorer les résultats. On peut donc considérer que ces premiers résultats encouragent à poursuivre le travail commencé.

ht

version ROUGE	**-1**	**-2**	**-3**	**-4**	**-L**	**-W-1-2**	**SU4**
Max	0,427	0,113	0,038	0,023	0,393	0,149	0,166
Min	0,279	0,037	0,010	0,001	0,235	0,086	0,083
Mediane	0,397	0,089	0,027	0,012	0,365	0,138	0,146
WikiRI_1 p=0,85	0,399	0,092	0,030	0,016	0,366	0,139	0,146
rang (/33)	16	11	11	6	17	15	16

TABLE 5.8: Résultats du système sur les données DUC 2007.

Par ailleurs, nous n'avons pas pu, faute de temps, obtenir les résultats de WikiRI_2. L'algorithme mis en œuvre dans WikiRI_2 est en effet plus coûteux en temps que celui mis en œuvre dans WikiRI_1. Ce défaut s'était révélé peu sensible dans les précédents tests de similarité ou de résumé. Mais la taille des documents de tests proposées dans DUC 2007 est notablement plus importante que celle des données de tests du corpus RPM2 et cette différence suffit à rendre WikiRI_2 peu opérationnel avec les moyens de calcul dont nous disposions.

5.3 Conclusion

Les travaux entrepris sur le résumé multi-documents n'ont pas été entièrement finalisés, faute de temps. Cependant, ils offrent des perspectives intéressantes pour des travaux de recherche ultérieurs. Par ailleurs, les premiers résultats obtenus sont instructifs et positifs. La comparaison des résultats de WikiRI_1 et de WikiRI_2 sur le corpus RPM2 montrent l'importance de la tâche de similarité en sous-tâche de celle de résumé automatique lorsque l'on veut s'appuyer sur un algorithme de type PageRank. En outre, le niveau de résultat obtenu à partir de WikiRI_1 sur les données de DUC 2007 est tout à fait satisfaisant et, compte tenu de l'absence d'optimisation, ils prouvent la validité de l'approche pour réaliser un système de résumé multi-documents générique, léger et robuste.

Cependant, en même temps que l'on peut se satisfaire du fait que des stratégies telles que représenter une phrase par la somme de ses vecteurs de termes fonctionnent (du moins jusqu'à un certain point), on ne peut que s'interroger sur le « *pourquoi* » et le « *comment* ». Pratiquement, l'aspect « boîte noire » de l'approche rend difficile les améliorations de la méthode : les tentatives pour améliorer WikiRI_2 sont réprésentatives de cette difficulté. Si donc, les méthodes statistiques

basées sur la vectorisation des mots sont efficaces, le problème de les associer avec une analyse plus profonde des textes pour prendre en compte les relations sémantiques créées en associant les mots et groupes de mots est actuellement un problème qui reste ouvert.

Chapitre 6

Corpus, annotations et évaluations

6.1 Introduction

Les travaux en traitement automatique des langues utilisent des corpus pour détecter des patrons, entraîner les systèmes ou les évaluer. Par exemple, les travaux décrits dans les chapitres précédents ont tous utilisé des corpus représentatifs de la tâche. Certains de ces corpus, et particulièrement ceux destinés à une évaluation, avaient fait l'objet d'annotations de qualité diverse et surtout, plus ou moins adaptées au système en cours de construction. Presque toujours, la rareté de ces ressources a posé problème. Par ailleurs, le développement des corpus est déjà en soi une tâche délicate : pour la langue orale, elle est extrêment coûteuse en temps et en ressources du fait de la difficulté de la tâche de transcription ; si, pour l'écrit, le problème semble faussement simple, dans la pratique, il ne suffit pas de réunir un ensemble de textes. Étant donné l'importance du type des textes traités sur les méthodes à utiliser, il convient également de les choisir de telle sorte qu'ils soient représentatifs de la tâche que l'on prétend traiter, ce qui ne va pas nécessairement de soi.

Si les difficultés précédentes sont réelles, il n'en reste pas moins que le problème le plus difficile reste celui de l'annotation. Or, la création de corpus annotés est nécessaire au développement du TAL, et ce à double titre.

1. Pour les systèmes qui utilisent des méthodes d'apprentissage supervisées, les corpus annotés sont vitaux et toujours en quantité insuffisante aux yeux de ceux qui conçoivent ces systèmes.
2. Les annotations, sous des formes très variables, sont en général nécessaires à la création des Gold Standard sur lesquels reposent les évaluations. Leur importance est alors cruciale puisque ce sont elles qui conditionnent les sorties même des systèmes évalués. Celles-ci doivent en effet être conformes aux sorties attendues lors des défis et autres évaluations comparatives, faute de quoi les systèmes risquent d'être relégués dans les fonds de classement.

Il n'est déjà pas facile de voir si un corpus annoté que l'on crée ou que l'on utilise est représentatif de la tâche visée et si les annotations correspondent aux objectifs attendus. Veiller à la qualité même des annotations proposées est un problème encore plus difficile. Bien sûr, une condition nécessaire est la mise au point des règles d'annotation : dans Media, de nombreuses réunions entre les partenaires ont été nécessaires à la rédaction de ces règles. Pour les classes de parenthétiques et alors même que les choix se jouaient entre deux personnes, leur mise au point a également demandé de nombreuses heures de travail. Quant aux consignes d'annotation concernant la similarité entre phrases, elles ont donné lieu à plusieurs versions, sans que nous soyons réellement parvenus à une solution qui nous satisfasse totalement.

Cependant, le soin apporté à préciser les règles ne résoud pas tous les problèmes. D'une part, la tâche d'annotation en elle-même met en œuvre une interprétation du texte qui ne peut pas être entièrement objective. D'autre part, depuis quelques anneés, le TAL s'est attaqué à des domaines où les jugements des annotateurs sont très largement subjectifs, tels que par exemple la détection de l'émotion ou des opinions. La sensibilité propre des annotateurs entre alors largement en jeu, indépendamment des consignes qui leur sont données.

Les accords inter-annotateurs permettent d'évaluer la cohérence des annotations et par là-même, dans une certaine mesure, leur qualité. On peut en effet penser que si plusieurs annotateurs sont largement d'accord sur leurs choix, i.e. si leurs annotations sont à peu près identiques, l'annotation est fiable. Artstein and Poesio (2008) fait cependant remarquer que cette fiabilité des annotations ne garantit pas pour autant leur validité : deux annotateurs peuvent être d'accord parce qu'ils se sont trompés ensemble. Ainsi, on peut objecter que ce critère de qualité suppose que les

annotateurs ont les compétences nécessaires pour pouvoir appréhender les règles de l'annotation ou, dans le cas d'annotations en opinion ou émotion, qu'ils soient représentatifs du public auquel va s'adresser la tâche envisagée. Ce problème est laissé de côté dans ce chapitre où nous présentons les expérimentations que nous avons menées concernant les mesures d'accords inter-annotateurs. Ce travail a été réalisé avec Jean-Yves Antoine et Anaïs Lefeuvre [Antoine et al. (2014)]. Il s'est imposé à nous comme un sujet essentiel lorsque nous avons construit les corpus destinés à évaluer le système Emotirob et dû établir ce que devait être l'annotation de référence par rapport aux annotations recueillies [Antoine et al. (2011)]. D'une manière générale, ce travail concerne les annotations où il est demandé aux annotateurs de choisir entre plusieurs classes prédéterminées.

6.2 Mesures d'accords inter-annotateurs sur des annotations ordinales : expérimentations

Plusieurs mesures d'accords ont été proposées. Artstein and Poesio (2008) en propose une étude théorique comparée extrêmement précise et fouillée. L'étude présentée se veut essentiellement expérimentale et rapporte différentes conclusions à partir de corpus annotés pour des tâches où les accords sont toujours relativement faibles : co-référence, émotion et opinion.

6.2.1 Les mesures comparées

Les accords inter-annotateurs sont presque toujours mesurés suivant la formule :

$$\mu = \frac{A_o - A_e}{1 - A_e} = \frac{D_e - D_o}{D_e} = 1 - \frac{D_o}{D_e}$$

où

- A_o et D_o sont respectivement l'accord et le désaccord observés entre les annotations,
- A_e et D_e sont respectivement une estimation de l'accord et du désaccord attendus par la seule intervention du hasard.

Les différentes mesures diffèrent essentiellement par la façon dont A_e ou D_e sont estimés. Par ailleurs, pour les tâches qui nous intéressent, deux problèmes se posent :

– d'une part, on peut estimer que deux annotations sont plus ou moins éloignées l'une de l'autre et que, dans ce cas, il est utile de pondérer le désaccord plutôt que de l'estimer de façon binaire ;
– d'autre part, si la formule précédente peut être appliquée telle quelle dans le cas de deux annotateurs, elle doit être généralisée lorsque l'on s'intéresse à des annotations qui impliquent un plus grand nombre d'annotateurs.

Les mesures qui ont été comparées sont les suivantes :

– Dans le κ proposé par Cohen (1960), qui prévaut dans la plupart des études [Carletta (1996)], A_e est estimé à partir d'une distribution générale des classes spécifique pour chacun des annotateurs. Le κ a été généralisé au cas de plusieurs annotateurs par Davies and Fleiss (1982) d'une part et Cohen (1968) a lui-même proposé un κ pondéré d'autre part. Par contre, il n'existe pas de mesure κ qui permette de pondérer les classes dans une annotation multi-annotateurs. Cette mesure du κ multi-annotateurs est appelée μ-κ par la suite.
– Dans le π de Scott (1955), A_e est estimé sur la distribution générale des classes observée dans l'annotation, tous annotateurs confondus. La mesure a été généralisée à une annotation multi-annotateurs par Fleiss (1971). Elle sera désignée par μ-π.
– La mesure α de Krippendorff (1980) est basée sur les mesures de désaccord. D_e est estimé, comme l'est A_e dans la mesure π, sur une distribution générale des classes. La mesure prévoit de prendre en compte un nombre quelconque d'annotateurs et n'importe quelle distance définie entre les classes.

  Dans la suite de ce chapitre, nous désignerons par α_b (pour α binaire) la mesure α de Krippendorff utilisée sans pondération des classes et α_p (pour α pondéré) la mesure α dans laquelle nous aurons introduit une distance entre les classes considérées.

Choisir entre calculer A_e (ou D_e) à partir de la distribution générale des classes obtenue en considérant l'ensemble des annotateurs, ou à partir des distributions spécifiques à chacun des annotateurs, a donné lieu à de nombreux débats [Di Eugenio and Glass (2004); Krippendorff (2004); Craggs and Wood (2005); Artstein and Poesio (2008)] avec des arguments contradictoires : les uns prétendent que κ est la seule mesure efficace lorsque les distributions varient beaucoup d'un annotateur à l'autre, d'autres font valoir que le κ récompense les annotateurs qui sont en désaccord, etc. Nous voulions tenter d'y voir plus clair en comparant les résultats expérimentaux à partir de plusieurs corpus et sur des tâches différentes.

6.2.2 Expérimentations et résultats

Les premières expérimentations ont été menées en comparant les 4 mesures μ-κ μ-π, α_b et α_p sur trois corpus différents :

- Le premier corpus annoté a été développé pour tester le système Emotirob (cf. page 24). L'annotation consistait à attribuer un score à la valeur émotionnelle des différentes phrases d'un énoncé, hors-contexte ou en contexte. Les 25 annotateurs devaient choisir entre 5 valeurs de -2 à +2. Le signe du score correspond à la valence émotionnelle de l'énoncé (positive, négative ou neutre) tandis que la valeur absolue précise l'intensité de la valeur émotionnelle (nulle, moyennement intense ou très intense).
- Le deuxième corpus est composé de 183 phrases qui correspondent à des opinions exprimées sur des films. Les 25 annotateurs ont utilisé la même échelle de score ; le score permet ainsi de préciser la polarité et l'intensité de l'opinion ressentie.
- Le troisième corpus est un corpus de la langue orale contenant 488 000 unités lexicales, le corpus ANCOR [Muzerelle et al. (2014)]. Les annotations faites sont de trois types : marquage des entités, marquage des relations référentielles et marquage du type de relations référentielles prédéterminées.

  Les études concernant l'accord inter-annotateurs ne concernent que le choix du type des relations référentielles déjà répertoriées. Les 9 annotateurs avaient le choix entre 5 classes de relations de co-référence pour les 384 relations concernées.

Trois études ont été menées qui permettent de comparer le comportement des 4 mesures d'accord ; elles concernent la stabilité des scores selon le nombre de classes, l'influence du nombre d'annotateurs sur les résultats trouvés et leur caractère interchangeable.

6.2.2.1 Nombre de classes

L'influence du nombre de classes a été testée sur les deux premiers corpus. En effet, il était facile de réduire la classification en 5 classes induite par le score de -2 à +2 en une classification qui ne prenne en compte que la valence (corpus annoté en émotion) ou la polarité (corpus annoté en opinion) pour obtenir une classification en 3 classes : positive, négative ou neutre.

Emotion *(conte de fées)*				
Métrique	μ-κ	μ-π	α_b	α_p
3 classes	0,41	0,41	0,41	0,57
5 classes	0,29	0,29	0,29	0,57
Écart	*0,12*	*0,12*	*0,12*	*0,0*

Opinion *(critiques de films)*				
Métrique	$\mu - \kappa$	$\mu - \pi$	α_b	α_p
3 classes	0,58	0,58	0,58	0,75
5 classes	0,45	0,45	0,45	0,80
Écart	*0,13*	*0,13*	*0,13*	*0,05*

Co-Reference *(dialogue oral)*				
Métrique	μ-κ	μ-π	α_b	α_p
5-classes	0,69	0,69	0,69	n.s.

FIGURE 6.1: Résultats généraux et influence du nombre de classes.

La figure 6.1 présente les tableaux des résultats généraux obtenus sur les 3 corpus, ainsi que les scores obtenus avec une réduction à 3 classes pour les deux premiers. Ils permettent de faire plusieurs remarques.

– D'une manière générale, et comme il est courant sur ce type de tâches, on peut observer que les mesures d'accord sont très faibles. Ces faibles accords justifient que l'on fasse appel à un grand nombre d'annotateurs pour tenter de parvenir à une annotation quelque peu stable.
– Une autre observation est que les mesures d'accord non pondérées, donnent, à la troisième décimale près, les mêmes valeurs, ce qui relativise l'importance de la controverse concernant la façon d'estimer la distribution des classes.
– La pondération du désaccord (α_p de Krippendorff) permet d'obtenir d'un score plus élevé (0,57). Par ailleurs, il permet également d'obtenir un résultat nettement plus stable par rapport au nombre de classes considéré.

6.2.2.2 Nombre d'annotateurs

Une façon de stabiliser une annotation dite de référence est de faire appel à un grand nombre d'annotateurs. Malheureusement, plus les annotateurs sont nombreux, plus l'annotation devient coûteuse en temps comme en ressources ; sans compter le problème déjà évoqué de leur représentativité ou de leur compétence.

Nous voulions tester la fiabilité de l'annotation en fonction du nombre n d'annotateurs, à partir des N annotations dont nous disposions, soit $N = 25$ dans les

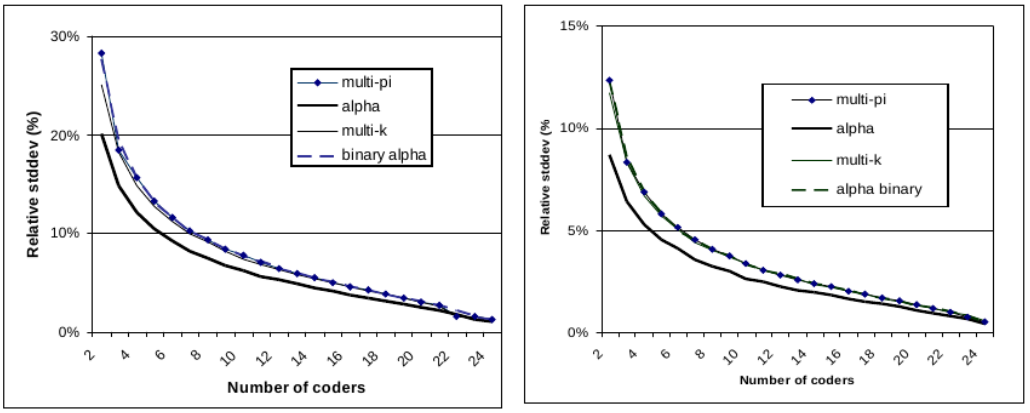


FIGURE 6.2: Moyenne des écarts-types (en pourcentage des valeurs moyennes de l'annotation) calculés à l'intérieur des groupes de n annotateurs.

corpus en émotion et en opinion et $N = 9$ pour le corpus de co-référence. Pour ce faire, nous avons considéré les sous-groupes de n annotateurs parmi N, calculé l'écart-type entre annotations à l'intérieur de ces sous-groupes et fait la moyenne de ces écarts-types.

Les résultats sont donnés dans la figure 6.2 pour les corpus en émotion et en opinion et pour une classification suivant 3 classes ; les courbes donnent la moyenne obtenue, rapportée à la valeur moyenne des annotations.

On peut constater qu'avec un faible nombre d'annotateurs, l'écart-type relatif est très important. Il décroît très vite mais il ne tombe au-dessous de 5% qu'avec 17 à 18 annotateurs pour le corpus en émotion et une quinzaine pour le corpus d'opinion. Par ailleurs, on peut observer les courbes identiques obtenues avec les mesures d'accord μ-κ, μ-π et α_b, alors que la prise en compte d'une distance entre classes permet d'atténuer les écarts d'appréciation entre annotateurs.

6.2.2.3 Interchangeabilité des annotateurs

Pour étudier jusqu'à quel point les groupes d'annotateurs étaient interchangeables, nous avons considéré toutes les combinaisons 10 codeurs parmi les 25 pour les deux premiers corpus et de 4 codeurs parmi 9 pour le dernier. Nous avons calculé la moyenne des scores de chacune d'entre elles et calculé l'écart-type entre ces différentes moyennes. La figure 6.3 donne les résultats obtenus avec les différentes mesures de score et en fonction du nombre de classes.

Corpus	**Émotion** ***(contes de fées)***				**Opinion** ***(critiques de films)***			
Métrique	μ-κ	μ-π	α_b	α_p	$\mu-\kappa$	$\mu-\pi$	α_b	α_p
3 classes	7.4%	7.7%	7.6%	6.2%	3.4%	3.3%	3.3%	2.6%
5 classes	9.0%	9.1%	9.1%	6.1%	4.0%	4.0%	4.1%	1.7%

Co-Référence ***(dialogue oral)***				
Métrique	μ-κ	μ-π	α_b	α_p
5-classes	4.6%	4.6%	4.6%	n.s.

FIGURE 6.3: Écart-type (rapporté à la moyenne des scores) entre les scores moyens de toutes les combinaisons de 10 annotateurs (parmi 25) sur les corpus **Émotion** et **Opinion** et 4 annotateurs (parmi 9) sur le corpus de **Co-Référence**.

Là encore, la pondération du désaccord autorisée par l'utilisation du α permet d'obtemir des scores plus optimiste pour ce qui est du caractère interchangeable des groupes d'annotateurs.

6.2.2.4 Stabilité de la référence

La dernière étude ne concerne pas les mesures d'accord mais la stabilité de la référence obtenue en fonction du nombre d'annotateurs. En considérant que la référence est obtenue à partir d'un vote majoritaire des annotateurs, nous avons évalué la moyenne du nombre de modifications de la référence entre groupes de n annotateurs indépendants. Ce calcul a été fait pour les corpus **Émotion** et **Opinion**, avec les deux options : classification en trois classes ou classification en cinq classes. Les résultats sont donnés sous forme de pourcentage et en fonction de n dans le tableau de la table 6.1.

Les résultats montrent à quel point la référence dépend du choix des annotateurs. Elle s'avère très instable, même avec seulement 3 classes, lorsque l'on passe d'un groupe de 5 annotateurs à un autre, puisqu'en moyenne, la référence est modifiée à plus de 20% dans le corpus **Émotion** et à 13,5% dans le corpus **Opinion**. Pour espérer avoir moins de 10% de modifications, il faut au moins un groupe de 13 annotateurs dans le corpus **Émotion** et 10 dans le corpus **Opinion**. Dès lors, on comprend l'importance du choix du nombre d'annotateurs quant à la stabilité et donc à la qualité de la référence proposée. L'enjeu est particulièrement important si cette référence doit servir à une évaluation et, en particulier, si elle sert de base à un classement des systèmes.

	Corpus Émotion		**Corpus Opinion**	
nb annot.	3-classes	5-classes	3-classes	5-classes
2	30,3	40,2	20,0	37,8
3	26,8	39,9	16,7	32,8
4	23,4	33,1	15,3	29,3
5	20,3	30,4	13,5	25,7
6	18,9	27,8	12,6	24,0
7	16,4	25,4	11,5	21,5
8	15,2	22,9	10,7	20,0
9	13,6	21,4	10,0	18,4
10	12,6	19,3	9,2	17,2
11	11,4	17,9	8,8	15,8
12	10,6	16,4	8,2	15,0
13	9,7	15,1	7,7	13,9
14	8,9	13,8	7,2	12,9
15	8,3	12,7	6,9	12,0

TABLE 6.1: Pourcentage du nombre de modifications de la référence en fonction du nombre d'annotateurs.

6.3 Conclusions et perspectives

Cette étude expérimentale concerne des annotations où la tâche de l'annotateur consiste à choisir entre plusieurs classes. Concernant le choix de la mesure d'accord, aucune différence notable n'a été observée entre les 3 mesures μ-κ, μ-π et α_b, malgré les choix divergents concernant la façon de mesurer les accords (ou désaccords) dus au seul hasard. En revanche, lorsque les désaccords peuvent être quantifiés, l'introduction d'une distance entre classes dans la mesure d'accord, permet d'améliorer la stabilité des résultats et leur cohérence. Or, la mesure κ qui prévaut actuellement ne permet pas cette option lorsque le nombre d'annotateurs est supérieur à deux.

Par ailleurs, les résultats liés au nombre d'annotateurs et à la stabilité de la référence montrent que recourir à un grand nombre d'annotateurs permet de stabiliser l'annotation de référence. Nous aimerions prolonger le travail qui a été fait en essayant de préciser les liens entre la valeur de la mesure de l'accord entre les annotateurs et la stabilité de l'annotation de référence. Les seuils qui définissent ce qu'il est convenu un « bon accord » sont particulièrement significatifs du vague qui entoure ce genre de question. Ils varient d'un auteur à l'autre et, pour un même auteur, d'un article à l'autre tant le sujet est complexe et dépendant du domaine

considéré. Actuellement, nous réfléchissons à la façon d'utiliser les annotations dont nous disposons déjà pour générer des annotations fictives sur lesquelles travailler.

Chapitre 7

Bilan et perspectives

Les travaux présentés peuvent apparaître éclectiques et dispersés : à juste titre, car la distance est grande entre la « compréhension » de l'oral spontané dans un système de Dialogue Oral Homme Machine et le résumé automatique de documents extraits du web dans un domaine défini par un ou plusieurs mots clefs. De plus, si les tâches sont éloignées, les techniques employées le sont encore davantage : Logus a été conçu et développé entièrement « à la main », lexique et règles inclus, à partir de quelques centaines de phrases imaginées par des experts et qui tenaient lieu de corpus ; le système de résumé automatique qui a été testé s'appuie sur des techniques presque exclusivement statistiques, en analysant les gros corpus que représentent les versions française et anglaise de l'encyclopédie Wikipédia. Il est vrai aussi que j'ai parfois eu le sentiment d'arrêter de travailler sur un projet au moment même où je commençais à être capable d'y apporter des idées nouvelles. Mais c'est un peu l'écosystème dans lequel évolue l'enseignant-chercheur qui aboutit à cela...

En même temps et comme il avait été annoncé dans l'introduction, le lien qui réunit l'ensemble de ces travaux est la quête du sens, au sens large du terme, et la façon de le modéliser. Dans le cas de Logus et d'Emologus, il s'est agi d'interpréter le message porté par les mots, d'un point de vue pragmatique et d'un point de vue émotionnel. Dans les travaux concernant la recherche de patrons sémantiques, les approches décrites essaient de combiner les analyses linguistique et statistique pour essayer de mieux appréhender comment les mots ou groupes de mots s'associent pour « faire sens ». Avec le résumé automatique, la méthode est entièrement basée sur une approche statistique, basée sur le fait que l'analyse de très gros corpus est

actuellement la plus efficace pour obtenir une représentation sémantique des mots à la fois générique et robuste.

Les approches statistiques sont devenues prédominantes, voire hégémoniques, en Traitement Automatique des Langues, au détriment des approches logiques ou algébriques, actuellement souvent considérées comme inaptes à un traitement efficace des flux de documents textuels que l'on peut trouver sur le Web. Par ailleurs, alors même que l'objet d'études du TAL est la langue, la sophistication des nouveaux outils statistiques fait que parfois les linguistes ne se les sont pas totalement appropriés.

Certes, ce succès des approches statistiques prouvent leur efficacité ; de plus, il est difficile de ne pas reconnaître que les énoncés que tout un chacun dit ou écrit sont composés de groupes de mots qui ont déjà été utilisés précédemment des milliers ou des millions de fois. Cependant, il est permis de penser que le « tout statistique » a ses limites. Dans le cas, par exemple, du résumé automatique, si la méthode statistique est celle qui fonctionne le mieux, on a plutôt envie de dire, au vu des résultats, que c'est celle qui fonctionne le moins mal. L'amélioration de notre système de résumé multi-textes passerait probablement par une meilleure adaptation aux types des textes et aux domaines concernés ; et, pour ce faire, il faudrait probablement faire une analyse linguistique des corpus correspondants. Par ailleurs, on voit émerger de nouvelles approches issues de la branche du TAL qui s'intéresse aux méthodes formelles ; certains proposent des représentations plus complexes des mots prédicatifs (sous forme de matrices par exemple) afin de combiner analyse vectorielle statistique et compositionnalité du langage. De fait, le fait qu'une simple approche en « sac de mots » soit compétitive laisse penser qu'il existe encore d'importantes marges de progrès. Elles laissent largement la place à une combinaison possible d'approches et de méthodes, non encore bien définie.

Perspectives

L'intérêt pour Logus a été réactivé fin 2014 par les roboticiens qui espèrent l'utiliser pour des interactions spécifiques entre personnes âgées et robots-compagnons, telles que par exemple la gestion d'un agenda. Les derniers travaux auxquels j'ai participé ont consisté à proposer une réorganisation des concepts et de la connaissance de Logus qui permette de faciliter l'adaptation du système à

un domaine spécifique, et ce, en co-encadrant le stage de fin d'études d'un étudiant de l'ENSIBS.

Le principal chantier qui s'ouvre actuellement, et qui est la raison d'être de cette candidature à l'HDR, est lié aux travaux de la thèse que débute actuellement Stefania Pecore. Son stage de fin d'étude de master, proposé par Jean-Yves Antoine et que j'ai co-encadré, consistait à expérimenter différentes techniques pour étendre les bases lexicales en émotion ou en opinion. Comme souvent en TAL, les ressources existantes pour la langue française sont très pauvres. À preuve, le challenge DEFT 2015 où les participants devaient évaluer l'émotion portée par les tweets. Beaucoup d'entre eux ont utilisé un lexique émotionnel traduit de l'anglais. On sait que le décalage entre les différentes langues fait qu'une traduction ne restitue que partiellement le sens d'un mot alors que déjà, la relation de synonymie est loin de correspondre à une égalité en terme de valeurs émotionnelles : les mots *mère* et *maman* en sont un exemple ; il est permis de penser que la qualité d'un lexique émotionnel obtenu par traduction est très probablement médiocre.

Après un an passé à Milan dans un organisme de recherche italien, Stefania rejoint l'équipe Expression de l'IRISA en tant que doctorante pour travailler sur un sujet lié à celui de son stage, à savoir l'analyse de sentiment et la fouille d'opinion sur les réseaux sociaux. Il est prévu que je la co-encadre, ainsi que Farida Saïd-Hocine (du LMAM, Laboratoire de Mathématiques), sous la direction de Pierre-François Marteau (IRISA, équipe Expression). Le domaine est actuellement extrêmement porteur et les enjeux sont énormes, tant au point de vue économiques que politiques. La tâche est complexe, les méthodes devant très certainement dépendre du style de texte (tweet, forums, etc.) et du domaine testé. J'espère que nous pourrons précisément y mettre en œuvre une combinaison d'approches statistiques avec des approches linguistiques : les compétences de Stefania devraient en offrir l'occasion.

À côté de ce travail, j'espère également qu'avec Jean-Yves Antoine et Anaïs Lefeuvre, nous pourrons poursuivre nos travaux concernant les annotations, les accords inter-annotateurs et la qualité des corpus annotés.

Et, pour finir, il y a également le projet TAL-breizh déposé à la MSHB de Rennes porté par Annie Foret (IRISA, équipe LIS) auquel Farida et moi devrions participer s'il est accepté ; sans oublier le projet Asialog consacré à la numérisation des journaux de bord de la compagnie des Indes, et l'extraction des informations historiquement intéressantes qu'ils contiennent dans leurs parties purement

textuelles ; et sans oublier non plus les travaux liés aux cartes de Kohonen et à la modélisation du processus cognitif chez les enfants que Farida et moi aimerions reprendre et prolonger.

Trop de projets, pensez-vous ? Peut-être mais peut-être pas. La question est-elle vraiment importante ?

Bibliographie

ACE (2008). Automatic content extraction 2008 evaluation plan. Assessment of Detection and Recognition of Entities and Relations Within and Across Documents. cité page 60

Achour, A. (2010). *Contribution à l'étude de la modélisation des connaissances enfantines : mise en oeuvre d'une approche utilisant les cartes auto-organisatrices pour la classification et l'apprentissage.* PhD thesis, Université de Bretagne Sud. cité page 28

Achour, A., Villaneau, J., and Duhaut, D. (2008a). Cognitive and Emotional Interaction. In Sojka, P., Horàk, A., Kopecek, I., and Pala, K., editors, *Text, Speech and Dialogue*, pages 553–560. Springer Berlin / Heidelberg. https://hal.archives-ouvertes.fr/hal-00515228/file/TSD_CognitiveEmotionalInteraction.pdf. cité page 28, 29

Achour, A., Villaneau, J., Duhaut, D., and Saïd, F. (2008b). Cognitive and Emotional linguistic Interaction. In *2008 International Conference on Multimodal Interfaces*, Chania, Crete, Greece. https://hal.archives-ouvertes.fr/hal-00515233. cité page 29

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., and Wiebe, J. (2014). Semeval-2014 task 10 : Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics and Dublin City University. cité page 73

Antoine, J.-Y., Le Tallec, M., and Villaneau, J. (2011). Evaluation de la détection des émotions, des opinions ou des sentiments : dictatute de la majorité ou respect de la diversité d'opinions ? In *TALN'2011*, volume 2, Montpellier, France. https://hal.archives-ouvertes.fr/hal-00625727. cité page 86

Antoine, J.-Y., Villaneau, J., and Lefeuvre, A. (2014). Weighted Krippendorff's alpha is a more reliable metrics for multi- coders ordinal annotations : experimental studies on emotion, opinion and coreference annotation. In *EACL 2014*, Gotenborg, Sweden. http://www.aclweb.org/anthology/E14-1058. cité page 86

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4) :555–596. cité page 85, 86, 87

Bassano, D., Labrell, F., and Champaud, C. (2005). Le dlpf, un nouvel outil pour l'évaluation du développement du langage de production en français. *Enfance*, 2(5) :171–208. cité page 22, 29, 42, 50

Bonneau-Maynard, H., C., A., Béchet, F., Denis, A., Kuhn, A., Lefevre, F., Mostefa, D., Quignard, M., Rosset, S., Servan, C., and Villaneau, J. (2006). Results of the french evalda-media evaluation campaign for literal understanding. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. cité page 10

Boudreau, S. and Kittredge, R. (2005). Résolution des anaphores et détermination des chaînes de coréférences. *Traitement Automatique des Langues (TAL)*, 46(1) :41–69. cité page 11

Boyd, A., Gegg-Harrison, W., and Byron, D. (2005). Identifying non-referential it : a machine learning approach incorporating linguistically motivated patterns. *Traitement Automatique des Langues (TAL)*, 46(1) :71–90. cité page 14

Bretonnel Cohen, K., Johnson, H. L., Verspoor, K., Roeder, C., and Hunter, L. E. (2010). The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics, 11 :492*. cité page 57

Callejas, Z. and Lopez-Cozar, R. (2008). Influence of contextual information in emotion annotation for spoken dialogue systems. *Speech Communication*, 50 :416–433. cité page 19

Cao, G., Gao, J., Nie, J.-Y., and Redmond, W. (2007). A system to mine large-scale bilingual dictionaries from monolingual web. *Proc. of MT Summit XI*, pages 57–64. cité page 57

Carletta, J. (1996). Assessing agreement on classification tasks : the kappa statistic. *Computational Linguistics*, 22(2) :249–254. cité page 87

Chatterjee, N. and Mohan, S. (2007). Extraction-based single-document summarization using random indexing. In *ICTAI(2)*, pages 448–455. IEEE Computer Society. cité page 67

Church, K. and Hanks, P. (1996). Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, 16(1) :22–29. cité page 48

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 :37–46. cité page 87

Cohen, J. (1968). Weighted kappa : nominal scale agreement with provision for scaled disagreement of partial credit. *Psychological Bulletin*, 70 :213–220. cité page 87

Cottrell, M., Ibbou, S., and Letrémy, P. (2003). Traitement des données manquantes au moyen de l'algorithme de Kohonen. In *Jacques Marie Aurifeille*, pages 201–217, Université de Nantes. https://hal.archives-ouvertes.fr/hal-00141475. cité page 39

Cowie, R. and Cornelius, R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, 40 :5–32. cité page 18

Craggs, R. and Wood, M. M. (2005). Evaluating discourse and dialogue coding schemes. *Computational Linguistics*, 31(3) :289–295. cité page 87

Dang, H. (2006). Overview of duc 2006. In *HLT-NAACL*. Document Understanding Workshop. cité page 78

Davies, M. and Fleiss, J. (1982). Measuring agreement for multinomial data. *Biometrics*, 38 :1047–1051. cité page 87

de Loupy, C., Guégan, M., Ayache, C., Seng, S., and Torres Moreno, J.-M. (2010). A french human reference corpus for multi-document summarization and sentence compression. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA). cité page 79

Devillers, L., Maynard, H., Rosset, S., Paroubek, P., McTait, K., Mostefa, D., Choukri, D., Charnay, L., Bousquet, C., Vigouroux, N., Béchet, F., Romary, L., Antoine, J.-Y., Villaneau, J., Vergnes, M., and Goulian, J. (2004). The french media/evalda project : the evaluation of the understanding capability of

spoken language dialogue systems. In *Proceedings of the fourth International Conference on Language Resources and Evaluation (LREC'04)*. cité page 8

Devillers, L. and Vasilescu, I. (2005). Emotion detection in task-oriented spoken dialogs. *Journal of Neural Networks.*, 18(4). cité page 19

Di Eugenio, B. and Glass, M. (2004). The Kappa statistic : A second look. *Computational Linguistics*, 30(1) :95–101. cité page 87

Ekman, P. (1999). *Handbook of Cognition and Emotion.*, chapter Basic Emotions, pages 45–60. NY : John Wiley & Sons Ltd, New York, Dalgleish, T. & Power, M. J. edition. cité page 18

El Maarouf, I. (2011). *Corpus-based knowledge formalization : context linguistic modeling for automatic semantic relation extraction.* Theses, Université de Bretagne Sud. https://tel.archives-ouvertes.fr/tel-00657708/file/These_Ismail_El-Maarouf_2011.pdf. cité page 47

El Maarouf, I. and Villaneau, J. (2012a). A french fairy tale corpus syntactically and semantically annotated. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA). cité page 50

El Maarouf, I. and Villaneau, J. (2012b). Parenthetical classification for information extraction. In *Proceedings of COLING 2012 : Posters*, pages 297–308, Mumbai, India. The COLING 2012 Organizing Committee. cité page 57

El Maarouf, I., Villaneau, J., and Rosset, S. (2011). Extraction de patrons sémantiques appliquée à la classification d'entités nommées. In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Montpellier, France. Association pour le Traitement Automatique des Langues. cité page 53

El Maarouf, I., Villaneau, J., Saïd, F., and Duhaut, D. (2009). Comparing child and adult language : Exploring semantic constraints. In *Wocci 2009, 2nd Workshop on Child, Computer and Interaction*, Boston (USA). ACM. cité page 50

Firth, J. (1957). A synopsis of linguistic theory 1930–1955. *Studies in linguistic analysis*, Special volume of the Philological Society. cité page 47

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76 :378–382. cité page 87

Forbes-Riley, K. and Litman, L. (2004). Predicting emotion in spoken dialogue from multiple knowledge sources. In *HLT/NAACL'2004*, pages 161–176. cité page 19

Gardent, C. and Manuelian, H. (2005). Création d'un corpus annoté pour le traitement des descriptions définies. *Traitement Automatique des Langues (TAL)*, 46(1) :115–139. cité page 10

Goldstein, J. and Carbonell, J. (1998). Summarization : (1) using MMR for diversity - based reranking and (2) evaluating summaries. In *Proceedings of a Workshop on Held at Baltimore, Maryland : October 13-15, 1998*, TIPSTER '98, pages 181–195, Stroudsburg, PA, USA. Association for Computational Linguistics. cité page 78

Grishman, R. (2010). The impact of task and corpus on event extraction systems. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. cité page 54

Hanks, P. (2008a). Lexical patterns : From hornby to hun- ston and beyond. In *Euralex*, pages 89–129. cité page 48

Hanks, P. (2008b). Mapping meaning onto use : a pattern dictionary of english verbs. In *AACL 2008*, Utah. cité page 23, 48

Harris, Z. (1954). Distributional structure. *Word*, 10(3) :146–162. cité page 47

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, France. cité page 53

Higgins, D. and Burstein, J. (2007). Sentence similarity measures for essay coherence. In *Proceedings of the 7th International Workshop on Computational Semantics*, pages 1–12. cité page 68

Hovy, E., Lin, C.-Y., Zhou, L., and Fukumoto, J. (2006). Automated summarization evaluation with basic elements. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. cité page 78

Ibbou, S. (1998). *Classification, analyse des correspondances et méthodes neuronales*. PhD thesis, Université de Paris 1 Panthéon Sorbonne. cité page 39

Inderjeet, M. (2001). *Automatic Summarization*. John Benjamins Publishing. cité page 78

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 46 :59–69. cité page 32

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78 :1464–1480. cité page 33

Kohonen, T. (2001). Self-organizing maps, third edition. *Springer Series in Information Sciences.*, 30. cité page 32

Krippendorff, K. (1980). *Content Analysis : An Introduction to Its Methodology*, chapter 12. Sage, Beverly Hills, CA. cité page 87

Krippendorff, K. (2004). Reliability in content analysis : Some common misconceptions and recommendations. *Human Communication Research*, 30(3) :411–433. cité page 87

Lakoff, G. (1987). *Women, fire and dangerous things : What categories reveal about the mind*. University of Chicago Press. cité page 30

Landis, J. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33 :159–174. cité page 19

Le Tallec, M., Antoine, J.-Y., Villaneau, J., and Duhaut, D. (2011). Affective Interaction with a Companion Robot for Hospitalized Children : a Linguistically based Model for Emotion Detection. In *5th Language and Technology Conference (LTC'2011)*, Poznan, Poland. https://hal.archives-ouvertes.fr/hal-00664618. cité page 19

Le Tallec, M., Villaneau, J., Antoine, J.-Y., Savary, A., and Syssau-Vaccarella, A. (2009). Détection des émotions à partir du contenu linguistique d'énoncés oraux : application à un robot compagnon pour enfants fragilisés. In *Actes de la 16e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*. cité page 24

Le Tallec, M., Villaneau, J., Antoine, J.-Y., Savary, A., and Syssau-Vaccarella, A. (2010). Emologus - A Compositional Model of Emotion Detection based on the Propositionnal Content of Spoken Utterances. In *Text Speech and Dialogue 2010*, volume 6231 of *LNCS/LNAI*, page 8 pages, Brno, Czech Republic. Springer. https://hal.archives-ouvertes.fr/hal-00536786. cité page 18

Lee, C. and Narayanan, S. (2005). Towards detecting emotions in spoken dialogs. *IEEE Transactions On Speech and Audio Processing.*, 13(3) :293–303. cité page 19

Li, Y., McLean, D., Bandar, Z. A., O'shea, J. D., and Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *Knowledge and Data Engineering, IEEE Transactions on*, 18(8) :1138–1150. cité page 76

Lin, C.-Y. (2004). Rouge : a package for automatic evaluation of summaries. In *Workshop on Text Summarization*, pages 25–26. cité page 78

Marcu, D. (2000). The rhetorical parsing of unrestricted texts : A surface-based approach. *Computational Linguistics*, 26(3) :395–448. cité page 53

Mareschal, D. and Quinn, P. (2001). Categorization in infancy. *TRENDS in Cognitive Sciences*, 5(10) :443–450. cité page 30

Mei, Q., Guo, J., and Radev, D. R. (2010). Divrank : the interplay of prestige and diversity in information networks. In Rao, B., Krishnapuram, B., Tomkins, A., and Yang, Q., editors, *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pages 1009–1018. ACM. cité page 79

Mihalcea, R., Corley, C., and Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *IN AAAI'06*, pages 775–780. cité page 71

Morin, E. (1998). Prométhée : un outil d'aide à l'acquisition de relations sémantiques entre termes. In *5ème Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 172–181, Paris, France. cité page 53

Muzerelle, J., Lefeuvre, A., Schang, E., Antoine, J.-Y., Pelletier, A., Maurel, D., Eshkol, I., and Villaneau, J. (2014). ANCOR_Centre, a Large Free Spoken French Coreference Corpus : description of the Resource and Reliability Measures. In ELRA, editor, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reyjavik, Iceland. https://hal.archives-ouvertes.fr/hal-01075679. cité page 88

Nenkova, A. and Passonneau, R. (2004). Evaluating content selection in summarization : the pyramid method. In *NAACL-HLT*. cité page 78

Neto, J. L., Freitas, A. A., and Kaestner, C. A. (2002). Automatic text summarization using a machine learning approach. In *Advances in Artificial Intelligence*, pages 205–215. Springer. cité page

Neto, J. L., Santos, A. D., Kaestner, C. A., and Freitas, A. A. (2000). Generating text summaries through the relative importance of topics. In *Advances in Artificial Intelligence*, pages 300–309. Springer. cité page

Okazaki, N., Ishizuka, M., and Tsujii, J. (2008). A discriminative approach to japanese abbreviation extraction. In *IJCNLP proceedings*, pages 889–894. cité page 57

Pingali, P., K, R., and Varma, V. (2007). Iiit hyderabad at duc 2007. In *NAACL-HLT 2007*. cité page 79

Pustejovsky, J. (1998). *The generative lexicon.* MIT Press, Cambridge (Ma) edition. cité page 48

Reed, J. W., Jiao, Y., Potok, T. E., Klump, B. A., Elmore, M. T., and Hurson, A. R. (2006). Tf-icf : A new term weighting scheme for clustering dynamic data streams. *Machine Learning and Applications, Fourth International Conference on*, 0 :258–263. cité page 67

Rosset, S., Galibert, O., Bernard, G., Bilinski, E., and G., A. (2008). The LIMSI participation to the QAst track. In *Actes de Working Notes of CLEF 2008 Workshop.* cité page 53, 54

Sahlgren, M. (2005). An introduction to random indexing. In *Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering, TKE*, volume 5. cité page 67

Sahlgren, M. (2006). *The Word-Space Model : Using distributional analysis to represent syn- tagmatic and paradigmatic relations between words in high-dimensional vector spaces.* PhD thesis, Department of Linguistics, Stockholm University. cité page 66

Scott, W. (1955). Reliability of content analysis : the case of nominal scale coding. *Public Opinions Quaterly*, 19 :321–325. cité page 87

Sjöbergh, J. (2007). Older versions of the rougeeval summarization evaluation system were easier to fool. *Inf. Process. Manage.*, 43(6) :1500–1505. cité page 78

Sloutsky, V. (2003). The role of similarity in the development of categorization. *Opinion TRENDS in Cognitive Sciences*, 7(6) :246–251. cité page 30

Sowa, J. (2001). Conceptuals Graphs. http://users.bestweb.net/~sowa/cg/cgstand.htm. cité page 6

Syssau, A. and Monnier, C. (2009). Children's emotional norms for six hundred french words. *Behavior, Research, and Methods*, 41 :213–219. cité page 23

Torres-Moreno, J.-M. (2011). *Résumé automatique de documents : une approche statistique*. Recherche d'information et Web. Hermès. cité page 78

Toutanova, K., Brockett, C., Gamon, M., Jagarlamudi†, J., Suzuki, H., and Vanderwende, L. (2007). The PYTHY summarization system : Microsoft research at DUC 2007. In *NAACL-HLT 2007*. cité page 78

Van Noord, G., Bouma, G., Koeling, R., and Nederhof, M.-J. (1999). Robust grammatical analysis for spoken dialogue systems. *Natural Language Engineering*, 5 :45–93. cité page 5

Vanderveken, D. (2001). *Essays in Speech Act Theory.*, chapter Universal Grammar and Speech Act Theory, pages 25–62. John Benjamin, Amsterdam Philadelphia, D. Vanderveken and Susumu Kubo edition. cité page 6

Vasa, R., Carlino, A., London, K., and Min, C. (2006). Valence ratings of emotional and non-emotional words in children. *Personality and Individual Differences*, 41 :1169–1180. cité page 23

Villaneau, J. (2003). *Contribution au traitement syntaxico-pragmatique de la langue naturelle parlée : approche logique pour la compréhension de la parole.* PhD thesis, Université de Bretagne Sud, Vannes, France. cité page 6

Villaneau, J. and Antoine, J.-Y. (2004). Categorials grammars used to partial parsing of spoken language. In *Actes of CG2004*, pages 244–258, Montpellier, France. cité page 6

Villaneau, J., Ridoux, O., and Antoine, J.-Y. (2004). Logus : compréhension de l'oral spontané. présentation et évaluation des bases formelles de logus. *Revue d'intelligence artificielle*, 18(5-6) :709–742. cité page 6

Vu, H.-H., Villaneau, J., Saïd, F., and Marteau, P.-F. (2014). Sentence Similarity by Combining Explicit Semantic Analysis and Overlapping N-Grams.

In Springer, editor, *Text, Speech and Dialogue*, volume 8655, pages 201–208, Brno, Czech Republic. https://hal.archives-ouvertes.fr/hal-01066170. cité page 65

Vu, H.-H., Villaneau, J., Saïd, F., and Marteau, P.-F. (2015). Mesurer la similarité entre phrases grâce à Wikipédia en utilisant une indexation aléatoire. In *TALN 2015*, Caen, France. https://hal.archives-ouvertes.fr/hal-01167929. cité page 65

Wilks, Y. (1975). A preferential, pattern-seeking, semantics for natural language inference. *Artif. Intell.*, 6(1) :53–74. cité page 48

Wong, S. K. M., Ziarko, W., and Wong, P. C. N. (1985). Generalized vector spaces model in information retrieval. In *SIGIR ACM*. cité page 66

Zhou, G., Su, J., Zhang, J., and Zhang, M. (2005). Exploring various knowledge in relation extraction. In *proceedings of ACL*. cité page 62

Table des figures

Liste des tableaux

Publications

PUBLICATIONS

Conférences internationales avec Comité de lecture

[2014a] H. H. Vu, J. Villaneau, F. Saïd , P-F. Marteau (2014) *Sentence Similarity by combining Explic Semantic Analysis and overlapping n-grams.* Proc. of Text, Speech and Dialogue (TSD) 2014, Brno, Czech Republic HAL .

[2014b] J.-Y. Antoine, J. Villaneau, A. Lefeuvre (2014) *Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation.* Proc. 14th Conference of the European Chapter of the Association of Computational Linguistics, EACL 2014, Gothenburg, Suède HAL .

[2014c] J. Muzerelle, A. Lefeuvre, E. Schang, J-Y. Antoine, A. Pelletier, D. Maurel, I. Eskol, J. Villaneau (2014) *ANCOR_Centre, a Large Free Spoken French Coreference Corpus: Description of the Resource and Reliability Measures*. Proc. *LREC'2014*, Reykjavik, Islande HAL

[2012a] I. El Maarouf, J. Villaneau (2012) *A French Fairy Tale Corpus syntactically and semantically annotated*. Proceedings of the Eight International Conference on Language Resources and Evaluation, LREC'12, HAL .

[2012b] I. El Maarouf, J. Villaneau (2012) Parenthetical *Classification for Information Extraction.*Proceedings of COLING 2012: Posters. Decembre 2012, Mumbai, India HAL

[2011a] M. Le Tallec, J-Y. Antoine, J. Villaneau, D. Duhaut (2011) *Affective Interaction with a Companion Robot for vulnerable Children: a Linguistically based Model for Emotion Detection.* Proc. LTC'2011, Language Technology Conference, Poznan, Poland. 445-450, HAL .

[2010b] M. Le Tallec, J. Villaneau, J-Y. Antoine, A. Savary, A. Syssau-Vacarella (2010) *Emologus - a compositional model of emotion detection based on the propositionnal content of spoken utterances.*Proc. 13th International Conference on Text, Speech and Dialogue, TSD'2010, Brno, Czech Republic, sept. 2010 In LNCS/LNAI 6231, Springer, ISBN: 978-3-642-15759-2 HAL

[2010c] M. Le Tallec, S. Saint-Aime, C. Jost, J. Villaneau, J-Y. Antoine, S. Letellier-Zarshenas, B. Le Pevedic, D. Duhaut (2010) *From speech to emotional interaction: EmotiRob project.* Proc. 3rd International Conference on Human-Robot Personal Relationships, HRPR'2010, Leiden, NL, june 2010, pp. 57-64 HAL .

[2009a] I. El Maarouf, J. Villaneau, F. Saïd, D. Duhaut (2009) *Comparing Child and Adult Language : Exploring Semantic constraints.* ICMI-MLMI'09 Workshop on Child, Computer and Interaction, Cambridge, MA : États-Unis HAL .

[2009b] J. Villaneau, J-Y. Antoine *(2009) Deeper spoken language understanding for man-machine dialogue on broader application domains: a logical alternative to concept spotting.* Proc. Workshop on the Semantic Representation of Spoken Language, SRSL'2009, EACL'2009, Athens, Greece, April 2009 HAL .

[2009c] J-Y. Antoine, J. Goulian, J. Villaneau, M. Le Tallec (2009) *Word Order Phenomena in Spoken French: a Study on Four Corpora of Task-Oriented Dialogue and its Consequences on Language Processing.* Proc. Corpus Linguistics'2009, Liverpool, UK, July 2009 HAL

[2008a] A. Achour, M. Le Tallec, S. Saint-Aime, J. Villaneau, J-Y. Antoine, B. Le Pevedic, D. Duhaut (2008). *EMOTIROB : from understanding to cognitive interaction.* Proc. IEEE International Conference on Mechatronics and Automation, ICMA'2008. Takamatsu, Japon. p. 369-374, HAL

[2008b] A. Achour, J. Villaneau, D. Duhaut, et F. Said (2008) *Cognitive and Emotional Linguistic Interaction.* In Child, Computer and Interaction (ICMI'08 post-conference workshop), Chania, Crete, Greece, octobre 2008, HAL.

[2008c] A. Achour, J. Villaneau, et D. Duhaut. (2008) *Cognitive and Emotional Interaction.* In Text, Speech and Dialogue 2008 (TSD 2008), volume 5246 de LNAI, Brno, Czech Republic, septembre 2008. Springer-Verlag, HAL.

[2007a] J. Villaneau, S. Rosset, O. Galibert (2007) *Semantic Relations for an Oral and Interactive Question-Answering System.* SRSL7 (Semantic Representation of Spoken Language 2007), Salamanca : Espagne (2007) HAL

[2006a] J. Villaneau, O. Ridoux (2006) *Computation and Representation of Meaning in a Man-Machine Dialogue: a pragmatic Combination of logical Formalisms.* Computers and Philisophy, an international conference (i-C&P 2006).

[2006b] H. Bonneau-Maynard, C.Ayache, F. Bechet, A. Denis, A.Lefebvre, D. Mostafa, M. Quignard, S. Rosset, C. Servan, J. Villaneau (2006) *Results of the French Evalda-Media evaluation campaign for literal understanding.* Proc of LREC (2006) HAL.

[2004a] J. Villaneau, J-Y. Antoine, O. Ridoux (2004) *Logical approach to natural language understanding in a spoken dialog system.* Proc. 7th International Conference on Text Speech and Dialog, TSD'2004, Brno, République tchèque. pp. 637-644. In LNCS/LNAI 3206, Springer HAL.

[2004c] J. Villaneau, J-Y. Antoine (2004) *Categorial grammars used to partial parsing of spoken language.* Proc. Categorial Grammars'2004, Montpellier, France.

[2004d] L. Devillers, H. Bonneau-Maynard, S. Rosset, P. Paroubek, K. McTait, D. Mostefa, K. Choukri, L. Charnay, C. Bousquet, N. Vigouroux, F. Béchet, L. Romary, J-Y. Antoine, J. Villaneau, M. Vergnes, J. Goulian (2004) The French MEDIA/EVALDA Project: the Evaluation of the Understanding Capability of Spoken Language Dialogue Systems, *In International Conference on Language Resources and Evaluation*.

[2002a] J-Y. Antoine, C. Bousquet-Vernhettes, J. Goulian, M. Zakaria Kurdi, S. Rosset, N. Vigouroux, J. Villaneau (2002). *Predictive and objective evaluation of speech understanding: the challenge evaluation campaign of the I3 speech workgroup of the French CNRS.* Proc. 3rd International Conference on Language Resources & Evaluation, LREC'2002, Las Palmas de Gran Canaria, Espagne. pp.529-535

[2001a] J. Villaneau, J-Y. Antoine, O. Ridoux (2001). *Combining syntax and pragmatic knowledge for the understanding of spontaneous spoken sentences.* Proc. 4th Conference on Logical Aspects of Computational Linguistics, LACL'2001, Le Croisic, France. In P. de Groote, G. Morrill, C. Retore (Eds.) LNAI 2099, Springer Verlag, pp. 279-295.

[2000] J-Y. Antoine, J. Siroux, J. Caelen, J. Villaneau, J. Goulian, M. Ahafhaf (2000). *Obtaining predictive results with an objective evaluation of spoken dialogue systems: experiments with the DCR assessment paradigm*. Proc. 2nd International Conference on Language Resources & Evaluation, LREC'2000, Athenes,

Chapitre d'ouvrage

[2008d] H. Bonneau-Maynard, A. Denis, F. Béchet, L. Devillers, F. Lefèvre et al. (2008) *Media: évaluation de la compréhension dans les systèmes de dialogue.* Stéphane Chaudiron; Khalid Choukri. *L'évaluation des technologies de traitement de la langue, les campagnes Technolangue*, Hermès, Lavoisier, pp.209-232, 2008, Cognition et traitement de l'information, 978-2746219922 HAL

Revues francophones avec comité de lecture

[2012c] J-Y. Antoine, J. Villaneau, J. Goulian (2012) *Influence du genre applicatif sur la réalisation des extractions en dialogie oral : constantes et variations.* Revue Langages, septembre 2012, 187, p. 109-126 HAL

[2004b] J. Villaneau, O. Ridoux, J-Y. Antoine (2004). *LOGUS : un système formel de compréhension de l'oral spontané.* RIA, Revue d'Intelligence Artificielle, 18(5/6) Lien.

Conférences francophones avec comité de lecture

[2015a] H. H. Vu, J. Villaneau, F. Saïd , P-F. Marteau. (2015) *Mesurer la similarité entre phrases grâce à Wikipédia en utilisant une indexation aléatoire.* Actes de TALN'2015, Caen, juin 2015 HAL.

[2013a] J. Muzerelle, A. Lefeuvre, J-Y. Antoine, E. Schang, D. Maurel, J. Villaneau, I. Eshkol (2013) *ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement.* Actes TALN'2013, Les Sables d'Olonnes, juin 2013 HAL

[2013b] H. H. Vu, J. Villaneau, F. Saïd. (2013) *Utilisation des liens Wikipédia pour la détection automatique des concepts d'un domaine*. Journée de Linguistique de Corpus (JLC). Lorient, 2013.

[2011b] J-Y. Antoine, M. Le Tallec, J. Villaneau (2011) *Evaluation de la détection des émotions, des opinions ou des sentiments : dictature de la majorité ou respect de la diversité d'opinions ?* Actes TALN'2011, Montpellier, France, Juillet 2011, HAL

[2011c] I. El Maarouf, J. Villaneau, S. Rosset (2011).*Extraction de patrons sémantiques appliquée à la classification d'Entités Nommées***.** Actes TALN'2011

[2010a] M. Le Tallec, J. Villaneau, J-Y. Antoine, A. Savary, A. Syssau (2010) *Détection hors contexte des émotions à partir du contenu linguistique d'énoncés oraux : une approche compositionnelle.* Actes TALN'2010, Montréal, Québec, juillet 2010 HAL

[2009d] M. Le Tallec, J. Villaneau, J-Y. Antoine, A. Savary, A. Syssau (2009) *Détection des émotions à partir du contenu linguistique d'énoncés oraux : application à un robot compagnon pour enfants fragilisés.* Actes TALN'2009, Senlis, France HAL

[2009e] I. El Maarouf, M. Le Tallec, J. Villaneau, G. Williams (2009) *Ontologies Naturelles et Coercion : Formalisation de Connaissances à partir d'observations en Corpus.* Journées de Linguistique de Corpus, 2009, Lorient, France. 2009 HAL

[2007b] J. Villaneau (2007) *Une expérience de compréhension en contexte de dialogue avec le système LOGUS, approche logique de la compréhension de la langue orale.* Actes TALN'2007, Toulouse, France HAL **.**

[2005] J. Villaneau, S. Lamprier (2005) *Corpus de dialogue Homme-Machine : annotation sémantique et compréhension.* Journées de la Linguistique de Corpus (JLC), Lorient, 2005.

[2003a] J-Y. Antoine, J. Goulian, J. Villaneau (2003). *Quand le TAL robuste s'attaque au langage parlé : analyse incrémentale pour la compréhension de la parole spontanée.* Actes TALN'2003, Batz-sur-Mer, France. pp. 25-34.

[2002b] J. Villaneau, J-Y. Antoine, O. Ridoux (2002). *LOGUS : un système formel de compréhension du français parlé spontané : présentation et évaluation.* Actes TALN'2002, Nancy, France, pp. 165-174