Extraction et regroupement de descripteurs morpho-syntaxiques pour des processus de Fouille de Textes

Nicolas Béchet

Académie de Montpellier

Université Montpellier II

— Sciences et Techniques du Languedoc —

Thèse

présentée à l'Université des Sciences et Techniques du Languedoc
pour obtenir le diplôme de DOCTORAT

Spécialité : **Informatique**
Formation Doctorale : **Informatique**
École Doctorale : **Information, Structures, Systèmes**

Extraction et regroupement de descripteurs morpho-syntaxiques pour des processus de Fouille de Textes

par

Nicolas BÉCHET

Soutenue le 8 décembre 2009

27ème Section : Informatique devant le Jury composé de :

Catherine Berrut, Professeur, Université Joseph Fourier, ... Rapportrice
Christophe Roche, Professeur, Université de Savoie, ... Rapporteur
Violaine Prince, Professeur, Université Montpellier 2, ... Examinatrice
Anne Vilnat, Professeur, Université Paris-Sud, ... Examinatrice
Jacques Chauché, Professeur, Université Montpellier 2, ... Directeur de thèse
Mathieu Roche, Maître de conférences, Université Montpellier 2, ... Co-directeur de thèse

Académie de Montpellier

Université Montpellier II

— Sciences et Techniques du Languedoc —

Thèse

présentée à l'Université des Sciences et Techniques du Languedoc
pour obtenir le diplôme de DOCTORAT

Spécialité : **Informatique**
Formation Doctorale : **Informatique**
École Doctorale : **Information, Structures, Systèmes**

Extraction et regroupement de descripteurs morpho-syntaxiques pour des processus de Fouille de Textes

par

Nicolas BÉCHET

Soutenue le 8 décembre 2009

27ème Section : Informatique devant le Jury composé de :

Catherine Berrut, Professeur, Université Joseph Fourier, ... Rapportrice
Christophe Roche, Professeur, Université de Savoie, ... Rapporteur
Violaine Prince, Professeur, Université Montpellier 2, ... Examinatrice
Anne Vilnat, Professeur, Université Paris-Sud, ... Examinatrice
Jacques Chauché, Professeur, Université Montpellier 2, ... Directeur de thèse
Mathieu Roche, Maître de conférences, Université Montpellier 2, ... Co-directeur de thèse

Remerciements

Je tiens tout d'abord à remercier **Mathieu Roche**, mon co-directeur de thèse, pour ses remarques pertinentes et son optimisme qui me fait parfois défaut. En outre ses qualités d'encadrant, m'ont toujours permis d'avancer à un rythme régulier dans mes travaux, m'encourageant à persévérer lorsque les résultats obtenus n'étaient pas toujours ceux attendus.

Je remercie également **Jacques Chauché** en tant que directeur de ma thèse, d'avoir toujours été disponible lors de mes requêtes, incluant notamment celles propres à l'analyseur syntaxique qu'il a développé : SYGFRAN.

Je remercie **Violaine Prince** d'avoir accepté de présider mon jury de thèse. Je la remercie également de m'avoir fait part de ses remarques pertinentes tout au long de mes travaux de thèse.

Merci à mes rapporteurs **Catherine Bérrut** et **Christophe Roche**. Je les remercie d'avoir accepté de rapporter mon manuscrit de thèse. Je les remercie pour les remarques fondées et leurs questions avisées qui m'ont permis d'améliorer mes travaux en ouvrant des perspectives intéressantes.

Merci également à **Anne Vilnat** d'avoir accepté d'être examinatrice de mon jury de thèse. Sa spécialisation en TAL a permis d'avoir une autre vision de mes travaux de recherche.

Je remercie pour finir mes collègues de l'équipe TAL du LIRMM.
Merci tout d'abord à **Alexandre Labadié**, ancien collègue de bureau avec qui j'ai pu partager de nombreuses discussions aussi bien scientifiques que tout autre sujet, ayant un certain nombre de goûts communs, notamment musicaux...

Je remercie **Mehdi Yousfi-Monod**, ancien collègue de bureau également, mais également ancien Assadin, avec qui j'ai pu partager des points de vu, notamment sur le monde du logiciel libre.

Je remercie encore **Mathieu Lafourcade** pour sa sympathie et la justesse de ces remarques.

Merci également à mes nouveaux collègues de bureau, **Johan Segura** et **Cédric Lopez**

avec qui j'ai souvent des conversations captivantes. Je leur souhaite également une grande réussite dans leurs travaux de thèse.

Merci pour finir à toutes les personnes qui ont contribué de près ou de loin à faire que ma tâche d'enseignant-chercheur soit menée au mieux au cours de ces trois années : **Christophe Borelly**, **Sébastien Druon**, **Michel Facerias**, **Rémy Kessler**, **Jocelyne Nanard**, **Pascal Poncelet**, **Marc Plantevit**, **Patrice Ravel** et **Myrtille Vivien**.

Sommaire

Chapitre 1

Introduction

Sommaire

1.1 Problématique

1.1.1 Le besoin humain de communiquer

Il est un besoin intemporel et universel chez l'être humain que la nécessité de communiquer. Cette communication s'effectue par l'emploi de signes pouvant être des sons et/ou des symboles qui deviendront avec le temps la parole et l'écriture. Ce type de communication se définit comme une faculté qui nous est propre : *le langage*. Le fait de s'exprimer par un langage nécessite alors l'emploi d'une *langue*. Cette vision des langues fondée sur un système de signes peut être attribuée à Ferdinand de Saussure [de Saussure, 1916] qui est souvent considéré comme un des fondateurs de la linguistique contemporaine.

L'ère de l'informatique a engendré une problématique nouvelle : produire de manière automatique des tâches autrefois réalisées par l'homme. Les principales motivations à l'origine de la réalisation de ces traitements furent sans doute la seconde guerre mondiale et la guerre froide qui s'en suivit. Nous vîmes en effet apparaître à cette époque les premiers systèmes de traduction automatique qui se concrétisèrent dans les années 50 avec la première conférence sur ce sujet. Ces applications furent à l'origine du traitement de langues "naturelles". Le terme "naturel" s'oppose alors aux langues dites "formelles" ou "artificielles". En outre, si la linguistique est l'étude des langues naturelles et *a fortiori* du langage humain par l'homme, le Traitement Automatique des Langues Naturelles (TALN) en est l'application informatique. Cette discipline étudie

en effet un ensemble d'approches ou de techniques permettant de modéliser, d'analyser et d'interpréter le langage humain. Elle est de ce fait multidisciplinaire réunissant des psychologues, documentalistes, lexicographes, traducteurs, logiciens, et *a fortiori* des informaticiens et linguistes. L'une des principales forces du TALN est sa considération des données textuelles. Les textes ne sont en effet pas réduits à une succession de mots mais sont considérés d'un point de vue linguistique. La théorie de la linguistique cherchant à représenter ou modéliser les connaissances linguistiques peut se diviser en plusieurs sous-domaines que sont la phonologie, la morphologie, la sémantique et la syntaxe. Par héritage, le TALN utilise les connaissances apportées par ces sous-domaines en s'intéressant à différentes tâches comme la traduction précédemment évoquée mais également *le résumé automatique de textes, la synthèse de la parole, la reconnaissance vocale, la classification automatique de documents textuels, la recherche documentaire*, etc.

Comment pouvons-nous alors représenter un document textuel afin de pouvoir au mieux découvrir et exploiter les connaissances qu'il contient. C'est là le rôle du descripteur. Nous pouvons définir des descripteurs de documents textuels comme **un ensemble de caractères** (et *a fortiori* de mots) **permettant de caractériser les documents**. Ainsi, en utilisant le "*mot*" en tant que descripteur, nous supposons qu'un document textuel est défini par les mots qui le composent. Ce descripteur est le plus utilisé dans la littérature afin de décrire un document. Il est en effet l'élément permettant de s'exprimer, d'échanger des informations, des sentiments, des pensées, etc. Cependant, d'autres descripteurs sont possibles afin de caractériser des données.
Nous pouvons (1) choisir un type de descripteur et (2) sélectionner les plus discriminants. Il existe dans le domaine de la fouille de données un certain nombre d'approches permettant de sélectionner des descripteurs. Nous proposons ci-dessous dans un premier temps de définir ces approches de sélection de descripteurs puis de montrer comment les connaissances de TALN peuvent également enrichir cette sélection.

1.1.2 De la fouille de données aux descripteurs de données textuelles

La fouille de données s'intéresse à la découverte d'informations utiles et nouvelles dans une quantité importante de données. Cette discipline comprend plusieurs processus dont la collecte de données, la suppression de bruits, la sélection de descripteurs puis une tâche de découverte de connaissances, visualisation et évaluation de ces dernières. La tâche de sélection de descripteurs à laquelle nous nous intéressons dans ce mémoire est une étape essentielle en fouille de données. Diverses approches permettent en effet de choisir quelles vont être les entités décrivant au mieux les données traitées. Une image peut par exemple

être décrite par les pixels qui la composent, en fonction de leurs couleurs, de leurs textures, etc. Une donnée sonore comme la voix humaine peut quant à elle être décrite selon sa fréquence pour distinguer par exemple une voix féminine d'une voix masculine. Se pose alors la notion de sélection des descripteurs de données, qui sont la plupart du temps de nature statistique. Nous les représentons en trois catégories (telles que détaillées dans [Saeys *et al.*, 2007]) spécifiées ci-dessous.

1. Les sélections de descripteurs à base de filtres (*Filter*).
2. Les sélections englobant les tâches qu'elles tentent de résoudre (*Wrapper*) [Kittler, 1978], [Holland, 1975].
3. Les approches dites *Embedded* [Duda *et al.*, 2001]. Ces dernières sont cependant spécifiques à des tâches de classification car dépendantes de classificateurs, notion sur laquelle nous reviendrons au cours de ce mémoire.

Ce mémoire porte cependant sur la sélection de descripteurs de documents textuels dans le but de décrire des thématiques plus ou moins spécialisées. Ainsi, les tâches de fouille de données peuvent être appliquées aux données textuelles et peuvent également bénéficier des connaissances et des méthodes issues du TALN. De telles tâches peuvent être qualifiées de tâches de "fouille de textes". Notons néanmoins que la nature des données textuelles rend parfois les applications de fouilles de données incompatibles et *vice versa*. La tâche de classification conceptuelle est par exemple propre à la fouille de textes et n'a pas d'équivalents en fouille de données. D'autres étapes d'un processus de fouille de données sont compatibles avec la fouille de textes. Par exemple, le domaine de la fouille de données comprend les tâches de pré-traitements des données souvent associées au nettoyage de corpus en fouille de textes bien que la notion de bruit soit différente. Comment par exemple transposer le concept de fautes d'orthographes en fouille de données ?
La fouille de textes se décompose selon les tâches suivantes : pré-traitements de corpus, choix du type de descripteurs, sélection des descripteurs et regroupement de ces derniers sous forme de concepts. Enfin la découverte de règles d'association ou la construction de patrons d'extraction peuvent représenter des applications finales d'un processus de fouille de textes. La sélection de descripteurs de données textuelles est alors complexe car devant tenir compte des spécificités de la langue. C'est en ce sens que les techniques de TALN peuvent être utilisées pour mieux décrire les documents textuels. Certes, de nombreuses approches utilisent une sélection de descripteurs via des méthodes statistiques à l'instar des approches de fouilles de données. Cependant, d'autres types de sélection spécifique au TALN peuvent être effectués. Par exemple nous pouvons sélectionner des descripteurs de données textuelles fondées sur les caractéristiques morphosyntaxiques des mots ou bien encore utiliser des connaissances extérieures comme des thésaurus sur lesquels nous reviendrons dans ce mémoire.

1.1.3 Thèse défendue : l'apport de l'information syntaxique à la sélection de descripteurs

Bien que les ressources du TALN soient parfois utilisées pour la tâche de sélection de descripteurs, l'information syntaxique contenue dans des données textuelles est rarement employée afin de sélectionner des descripteurs pertinents. Nous trouvons le plus souvent des approches se fondant sur les propriétés lexicales, c'est-à-dire sur la présence ou l'absence d'un mot dans un document, en tenant compte ou non de sa fréquence. Nous proposons ainsi de définir des méthodes de sélection plus complexes se fondant sur l'information syntaxique contenue dans un corpus.

La syntaxe consiste en l'étude de séquences de mots formant des unités syntaxiques, tels que les phrases ou les groupes de mots[1] [Riegel *et al.*, 1999]. Elle identifie également les liens ou relations qu'entretiennent ces différents mots ou groupes de mots entre eux. Ces liens sont nommés "relations syntaxiques". Ainsi, utiliser ces connaissances permet de prendre en compte l'ordre des mots et de désambiguïser lexicalement et sémantiquement des mots ou groupes de mots. D'un point de vue lexical, le mot "souris" sera identifié comme un nom ou une forme verbale de "sourire". Notons que cet exemple montre également une forme de désambiguïsation sémantique. Citons également le mot "avocat" qui est assez polysémique. Dans une relation syntaxique de type "verbe-objet", la sémantique du terme est alors précisée. Ainsi, le sens de ce dernier est très différent s'il est employé avec le verbe "manger" ou "appeler". Cet exemple montre que les informations syntaxiques sont cruciales pour découvrir des connaissances sémantiques ou pour désambiguïser un mot.
Nous présentons dans ce mémoire deux approches visant à extraire des descripteurs se fondant sur la syntaxe et permettant de découvrir des informations sémantiques. Deux méthodes sont proposées du fait de la nature des relations syntaxiques utilisées afin de décrire un corpus. Les premières sont des relations syntaxiques pouvant être extraites avec des analyseurs standards. Le second type de relations syntaxiques, les relations dites "induites", se révelent plus complexe à obtenir.

1.1.4 Les tâches nécessitant des descripteurs

Faisant partie du processus de fouille de textes, la sélection de descripteurs est une étape essentielle d'un certain nombre de tâches plus ou moins connexes. Toute application utilisant en effet les ressources d'un corpus peut bénéficier d'une sélection de descripteurs

[1] Ces groupes de mots peuvent être des syntagmes, des locutions, etc. Nous reviendrons sur ces notions en section 2.1.2.2.

appropriés. Nous listons ci-dessous un certain nombre de tâches et applications pour lesquelles la notion de descripteurs est cruciale.

– *La recherche documentaire (moteur de recherche).* D'une manière générale toute tâche d'indexation comprend une étape de sélection de descripteurs ou index, décrivant les données à indexer. En recherche documentaire, les requêtes soumises sous forme de mots clés vont être comparées aux descripteurs afin de satisfaire un utilisateur.

– *La classification.* La classification de données textuelles est une des premières tâches introduisant la sélection de descripteurs. En effet, des documents mieux décrits sont des documents plus discriminants dans un corpus. Ainsi, identifier leurs classes est une tâche plus aisée.

– *L'extraction d'informations.* Cette tâche consiste, comme son nom l'indique, à extraire de l'information à partir notamment de corpus spécialisés. Ces informations peuvent alors permettre de "nourrir" des modèles décrivant les connaissances du domaine. Ces modèles peuvent être des formulaires ou des bases de données. La sélection de descripteurs peut jouer un rôle déterminant pour cette tâche.

Bien d'autres tâches peuvent bénéficier d'une sélection de descripteurs dont des applications propres au TALN. Citons par exemple le *résumé automatique* de données textuelles où l'identification de descripteurs en tant que termes clés d'un corpus permettant de mettre en relief une thématique. Nous reviendrons dans ce mémoire sur le bénéfice apporté par les descripteurs suivant le type de tâches effectuées. Nous présentons ci-dessous l'organisation de ce manuscrit.

1.2 Organisation du manuscrit

La section suivante de ce mémoire propose un état de l'art sur d'une part les types de descripteurs et d'autre part la sélection de ces derniers. Ces deux notions sont en effet à distinguer. Rappelons de ce fait que nous présentons dans ce manuscrit des méthodes de sélection de descripteurs et non pas un nouveau type de descripteurs. Dès lors, nous effectuons un rapide survol des approches permettant de traiter des données textuelles d'un point de vue informatique. La représentation la plus répandue, consistant en la projection des descripteurs dans un espace vectoriel, sera abordée plus spécifiquement. Nous spécifierons alors les différentes tâches dans lesquelles les descripteurs jouent un rôle déterminant, en mettant l'accent sur le domaine de la classification qui constitue

une part non négligeable de nos expérimentations.

Le chapitre 3 présente notre premier modèle de sélection de descripteurs : SelDe. Ce chapitre décrit dans un premier temps l'analyseur syntaxique sur lequel nous nous fondons pour la sélection de descripteurs : l'analyseur Sygfran. Nous évoquerons par la même un certain nombre d'analyseurs syntaxiques usuels et en présenterons les caractéristiques par rapport à Sygfran. Nous introduirons également une approche sur laquelle nous nous sommes appuyés afin de définir la proximité sémantique de relations syntaxiques : le système Asium. Les auteurs de ce système ont défini une mesure de similarité éponyme que nous avons sélectionnée pour notre modèle. Nous la comparerons alors à certaines mesures de la littérature afin d'en montrer la qualité et le comportement. Nous détaillons alors les fondements du modèle SelDe une fois les principaux outils nécessaires à sa mise en place définis. Ce modèle repose sur l'utilisation de relations syntaxiques sémantiquement proches afin de décrire un corpus.

Les expérimentations effectuées avec les descripteurs du modèle SelDe sont présentées dans le chapitre 4. Nous présentons dans ce dernier une méthode d'enrichissement de corpus nommée *ExpLSA* se fondant sur le modèle SelDe. Cette dernière est évaluée dans le cadre de tâches de classification conceptuelle et classification de textes.

Le chapitre 5 s'intéresse aux données textuelles dites complexes. De telles données ne pouvant être décrites par le modèle SelDe, il fut alors nécessaire de proposer une alternative afin de les traiter. Ce chapitre aborde trois types de données complexes, ayant fait l'objet d'expérimentations dans le contexte d'encadrements de stages et de collaborations. Les données étudiées sont des articles issus de blogs, des CV ou encore des textes bruités issus de rétro-conversions OCR.

Nous revenons sur les descripteurs de nature syntaxique en présentant dans le chapitre 6 un modèle plus riche : SelDeF. Ce dernier permet l'utilisation de relations syntaxiques plus évoluées, du fait qu'elles ne soient pas initialement présentes dans un corpus mais elles sont en quelque sorte construites à partir de celui-ci. Cependant, ces relations ne sont pas toujours de qualité et doivent être filtrées afin de décrire un corpus. Ainsi, nous proposons dans ce chapitre des approches validant ces relations, en nous fondant sur des ressources sémantiques externes et sur le Web en utilisant des mesures statistiques.

Nous abordons dans le chapitre 7 une nouvelle tâche afin d'exploiter les descripteurs de SelDeF. Nous proposons en effet une méthode de construction de classes conceptuelles et deux méthodes visant à les enrichir. La première utilise les connaissances

de SelDeF en sélectionnant des descripteurs précis pour enrichir les classes. La seconde est une approche originale se fondant sur la notion d'énumération des ressources du Web.

Nous reviendrons finalement dans le chapitre 8 sur les différentes propositions de ce mémoire en mettant en relief les contributions essentielles. Ces travaux ouvrent de nouvelles pistes et réflexions qui pourront être menées dans la continuité du travail réalisé au cours de cette thèse.

Chapitre 2

État de l'art sur les descripteurs de textes et leur utilisation

Sommaire

2.1 Le choix du descripteur

Dans un processus de fouille de textes, un descripteur, tel que précédemment explicité est employé en tant que constituant d'un corpus. La figure 2.1 présente une manière de situer le descripteur dans un tel processus. Ainsi, une première tâche afin d'utiliser

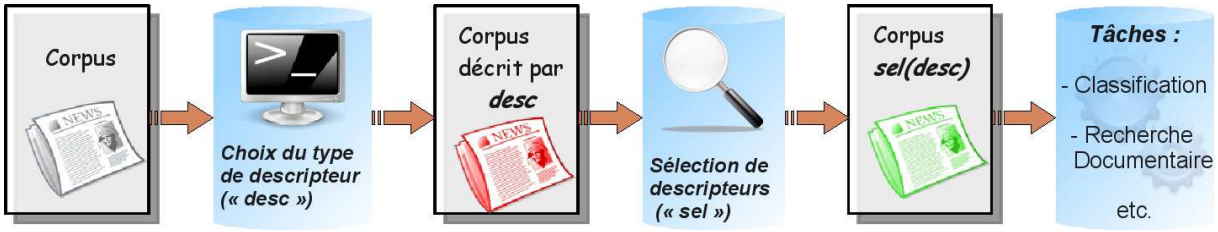

Fig. 2.1 – Choix et sélection de descripteurs dans un processus de fouille de textes.

un descripteur consiste à définir son type. Nous présentons ci-dessous quelles peuvent être ces définitions. Une fois le type de descripteur défini, il peut être également utile de sélectionner les descripteurs les plus représentatifs d'un corpus. Ainsi, plusieurs approches de la littérature permettent ce type d'extraction. Nous en présentons quelques unes dans la section 2.1.2.

2.1.1 Les types de descripteurs

2.1.1.1 Le mot ou lemme

Un corpus, écrit en langue naturelle indo-européenne, est constitué de *mots* afin de le décrire. Un mot peut être défini comme un son ou bien un ensemble de sons exprimant une *sensation*, une *conception* ou encore une *représentation*. Nous distinguons alors deux types de mots : les variables et invariables. Ces derniers peuvent être des adverbes, interjections, conjonctions ou prépositions. Nous nous focalisons sur les mots variables pouvant être des *adjectifs*, *substantifs* (ou *noms*), *articles*, *pronoms* et *verbes*. Les mots variables ont la propriété de pouvoir être déclinés ou conjugués (dans le cas de langues indo-européennes). Nous parlons alors de forme fléchie du mot. Notons que le "*mot*" tel que nous l'avons défini peut également se nommer *lemme*. Les lemmes sont en d'autres termes les entrées de dictionnaires.
Un exemple de mot, ou forme lemmatisée de ce mot peut être le verbe "*faire*". L'intérêt de ce type de descripteurs est double. Il peut dans un premier temps réduire de manière non négligeable le nombre de descripteurs utilisés pour par exemple une tâche d'apprentissage. Les lemmes peuvent également permettre d'associer des termes ayant une sémantique commune.

2.1.1.2 La forme fléchie

Outre le lemme, un mot peut être représenté sous différentes formes. En effet, la nature flexionnelle des langues indo-européennes introduit la notion de flexion comme nous l'évoquions ci-dessus. La forme *fléchie* d'un mot ou *flexion* est sa représentation "usuelle". Ainsi, il paraît naturel de décrire un corpus avec ce type de descripteurs issus du lemme. Nous distinguons deux types de flexions, les verbales ou *conjugaisons* propres aux verbes et les nominales ou *déclinaisons* propres aux noms mais également aux adjectifs, articles et pronoms. Les formes fléchies sont caractérisées par un certain nombre de *traits morphologiques* pouvant être le genre, le nombre, le temps, le mode, etc. en fonction du type de flexion. Elles sont également caractérisées par un lemme. Ainsi, en définissant une flexion avec le lemme "faire" et les traits "subjonctif présent, première personne du pluriel", nous obtenons la flexion "*fassions*".

2.1.1.3 Le radical

Une flexion peut être décomposée en deux entités, un radical et un (ou des) affixe(s). Le radical a des avantages similaires au lemme en ce sens qu'il réduit le nombre de descripteurs et permet de rapprocher des termes ayant une base commune, mais il le fait de manière moins systématique. En reprenant la flexion précédente *fassions*, le radical sera "*fass*".

Notons que ce descripteur est traduisible en anglais par le terme *stem* tel qu'employé par exemple dans [Lovins, 1968]. Ce dernier est cependant très souvent confondu avec le lemme (*lemma* en anglais) qui peut se résumer à la forme canonique d'un mot et non pas à son radical. Il a notamment été introduit dans le domaine de la recherche d'information par [Porter, 1980].

2.1.1.4 Les descripteurs phonétiques

Les descripteurs phonétiques sont identifiés par des syllabes, responsables des sons. La définition de "syllabe" peut être : structure de type consonne+voyelle. Cependant, cette notion est ambigüe et a été discutée dans la littérature. Citons par exemple [Laueufer, 1992] et [Pallier, 1994] qui proposent un certain nombre de définitions permettant de caractériser une syllabe.
Bien que peu utilisés dans le cadre d'application typiquement TAL, ces descripteurs sont surtout employés pour des approches de synthèses vocales comme le montrent [Dutoit, 1997] et [Bagein *et al.*, 2001]. Les approches de synthèse vocale ne se limitent pas aux descripteurs phonétiques et sont souvent des méthodes complexes comme c'est le cas du système MARY TTS présenté dans [Schröder *et al.*, 2003] qui traite l'allemand.
Les descripteurs qui ont été présentés jusqu'ici ont des avantages certains mais ne permettent pas de résoudre le cas de termes polysémiques comme le mot "livre", ayant un certain nombre de sens distincts. Un type de descripteur peut permettre de lever en partie les ambigüités sémantiques : les n-grammes qui sont présentés ci-dessous.

2.1.1.5 Les n-grammes

La notion de n-grammes et plus particulièrement bi-grammes et tri-grammes (c'est-à-dire avec respectivement n=2 et n=3) est apparue à l'origine dans [Pratt, 1939] selon [Shannon, 1948]. Ce dernier introduisit la notion de n-grammes dans le cadre de systèmes de prédiction de caractères en fonction des autres caractères précédemment entrés. Nous pouvons définir un *n-gramme* de X comme une séquence de n X consécutifs. X peut alors être un *caractère* ou bien un *mot*.
La figure 2.2 illustre la construction de n-grammes de caractères et de mots par la notion de déplacement de fenêtre. Ce déplacement se fait par étape, une étape étant soit un caractère ou bien un mot. Notons que les n-grammes (de mots et de caractères) sont construits à partir de flexions. Les caractères (ou mots) contenus dans la fenêtre ainsi définie constituent les descripteurs d'un corpus. Nous avons par exemple dans la figure 2.2 les descripteurs de la phrase sous forme de bi-grammes de mots qui sont : "Le choix", "choix du" et "du descripteur".
Nous présentons ci-dessous les deux types de n-grammes, caractères et mots, sous leurs

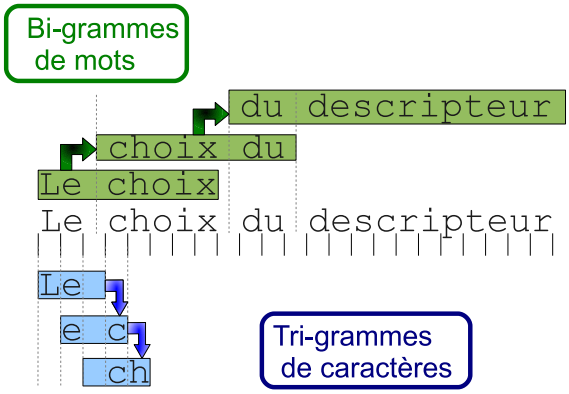

FIG. 2.2 – Exemple de N-grammes de mots et de caractères.

formes fléchies.

Les N-grammes de caractères

Les n-grammes de caractères sont les premiers à avoir été utilisés pour une tâche utilisant des données textuelles [Shannon, 1948]. La notion de n-grammes "seule" désigne en effet les n-grammes de caractères. Ce type de n-grammes est principalement utilisé dans l'identification de la langue ou encore la recherche documentaire sur laquelle nous reviendrons dans la section 2.3.3. Nous noterons que les n-grammes de caractères prennent en considération les espaces. En effet, il est assez trivial de montrer que la non prise en compte des espaces introduit du bruit, en considérant des mots qui n'existent pas.

Ce type de descripteurs a plusieurs avantages parmi lesquelles la non nécessité d'employer des descripteurs de type "*radical*". En effet, la description d'un corpus par les n-grammes de caractères prend automatiquement en compte les racines des mots les plus fréquents [Grefenstette, 1995]. L'utilisation de n-grammes de caractères introduit également la notion d'indépendance de la langue comme le montre [Dunning, 1994]. Notons pour finir que les n-grammes de caractères sont tolérants aux fautes d'orthographes et au bruit pouvant être causé par exemple par l'utilisation de numérisation de documents par OCR[2] (Optical Character Recognition). Par exemple [Miller *et al.*, 2000] expérimentent la robustesse de certains systèmes de recherches documentaires en utilisant des n-grammes de caractères. Ils montrent que les systèmes conservent leurs performances avec un taux de dégradation des données de l'ordre de 30%. Par ailleurs, ces descripteurs peuvent également permettre de détecter la langue d'un corpus comme dans les travaux de [Shen *et al.*, 2006].

[2]Nous reviendrons sur ce type de données dans le chapitre 5 de ce mémoire, consacré aux descripteurs adaptés aux données textuelles complexes.

Les N-grammes de mots

L'utilisation de n-grammes de mots est plus récente que l'utilisation de n-grammes de caractères dont par exemple [Solso, 1979] fut une des premières publications sur le sujet. Ce type de descripteurs est principalement employé dans le cadre de classification automatique de données textuelles[3] comme dans les travaux de [Fürnkranz, 1998]. Les n-grammes de mots ont l'avantage de désambigüiser des mots composés comme "carte de crédit" qui sera considéré comme un seul descripteur avec des tri-grammes de mots. Ces descripteurs sont particulièrement utilisés pour des tâches de modélisation de langage comme dans [Wang & Vergyri, 2006]. Citons également [Lei & Mirghafori, 2007] qui emploient les n-grammes de mots dans le cadre de reconnaissance d'interlocuteurs téléphoniques.

2.1.2 La sélection de descripteurs

Après avoir défini les différents types de descripteurs pouvant être employés afin de représenter un corpus, nous nous intéressons dans cette section à la sélection de descripteurs. En effet, bien que des descripteurs comme les lemmes ou les n-grammes soient généralement pertinents, il peut être nécessaire de ne sélectionner que les "meilleurs" descripteurs d'un corpus. Nous distinguons dans la littérature trois modes de sélection :

– Les approches statistiques
– Les approches se fondant sur des connaissances morphosyntaxiques
– L'utilisation de modèles de connaissance externes

2.1.2.1 Les approches statistiques

La sélection statistique de descripteurs est le type d'approche le plus répandu. Elle consiste à employer des mesures statistiques afin de donner un score de qualité à un descripteur. Ainsi, seuls les n premiers descripteurs seront conservés afin de décrire le corpus. Une première approche pourrait se fonder sur la loi de Zipf [Zipf, 1941] décrite en section 2.2.3.1 qui s'appuie sur le nombre d'occurrences d'un terme dans un document. Ainsi, nous pouvons considérer par exemple comme pertinents uniquement les descripteurs ayant un minimum de 3 occurrences et un maximum de 8. Notons que cette notion sera abordée dans le chapitre 3 dans lequel nous présentons notre premier modèle de sélection de descripteurs SELDE. Ce modèle se fonde entre autres sur la notion d'occurrence en proposant des paramètres précis de sélection de descripteurs.

[3] Nous reviendrons sur cette notion dans la section 2.3.1.

Nous présentons ci-dessous quelques mesures de sélection de descripteurs fréquemment employées dans la littérature. Ces mesures sont issues de la théorie de l'information et sont applicables à une tâche donnée. Nous nous référons dans cette section à la tâche de classification automatique avec apprentissage supervisé, détaillée dans la section 2.3.1.2. Nous définissons la tâche de classification de documents par le fait d'attribuer à un document donné (comme un paragraphe ou un texte) une catégorie (pouvant être un thème, un type, etc.). Avant de présenter les approches de sélection statistique de descripteurs, nous proposons de définir une *variable aléatoire* au sens statistique, objet de ces mesures. Un phénomène aléatoire peut se traduire par une "grandeur" mathématique représentée par un nombre réel ou entier. Citons par exemple le pourcentage de réponse "oui" à un sondage ou le nombre d'enfants d'un couple qui peuvent être représentés sous forme de variables aléatoires. De manière plus formelle, soit un univers Ω et ω un événement élémentaire de Ω. Nous appelons alors une variable aléatoire de l'univers Ω toutes applications X dans $\Re$ notées $X : \omega \mapsto X(\omega) \in \Re$. Nous définissons une variable aléatoire comme "discrète" (par opposition à "continue") lorsqu'elle ne prend que des valeurs discontinues dans un intervalle donné. Nous appelons finalement une loi ou distribution de probabilité d'une variable aléatoire discrète les valeurs de la probabilité P telle qu'une variable aléatoire discrète X soit égale à une quantité x_i, notée $P(X = x_i)$ par convention. Afin de présenter un exemple concret de sélection de descripteurs, nous présenterons ces mesures appliquées à une tâche de classification, en nous référant notamment à [Yang & Pedersen, 1997].

L'information mutuelle

L'information mutuelle [Fano & Hawkins, 1961] est une mesure permettant de mesurer la dépendance statistique entre deux variables aléatoires. Dans le cas discret, cette mesure se définit comme suit entre deux variables aléatoires X et Y et leurs distributions de probabilités respectives $P(x)$ et $P(y)$, et jointes $P(x, y)$:

$$I(X, Y) = \sum_{x,y} P(x, y) log \frac{P(x, y)}{P(x)P(y)} \tag{2.1}$$

Adaptée à la sélection de descripteurs pertinents, les variables aléatoires sélectionnées sont alors dépendantes de la tâche à effectuer. Par exemple, dans le cadre de classification de documents, nous allons mesurer la dépendance d'un descripteur en fonction d'un document. Soient un descripteur d et une classe c, l'information mutuelle se définit comme suit.

$$I(d, c) = log \frac{P(d \wedge c)}{P(d)P(c)} \tag{2.2}$$

avec $P(d \wedge c)$ pouvant être estimée par la quantité $A \times N$, $P(d)$ estimée par la quantité $A + C$ et $P(c)$ estimée par la quantité $A + B$ telles que :
– A = nombre de fois où d apparaît dans un document de classe c.
– B = nombre de fois où d apparaît dans une autre classe que c.
– C = nombre de documents de classe c sans le descripteur d.
– N = nombre total de documents.
Ainsi, non pouvons estimer la qualité d'un descripteur d d'un corpus en combinant le score d'un descripteur obtenu par les formules suivantes :

$$IM_{moyen}(d) = \sum_{i=1}^{m} P(c_i) IM(d, c_i) \tag{2.3}$$

$$IM_{max}(d) = max_{i=1}^{m} \{IM(d, c_i)\} \tag{2.4}$$

Plusieurs approches peuvent alors être utilisées afin de sélectionner les descripteurs. Nous pouvons par exemple retenir ceux ayant un score d'information mutuelle variant de la moyenne au maximum pour une classe donnée. Pour une tâche de classification, les descripteurs sélectionnés vont alors permettre de construire une base d'apprentissage plus riche.

χ^2 (CHI carré ou CHI deux)

La mesure statistique du χ^2 s'inspire de la loi éponyme qui serait due à Ernst Abbe en 1863 selon [Sheynin, 1977]. De manière formelle, cette loi est associée aux variables aléatoires distribuées selon une loi *normale* (cf. [Dodge, 2007] pour plus de détails sur les définitions des approches statistiques nommées ici) centrée réduite (de moyenne égale à 0 et de variance = 1). Soient Z_1, Z_2, ...,Z_n, n variables aléatoires centrées réduites indépendantes. La somme du carré de ces variables aléatoires suit alors une distribution selon une loi du χ^2 avec n degrés de liberté :

$$\chi^2 : Z_1^2 + Z_2^2 + ... + Z_n^2 = \sum_{i=1}^{n} Z_i^2 \tag{2.5}$$

Adaptée à notre problématique de **sélection de descripteurs**, nous utilisons cette loi dans le cadre du **test d'indépendance du** χ^2. Le principe est de déterminer si deux variables aléatoires étudiées sur un certain échantillon sont indépendantes en mesurant

l'écart obtenu entre une distribution indépendante et la distribution réelle. Soient n individus d'une population P sélectionnés aléatoirement et les définitions de présences et absences des variables aléatoires A et B (qui sont des caractéristiques des individus n) suivantes :
– $P(A \cap B) = P_{11}$ la proportion d'individus répondant aux caractéristiques A et B
– $P(\overline{A} \cap B) = P_{21}$ = la proportion d'individus répondant aux caractéristiques B mais pas à celles de A
– $P(A) = P_1$ = la proportion d'individus répondant aux caractéristiques A
etc.
Dès lors, la mesure d'indépendance de ces deux caractéristiques A et B du χ^2 est définie comme suit :

$$\chi^2 = n \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(P_{ij} - P_i.P_j)^2}{P_i.P_j} \tag{2.6}$$

Adaptée à un problème de classification de textes, nous trouvons dans [Yang & Pedersen, 1997] la formule suivante pour un descripteur d et une classe c :

$$\chi^2(d, c) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \tag{2.7}$$

avec A, B, C et N les quantités précédemment évoquées et D correspondant au nombre de documents n'appartenant pas à la classe c et ne contenant pas le descripteur d. Nous calculons dès lors les mêmes quantités moyennes et maximums de manière similaire à l'IM afin d'estimer la qualité d'un descripteur.

La force d'un descripteur

Une dernière mesure, la force d'un descripteur ou terme ("term strength" en anglais) fut originalement utilisée dans le cadre de sélection de vocabulaire pertinent par [Wilbur & Sirotkin, 1992]. Cette approche estime l'importance d'un descripteur en se fondant sur la manière dont ce dernier apparaît habituellement dans des documents "étroitement liés". Cette mesure nécessite un apprentissage préalable afin d'estimer la proximité de documents (mesurée par une représentation vectorielle en utilisant

une mesure de similarité, le *cosinus*[4]). Dès lors, la force d'un descripteur se fonde sur l'estimation des probabilités qu'un descripteur apparaisse dans la seconde moitié d'une paire de documents connexes, sachant qu'il n'apparaît pas dans la première moitié. Soient x et y une paire de documents consécutifs et d un descripteur, alors la force f de ce descripteur est la suivante :

$$f(t) = P(d \in y | d \in x) \tag{2.8}$$

Il existe de nombreuses autres mesures statistiques permettant la sélection de descripteurs comme le gain d'information, des mesures d'entropies, etc. décrites par exemple dans [Mitchell, 1997]. Notre objectif fut de présenter dans cette section des approches issues de différents domaines sans être nécessairement exhaustif. D'autres types de méthodes permettent également de sélectionner des descripteurs comme la sélection utilisant des informations morphosyntaxiques telles que présentées ci-dessous.

2.1.2.2 Sélection morphosyntaxiques

D'une manière générale, un corpus est décrit par le vocabulaire de son corpus. Cependant, un certain nombre de ces descripteurs peuvent apporter du bruit et dégrader la qualité d'un corpus. Nous proposons dans cette section de présenter des approches de sélection de descripteurs se fondant sur des informations morphosyntaxiques nous donnant par exemple la catégorie lexicale d'un descripteur. Notons que cette sélection ne peut s'appliquer qu'avec des descripteurs de langues naturelles du type *flexion*. En effet, si nous prenons le cas de n-grammes, la notion de catégorie lexicale n'a pas de sens. Nous présentons dans un premier temps une approche de sélection de descripteurs en fonction de leurs catégories lexicales.

Les connaissances lexicales

L'utilisation de connaissances lexicales, nous informant sur la catégorie d'un mot[5], va permettre une sélection en fonction de l'importance des catégories lexicales d'un corpus. Par exemple, les noms (ou substantifs) sont des mots susceptibles de décrire au mieux un concept spécifique [Poudat *et al.*, 2006]. Les auteurs présentent dans cet article une méthode

[4]Ces notions d'apprentissage, de représentation vectorielle et de similarité seront définies dans les sections suivantes.

[5]Nous parlerons dans cette section de "mots" afin de nommer un descripteur de type flexion.

de classification de documents textuels de discours scientifiques. Ils se fondent sur l'association de descripteurs morphosyntaxiques, les substantifs, ainsi qu'à un certain nombre de descripteurs qu'ils qualifient de "caractéristiques" du discours scientifique comme des abréviations ou encore des ponctuations, des acronymes, etc. Leurs expérimentations ont montré la qualité de ces descripteurs. D'autres travaux de la littérature comme ceux de [Kohomban & Lee, 2007] montrent également que les noms sont des descripteurs de qualité. Citons également [Benamara *et al.*, 2007] qui montrent que les adjectifs et parfois les adverbes sont assez adaptés aux données d'opinions. Notons qu'un certain nombre de combinaisons et pondérations peuvent être utilisées avec ce type de descripteurs. Nous reviendrons dans ce mémoire sur ces approches dans le chapitre 5 traitant du problème de la sélection de descripteurs propres aux données textuelles dites "complexes".

Les connaissances syntaxiques

Outre les connaissances lexicales, nous pouvons également sélectionner des descripteurs par des approches utilisant les informations syntaxiques d'un corpus. Le principe est assez similaire à l'approche précédente en ne conservant uniquement les descripteurs comme des syntagmes ou des relations syntaxiques. Nous définissons ci-dessous ce type de descripteurs. Notons que l'obtention de ce type de descripteurs nécessite le plus souvent une analyse syntaxique là où les précédents descripteurs pouvaient être obtenus par le biais d'étiqueteurs grammaticaux. Nous reviendrons sur ces notions dans les sections 3.2.2.2 (analyseurs) et 5.2.2 (étiqueteurs).

Les syntagmes

La sélection de *syntagmes* est une extension logique de la sélection de catégories lexicales pouvant être des noms, des verbes, des adverbes, etc. En effet un syntagme peut se définir comme un groupe de mots formant une unité lexicale par son sens et par sa fonction. Un syntagme est formé d'un noyau et de satellites. Le noyau est l'élément qui va définir la catégorie lexicale du syntagme. Par exemple, dans le syntagme "*une jolie petite maison*", le noyau est le nom *maison*. Nous parlons alors de *syntagme nominal*. Une description plus complète de la notion de syntagme peut être trouvée dans [Bouquiaux, 1987]. Les syntagmes peuvent être obtenus par le biais de *patrons syntaxiques*, eux mêmes issus du domaine de l'extraction d'information et plus précisément de l'extraction de la terminologie.

La terminologie d'un corpus se définit comme l'ensemble des termes "techniques" décrivant le plus significativement le domaine du corpus. Les méthodes permettant d'extraire de la terminologie sont fondées sur des approches numériques ou linguistiques. Citons par exemple Termino [David & Plante, 1990] et Lexter [Bourigault, 1994] qui se fondent sur des méthodes linguistiques. Citons par ailleurs Mantex [Frath *et al.*, 2000] et Ana [Enguehard, 1993], [Enguehard, 2001] s'appuyant sur des outils numériques. Finalement,

les approches les plus abondantes dans la littérature sont mixtes, utilisant des méthodes numériques avec des ressources linguistiques. Citons par exemple ACABIT [Daille, 1994], EXIT [Roche *et al.*, 2004] ou encore SYNTEX [Bourigault & Fabre, 2000] qui fait suite à l'approche LEXTER.
Notons que chaque mot constituant un syntagme est dissociable. Un groupe de mots non dissociable est appelé un *mot composé*, formant ainsi un lemme à part entière (comme par exemple "*après-midi*"). Un type particulier de mot composé, appelé une *locution* est défini comme un mot composé contenant au moins un espace. Il s'agit la plupart du temps de syntagme qui se sont figés dont les mots ne sont plus dissociables comme la locution "*pomme de terre*". Dans cet exemple, le sens du syntagme ne peut être déduit du sens de "pomme" et de "terre" pris séparément.
Les syntagmes sont assez utilisés dans le domaine de la classification de textes comme dans [Kongovi *et al.*, 2002] ou encore [Fei *et al.*, 2004]. Ces derniers proposent de construire des patrons à base de syntagmes afin de classifier des sentiments.

Les relations syntaxiques

La syntaxe peut se définir comme un ensemble de règles régissant les relations entre les descripteurs d'un corpus (pouvant être des mots ou des syntagmes). Ces relations de dépendances sont appelées des relations syntaxiques. Il existe plusieurs types de relations syntaxiques comme les relations "verbe-objet" ou "sujet-verbe". Ainsi, de la phrase "Je mange une pomme", nous pouvons extraire la relation *sujet-verbe* "sujet :Je, verbe :mange" et *verbe-objet* "verbe :mange, objet :une pomme". Une description détaillée des relations syntaxiques peut être trouvée dans [Bowers, 2001]. Les descripteurs de type relations syntaxiques ne sont pas employés en tant que tels dans la littérature. Ils sont cependant utilisés de manière connexe à d'autres approches dans différents domaines comme la biomédecine. Citons par exemple [Kim, 2008] qui utilise des relations syntaxiques afin de détecter l'interaction entre gènes et protéines. [Shen *et al.*, 2005] présentent par ailleurs une approche construisant des patrons fondés sur des relations syntaxiques afin de produire un système de réponse automatique à des questions.
Nous proposons dans ce mémoire une approche de sélection de descripteurs en se fondant sur l'utilisation de relations syntaxiques. Ainsi, ce type de descripteurs et son utilisation seront développés dans le chapitre 3 dans lequel nous présentons notre modèle.

Une dernière méthode de sélection de descripteurs est présentée ci-dessous. Elle se focalise sur l'utilisation de ressources sémantiques.

2.1.2.3 Sélection par des modèles de connaissances

Le principe des approches de sélections de descripteurs utilisant des modèles de connaissances est de ne sélectionner que certains descripteurs propres à un domaine. Ce type d'approche est souvent utilisé comme ressource supplémentaire afin d'effectuer une tâche sur des données textuelles. Notons que certaines des approches de sélection de descripteurs morphosyntaxiques et/ou statistiques évoquées précédemment peuvent permettre de construire de telles ressources. Cependant, nous ne nous intéressons pas ici à la construction de telles ressources mais à l'utilisation de celles-ci. Nous distinguons deux principaux types de modèles de connaissance : les thésaurus et les ontologies.

Les thésaurus

Un thésaurus est présenté par la norme *ISO 2788* de 1986 comme définissant "un vocabulaire d'un langage d'indexation contrôlé, organisé formellement de façon à expliciter les relations *a priori* entre les notions (par exemple relation générique-spécifique) ". En d'autres termes, un thésaurus contient un ensemble de lemmes d'une langue de spécialité (appelé un lexique). Ces lemmes sont décrits par un ensemble de relations sémantiques avec les autres lemmes du lexique (relation de synonymie, de traduction, hiérarchiques ou encore de règles d'associations). Ainsi, le thésaurus est lié à l'étude terminologique d'un domaine général ou spécialisé, d'une langue comme nous pouvons le voir par exemple dans [Knapen & Briot, 1999]. Les termes d'un thésaurus ainsi définis cherchent alors à décrire des concepts. Deux notions caractérisent les concepts définis par les thésaurus selon [Maniez, 1999].
(1) Ces concepts sont définis afin de faciliter l'interrogation de bases de données textuelles (souvent des fonds documentaires). Ce critère va ainsi définir le choix de descripteurs par rapport à d'autres.
(2) Les concepts d'un thésaurus sont dépendants des langues qu'ils décrivent ainsi que du discours décrit.
C'est en ce second point que d'autres types de représentations sémantiques peuvent se distinguer des thésaurus dont notamment les ontologies. L'un des thésaurus les plus utilisés dans la littérature est sans doute le thésaurus Roget [Roget, 1852] qui vise à décrire de manière générale la langue anglaise.

Les ontologies

Une autre ressource permettant de sélectionner des descripteurs est l'ontologie. Ce terme issus du domaine de la philosophie désigne un "discours sur l'être en tant qu'être". Repris en informatique, l'objectif de ces dernières est de décrire des concepts pouvant être des représentations mentales ou encore des catégories issues de la philosophie de la connais-

sance [Guarino, 1998]. Les concepts d'une ontologie sont définis au delà des langues et caractérisent davantage un domaine de spécialité. Ces concepts sont organisés hiérarchiquement. Notons que nous n'évoquerons pas ici les méthodes de construction d'ontologies qui possèdent une littérature abondante. L'ontologie est ici présentée en tant que ressources pouvant être employée afin de sélectionner des descripteurs. Un exemple d'application est donné dans [Verma *et al.*, 2007]. Les auteurs proposent une méthode de résumé automatique de documents médicaux en se fondant sur une ontologie afin de sélectionner les termes discriminants.
Notons pour finir qu'il existe un grand nombre de modèles de connaissances pouvant permettre de sélectionner des descripteurs comme la ressource WordNet[6] [Miller, 1985] visant à décrire l'anglais. Citons également la ressource terminologique européenne IATE[7] (InterActive Terminology for Europe) également très employée.
Après avoir montré un aperçu des différents types de descripteurs et leur sélection, nous proposons dans la section suivante de considérer la représentation vectorielle de ceux-ci, étape indispensable afin de pouvoir les exploiter numériquement.

2.2 Représentation vectorielle

Afin de permettre l'exploitation des descripteurs, nous devons fournir un modèle permettant de les représenter numériquement. Les documents textuels ont dans un premier temps été considérés comme une séquence de caractères codée informatiquement via un encodage de caractères de type *ASCII* (permettant 256 caractères différents) ou plus récemment *Unicode* (65 536 caractère permis). Cette dernière représentation permet de traiter entre autres des langues naturelles non alphabétiques comme le chinois. Rappelons cependant que nous nous intéressons dans ce mémoire uniquement aux langues alphabétiques et plus précisément au français.
Nous avons fait le choix d'utiliser des techniques numériques. D'autres approches existent, dont notamment des approches dites symboliques. Nous nous focalisons dans ce mémoire plus particulièrement sur la représentation vectorielle dont nous présentons les principales approches dans la section suivante. Ces dernières reposent sur les théories des espaces vectoriels que nous présentons ci-dessous.

2.2.1 Espaces vectoriels

Le modèle vectoriel est issu de la théorie des espaces vectoriels. Cette théorie, provenant de concepts issus de l'algèbre linéaire et de la géométrie s'est appliquée

[6] wordnet.princeton.edu
[7] http ://iate.europa.eu

à de nombreux domaines comme l'analyse factorielle de données (citons par exemple [Jolliffe, 1986]). Notre objectif dans cette section est de rappeler les fondements des espaces vectoriels sans pour autant en décrire précisément l'historique ni les principes mathématiques de manière formelle. Le lecteur intéressé par l'aspect historique de cette théorie pourra se référer à [Dorier, 1996] qui propose une rétrospective historique ainsi qu'une analyse des fondements de la théorie des espaces vectoriels. L'article de [Jordan, 1986] présente quant à lui une introduction plus théorique des espaces vectoriels.
Un espace vectoriel peut être vu comme une généralisation de l'espace vectoriel géométrique à 3 dimensions. Le nombre de dimensions d'un tel espace est alors le nombre maximal d'axes de coordonnées utiles afin de définir tous points de cet espace. La notion d'indépendance linéaire signifie alors la non dépendance des axes entre-eux. Cette notion est indispensable dans le fondement de l'algèbre linéaire et *a fortiori* de la théorie des espaces vectoriels. De manière formelle, un espace vectoriel se définit comme un ensemble d'éléments, appelés des *vecteurs*, munis de deux opérations internes particulières (l'addition au sens vectoriel et la multiplication par un scalaire). En sortie d'une telle opération se trouve systématiquement produit un vecteur (du fait que l'ensemble d'éléments caractérisant un espace vectoriel soit *fermé*).
De manière plus concrète et adaptée aux documents, un vecteur provenant d'un espace vectoriel peut se définir comme un ensemble de descripteurs textuels.
Nous décrivons ci-dessous divers types de modèles vectoriels, couramment utilisés dans la littérature afin de représenter un document sous forme numérique. Nous présentons en effet dans le chapitre 4 des expérimentations de classification automatique des données textuelles afin d'évaluer la qualité des descripteurs produits par notre modèle de sélection (présenté dans le chapitre 3).

2.2.2 Modèle vectoriel

L'idée d'une représentation de documents dans un espace vectoriel introduit la notion de matrice numérique. Cette dernière va être constituées de $m \times n$ cellules. En d'autres termes, elle possèdera n colonnes de dimension m ou bien m rangs de dimension n. Outre la facilité de l'approche afin de représenter des données, cette matrice va également permettre d'effectuer un certain nombre de calculs vectoriels.
Appliqué aux documents textuels, nous nous fonderons sur le principe de *sac de mots*. Ainsi, les approches de ce type considèrent les mots comme de simples séquences non ordonnées. Le principe général est de représenter un corpus, contenant un ensemble de documents formés de descripteurs (souvent le mot) par une matrice de type documents/descripteurs telle que représentée dans la figure 2.3.
Le principe est de définir chaque vecteur de descripteurs en fonction des documents dans

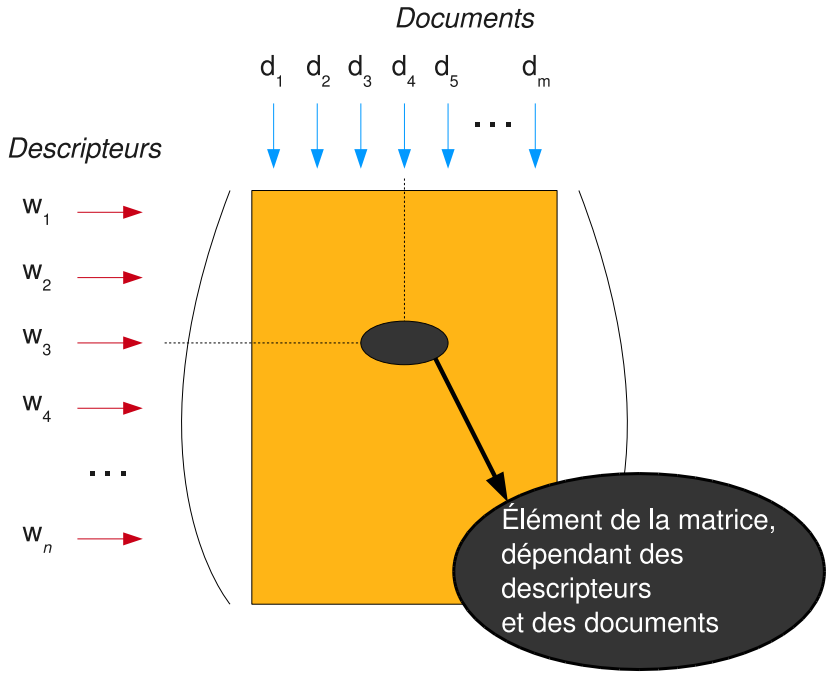

FIG. 2.3 – Représentation vectorielle d'un document.

lesquels ils apparaissent. La définition inverse peut également être utilisée suivant le type de besoin (cf section 2.3.1.1) à savoir un document est représenté par les descripteurs qu'il possède. Cette représentation est fréquemment nommée la représentation ou matrice de *Salton* et fut pour la première fois utilisée dans le cadre du système de recherche d'information *SMART* [Salton & Lesk, 1965]. Le modèle vectoriel de Salton est décrit précisément dans [Salton *et al.*, 1975b].

2.2.2.1 Booléen ou binaire

Le vecteur booléen figure parmi les plus anciennes représentations vectorielles de documents textuels. Une telle représentation considère seulement deux valeurs possibles pour les composantes d'un vecteur, 0 ou 1. Ces derniers chiffres correspondent respectivement à la présence ou l'absence d'un terme dans un document. Nous présentons ci-dessous un exemple de ce type de représentation. Nous utiliserons comme descripteurs le *lemme* et comme document la *phrase* et supprimons les mots outils, mots fonctionnels se répétant souvent comme des articles ou déterminants mais également des verbes trop communs comme *avoir* et *être*. En considérant les phrases suivantes :

- `Le chat mange la souris qui n'a pas eu le temps de manger son fromage.`
- `Le chat n'a plus faim et va rejoindre les autres chats.`
- `Le chien aboie après les chats et les souris mangent le fromage.`

La matrice correspondante avec une représentation booléenne est donnée dans le tableau

2.1.

	aboyer	*chat*	*chien*	*faim*	*fromage*	*manger*	*rejoindre*	*souris*	*temps*
document1	0	1	0	0	1	1	0	1	1
document2	0	1	0	1	0	0	1	0	0
document3	1	1	1	0	1	1	0	1	0

TAB. 2.1 – Représentation vectorielle booléenne.

Nous pouvons définir formellement cette représentation comme suit avec w_{ij} étant le poids d'une cellule pour le descripteur i et le document j et tf_{ij} étant le nombre d'occurrences de i dans j.

$$w_{ij} = \begin{cases} 1 & \text{si } tf_{ij} > 0 \\ 0 & \text{sinon} \end{cases} \tag{2.9}$$

Bien qu'assez simpliste, cette représentation fournit parfois des résultats plus performants que la représentation fréquentielle, notamment sur des corpus de critiques de films [Pang & Lee, 2002]. La représentation fréquentielle, tf_{ij}, est présentée ci-dessous.

2.2.2.2 Fréquentielle

La représentation dite fréquentielle, notée tf, est une extension assez naturelle de la représentation binaire en ne considérant plus uniquement la présence ou l'absence d'un descripteur dans un document mais en prenant en compte sa fréquence d'apparition. En d'autres termes, cela revient à considérer le nombre d'occurrences d'un terme i dans un document j. En reprenant les trois phrases de l'exemple précédent, nous obtenons la représentation du tableau 2.2.

	aboyer	*chat*	*chien*	*faim*	*fromage*	*manger*	*rejoindre*	*souris*	*temps*
document1	0	1	0	0	1	2	0	1	1
document2	0	2	0	1	0	0	1	0	0
document3	1	1	1	0	1	1	0	1	0

TAB. 2.2 – Représentation vectorielle fréquentielle.

Nous montrons par cet exemple l'intérêt de cette approche avec le terme lemmatisé *chat*. Ce dernier n'a en effet plus le même poids pour chaque document, contrairement à sa représentation binaire. Notons par ailleurs qu'une telle matrice est généralement normalisée afin d'éviter de défavoriser les documents les plus longs, contenant ainsi plus de termes

et possédant une norme plus importante que des documents plus courts. Cela influe notamment sur les calculs de produits scalaires sur lesquels nous reviendrons en section 2.2.5. Une représentation formelle de l'approche fréquentielle avec une norme euclidienne est donnée ci-dessous, avec $NbMot_j$ le nombre de mots distincts dans le document j.

$$w_{ij} = \frac{tf_{ij}}{\sqrt{\sum_{k=1}^{NbMot_j} tf_{kj}}} \tag{2.10}$$

Bien que cette dernière représentation soit efficace, elle ne permet pas de distinguer un mot fréquent dans tout le corpus d'un mot fréquent uniquement dans quelques documents. Les méthodes de pondération présentées ci-dessous proposent des solutions à ces limitations.

2.2.2.3 Représentations vectorielles par vecteurs d'idées

La notion de vecteurs d'idées est issue de la linguistique componentielle provenant entre autres des travaux de [Pottier, 1964], [Le Ny, 1979] et [Hjelmslev, 1968]. Cette dernière émet le principe qu'un terme peut être décrit par la combinaison des sens de termes plus généraux. Le principe des vecteurs d'idées en résultant est de projeter un espace ou "champ sémantique" dans un espace vectoriel. Plusieurs approches fondées sur des vecteurs d'idées ont été proposées par l'équipe TAL du LIRMM dont les vecteurs sémantiques [Chauché, 1990] qui seront développés dans la section 6.3.1, et deux types de vecteurs dits conceptuels [Lafourcade & Prince, 2001] et [Schwab, 2005]. Notons que les approches permettant de construire des vecteurs conceptuels ne sont pas utilisées dans ce mémoire. Elles sont toutes deux détaillées dans [Schwab, 2005]. Ainsi, nous présentons ci-dessous différentes techniques permettant de pondérer le poids de termes dans une représentation vectorielle de type *Saltonième*.

2.2.3 Pondérations statistiques

Les pondérations statistiques peuvent être vues comme des modèles de représentations vectorielles à part entière. Nous les nommons cependant pondérations statistiques du fait qu'elles pondèrent les poids des descripteurs utilisant des approches vectorielles. Notons également que de telles pondérations sont considérées parfois dans la littérature comme des approches de sélection de descripteurs pertinents de type statistique comme celles présentées en section 2.1.2.1.

2.2.3.1 Le tf-idf

L'approche la plus utilisée dans la littérature est sans doute le *tf-idf* [Salton & Yang, 1973], [Salton *et al.*, 1975a] pour *term frequency - inverse document fre-*

quency. Le principe de cette approche est de pondérer la méthode fréquentielle *tf* présentée précédemment par le nombre de documents dans lesquels ce terme apparaît *df*. Cette pondération issue du domaine de la Recherche d'Informations (RI) tire son inspiration de la loi de *Zipf* [Zipf, 1941] introduisant le fait que les termes les plus informatifs d'un corpus ne sont pas ceux apparaissant le plus dans ce corpus. Ces mots sont la plupart du temps des mots outils. Par ailleurs, les mots les moins fréquents du corpus ne sont également par les plus porteurs d'informations. Ces derniers peuvent en effet être des fautes d'orthographe ou encore des termes trop spécifiques à quelques documents du corpus étudié. Notons que certains des paramètres de notre premier modèle de sélection de descripteurs se fondent également sur cette loi empirique (cf. section 3.4.3.3). Le *tf-idf* peut se décrire formellement comme suit pour un descripteur i dans un document j parmi les N documents du corpus.

$$w_{ij} = tf_{ij} \times idf_i \tag{2.11}$$

avec

$$idf_i = log\frac{N}{n_i} \tag{2.12}$$

où n_i est le nombre de documents dans lesquels apparaît le descripteur i.
Une telle représentation avec l'exemple de nos trois phrases est donnée dans le tableau 2.2.

	aboyer	*chat*	*chien*	*faim*	*fromage*	*manger*	*rejoindre*	*souris*	*temps*
document1	0	0	0	0	0,18	0,35	0	0,18	0,48
document2	0	0	0	0,48	0	0	0,48	0	0
document3	0,48	0	0,48	0	0,18	0,18	0	0,18	0

Tab. 2.3 – Représentation vectorielle fréquentielle pondérée par tf-idf.

Nous constatons dans cet exemple son pouvoir discriminant. Notons qu'en effet, avec la simple représentation fréquentielle, le descripteur *chat* dans le second document avait le même poids que le descripteur *manger* dans le premier document. Avec le *tf-idf*, le descripteur *chat* se voit attribuer un poids nul, du fait qu'il apparaisse dans tous les documents là où le descripteur *manger* est plus discriminant à l'inverse, car apparaissant dans les documents 1 et 3.

2.2.3.2 L'entropie

Une dernière approche de pondération significative s'appuie sur l'utilisation de l'entropie. Cette dernière mesure la dispersion d'un descripteur dans un corpus et peut s'avérer une information importante dans le cadre de la sélection de descripteur et/ou de pondération de la représentation fréquentielle d'un corpus. Nous nous fondons ici sur la *log_entropy* telle que décrite dans [Dumais, 1991] qui révèle par ailleurs que cette approche obtient les meilleurs résultats pour des tâches de RI, associée à l'approche LSA que nous décrirons dans la section suivante. L'entropie E pour un descripteur i est décrite par la formule ci-dessous.

$$E(i) = \sum_j \frac{p_{ij}\ log_2\ (p_{ij})}{log_2\ N} \tag{2.13}$$

avec

$$p_{ij} = \frac{tf_{ij}}{gf_i} \tag{2.14}$$

où gf_i représente le nombre total de fois où le descripteur i apparaît dans le corpus de N documents.
Une représentation avec l'approche fréquentielle (tf) peut alors être la suivante avec pour un terme i et un document j :

$$w_{ij} = (1 + E(i))log(tf_{ij} + 1) \tag{2.15}$$

Les résultats obtenus avec notre exemple sont donnés dans le tableau 2.4.

	aboyer	*chat*	*chien*	*faim*	*fromage*	*manger*	*rejoindre*	*souris*	*temps*
document1	0	0,20	0	0	0,20	0,35	0	0,20	0,30
document2	0	0,31	0	0,30	0	0	0,30	0	0
document3	0,30	0,20	0,30	0	0,20	0,30	0	0,20	0

TAB. 2.4 – Représentation vectorielle fréquentielle pondérée par *log_entropy*.

Avec cette dernière approche, les descripteurs *chat* et *manger*, précédemment discutés ont une meilleure répartition avec un écart mieux mesuré.
Après avoir présenté quelques approches de pondération d'approches vectorielles, nous présentons ci-dessous une méthode dite de réduction.

2.2.4 La réduction / projection

Le principe de l'approche présentée dans cette section est non plus de pondérer les termes d'une matrice *Saltoniène*, mais d'en réduire la dimension en projetant les composantes de la matrice originale dans un espace vectoriel réduit.

Latent Semantic Analysis (LSA)

L'approche LSA est issue des laboratoires BellCore en 1989. Originalement, cette analyse représentait une aide à la recherche documentaire [Deerwester *et al.*, 1990]. Au fil du temps, son utilisation s'est étendue à des domaines plus variés comme le filtrage d'information [Foltz & Dumais, 1992], l'évaluation automatique de copies [Foltz, 1996], [Schreiner *et al.*, 1998], [Wiemer-Hastings *et al.*, 1999] ainsi que dans le domaine psycholinguistique par le biais de modélisation de l'acquisition [Landauer & Dumais, 1997], l'apprentissage des connaissances de l'apprenant [Zampa & Lemaire, 2002].

La méthode LSA qui s'appuie sur l'hypothèse distributionnelle émise par Harris [Harris, 1951], se fonde sur le fait que des mots qui apparaissent dans un même contexte sont sémantiquement proches. Le corpus est représenté sous forme matricielle. Les lignes sont relatives aux mots et les colonnes représentent les différents contextes choisis (un document, un paragraphe, une phrase, etc.). Chaque cellule de la matrice représente le nombre d'occurrences des mots dans chacun des contextes du corpus. Deux mots proches au niveau sémantique sont représentés par des vecteurs proches. La mesure de proximité est généralement définie par le *cosinus* de l'angle entre les deux vecteurs.
La théorie sur laquelle s'appuie LSA est la décomposition en valeurs singulières (SVD). Une matrice $A = [a_{ij}]$ où a_{ij} est la fréquence d'apparition du mot i dans le contexte j, se décompose en un produit de trois matrices TSD^t. T et D sont des matrices orthogonales et S une matrice diagonale. La figure 2.4 représente le schéma d'une telle décomposition où r représente le rang de la matrice A.

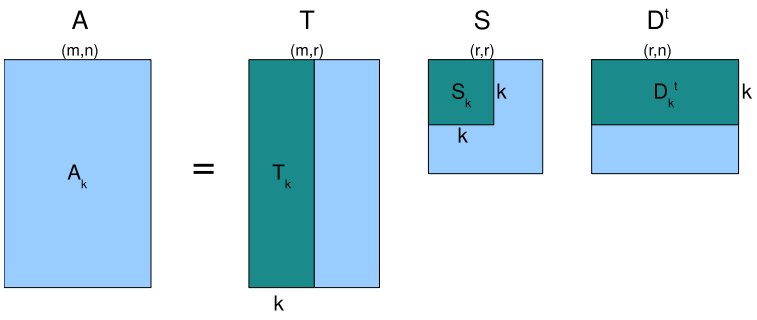

Fig. 2.4 – Décomposition en valeurs singulières.

Soit S_k où $k < r$ la matrice produite en enlevant de S les $r - k$ colonnes qui ont les plus

petites valeurs singulières. Soit U_k et V_k les matrices obtenues en enlevant les colonnes correspondantes des matrices T et D. La matrice $T_k S_k D_k^t$ peut alors être considérée comme une version compressée de la matrice originale A.

2.2.5 La similarité

Outre une représentation vectorielle de qualité, il est également nécessaire de définir une mesure permettant de comparer la proximité des différents vecteurs (issus des matrices de Salton, LSA) afin par exemple de regrouper des termes ou des documents. Ainsi, une telle proximité correspondra à des vecteurs proches. En d'autres termes, ces vecteurs auront des directions semblables ou bien encore leurs extrémités proches. Nous présentons dans cette section deux mesures classiques de proximité.

2.2.5.1 Une mesure binaire : le coefficient de Jaccard

Le coefficient de Jaccard introduit par Paul Jaccard fut à l'origine destiné à l'étude de la diversité d'espèces. Dans le cadre de la similarité vectorielle, cette mesure traduit le nombre de descripteurs (respectivement documents) communs à deux documents (respectivement descripteurs) par le nombre total de descripteurs non communs à ces deux documents (respectivement descripteurs). De manière plus formelle, ce coefficient est défini comme suit.

$$Sim_{Jaccard} = \frac{D_{commun}}{D_{total} - D_{commun}} \tag{2.16}$$

avec D_{commun} le nombre de descripteurs communs aux deux documents dont la similarité est mesurée, et D_{total} le nombre total de descripteurs des deux documents (et respectivement dans le cadre de la mesure de deux descripteurs). Nous remarquons ainsi qu'une telle approche ne se fonde que sur la *présence/absence* de descripteurs (ou documents). Le calcul du coefficient de Jaccard avec les documents 1 et 3 présentés en section 2.2.2.1 est donné ci-dessous.

$$Sim_{Jaccard}(doc1, doc3) = \frac{4}{9-4} = 0,8 \tag{2.17}$$

Ainsi, elle ne traite que des cas booléens explicités ci-dessus. Nous présentons dans la section suivante une approche plus adaptée à la représentation fréquentielle.

2.2.5.2 Produit scalaire, angle et cosinus

Le *cosinus* est l'une des premières mesures à avoir été utilisée dans le domaine de la recherche d'information et est régulièrement employée afin de déterminer le degré de

similarité de vecteurs. Cette mesure fut notamment la première à être utilisée dans le système de recherche d'information SMART [Salton, 1971].
Les résultats obtenus avec cette mesure varient entre *0* et *1*. Un score de 1 signifie que l'*angle* formé entre les deux vecteurs est très faible, indiquant une forte proximité des vecteurs. Le cosinus entre deux vecteurs est obtenu en calculant le *produit scalaire* entre ces deux vecteurs, que nous divisons par le produit de la norme des deux vecteurs. Le cosinus entre deux vecteurs $\overrightarrow{u}$ et $\overrightarrow{v}$ de telle sorte que θ soit l'angle formé par ces deux vecteurs est défini par l'équation suivante :

$$\theta = arccos \frac{\overrightarrow{u} . \overrightarrow{v}}{\| \overrightarrow{u} \| . \| \overrightarrow{v} \|} \tag{2.18}$$

En considérant un espace sémantique à deux dimensions représentant la fréquence d'apparition de descripteurs dans des documents, nous avons la représentation graphique fournie par la figure 2.5.

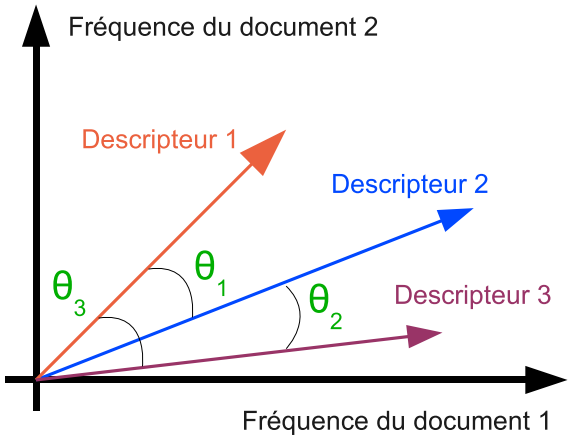

FIG. 2.5 – Proximité de descripteurs obtenue via l'angle résultant de leurs représentations vectorielles, en fonction des documents dans lesquels ils apparaissent.

Cet exemple montre qu'avec la mesure cosinus (et *a fortiori* l'angle entre les deux vecteurs), les descripteurs 2 et 3 sont les plus proches sémantiquement.
Notons que la mesure cosinus, qui est couramment employée avec des représentations vectorielles fréquentielles peut être utilisée avec d'autres représentations, dont notamment LSA. Nous montrons ci-dessous le calcul du cosinus entre les documents 1 et 3 tels que décrit en section 2.2.2.1.

$$cos(\theta) = \frac{5}{\sqrt{6} \times \sqrt{6}} \approx 0,83 \tag{2.19}$$

L'angle entre les deux vecteurs vaut alors :

$$\theta \approx arccos(0,83) \approx 0,58 \tag{2.20}$$

2.2.5.3 D'autres mesures de similarité

Bien d'autres mesures de similarité entre termes et/ou documents de corpus sont proposées dans la littérature. Citons l'information mutuelle au cube [Daille, 1994] et le coefficient de Dice décrit dans [Smadja *et al.*, 1996] dont les objectifs sont de renforcer les termes/documents fréquents et rares. Ces deux mesures seront abordées dans la section 3.3.4.1 et comparées à d'autres dans le cadre de la sélection de descripteurs pertinents de notre modèle de sélection de descripteurs SELDE. Nous utiliserons également la mesure de *Minkowski* [Sokal, 1977] et la mesure *Okabis* [Bellot & El-Bèze, 2001] qui seront explicitées en section 5.5.2. Notons que ces mesures ne peuvent être appliquées directement aux vecteurs définis dans cette section tel que nous le montrerons en section 3.3.4.1. D'autres mesures n'ont pas été abordées dans ce mémoire comme le *rapport de vraisemblance* [Dunning, 1993] utilisé dans le domaine de l'extraction terminologique, des mesures utilisées principalement avec des règles d'associations : la mesure de *Sebag-Schœnauer* [Sebag & Schoenauer, 1988] et la *J-mesure* [Goodman & Smyth, 1988] ou encore la *conviction* [Silverstein *et al.*, 1998], etc. Notons finalement qu'il peut être également employé la mesure du χ^2 détaillée en tant que mesure de sélection de descripteurs dans la section 2.1.2.1. Les mesures de similarité peuvent en effet être utilisées dans plusieurs tâches comme la sélection de descripteurs pertinents, l'étude de la proximité de termes ou documents, etc. Le lecteur intéressé par une liste plus conséquente de mesures de similarité de vecteurs pourra se référer à [Besançon, 2001].

2.2.6 Les autres modèles de représentation

La représentation vectorielle de documents textuels est assez répandue dans la littérature mais n'est cependant pas exclusive. Il existe en effet un certain nombre d'autres approches comme des modèles dits probabilistes ou encore séquentiels. Ce dernier est assez complexe et nécessite la mise en place de modèles évolués comme des modèles de Markov cachés. Un document textuel est alors représenté de manière générique par une séquence telle que le montre la formule suivante.

$$d = (w_1^d, ..., w_{|d|}^d) \tag{2.21}$$

avec $|d|$ le nombre de mots du document d et w_i^d représentant le $i^{\text{ème}}$ mot du document. Une telle représentation a l'avantage de conserver l'ordre des mots mais ne donne pas

toujours des résultats pertinents [Denoyer, 2004].
Le modèle probabiliste est quant à lui principalement utilisé pour des tâches de classifications qui seront évoquées ci-dessous, comme par exemple dans [Robertson & Jones, 1976]. La représentation probabiliste, détaillée dans [Spärck-Jones, 1999], est à l'origine d'un certain nombre de méthodes comme le système Okapi [Robertson *et al.*, 1996]. Cette approche fut l'un des systèmes les plus performants lors d'évaluations de *TREC*[8] (Text REtrieval Conference) avec le modèle vectoriel [Robertson *et al.*, 1996].

Rappelons que nous utilisons dans ce mémoire l'approche vectorielle principalement. Nous présentons dans la section suivante diverses tâches utilisant les descripteurs précédemment décrits et montrons dans quelles mesures ils se révèlent performants.

2.3 Comment sont utilisés ces descripteurs

2.3.1 Utilisation des descripteurs pour des tâches de classification

La sélection de descripteurs pertinents est une étape indispensable pour une tâche de classification de données textuelles. Nous présentons dans cette section le principe ainsi que les différentes approches décrivant cette tâche.

2.3.1.1 Principe

La classification de données textuelles consiste à attribuer à un document une ou plusieurs catégories (ou classes). Plusieurs modes de représentations de données peuvent être employés afin de réaliser une classification. Nous nous intéressons dans ce mémoire aux représentations vectorielles de type Salton et LSA. Ainsi, en nous appuyant sur une représentation classique (cf. figure 2.6), nous distinguons deux types de classification de données textuelles : la **classification de textes** et la **classification conceptuelle**.
Rappelons qu'une représentation matricielle classique d'un corpus (de type Salton) représente en lignes les mots du corpus et en colonnes les documents. Ainsi, nous associons l'étude de la proximité des mots d'un corpus à une classification conceptuelle et l'étude de la proximité des documents à la classification de textes.

De nombreuses applications en découlent comme la veille technologique ou bien encore la recherche documentaire. La classification de document peut faire intervenir la notion d'apprentissage ou non. Nous présentons ci-dessous une telle notion.

[8] http ://trec.nist.gov/

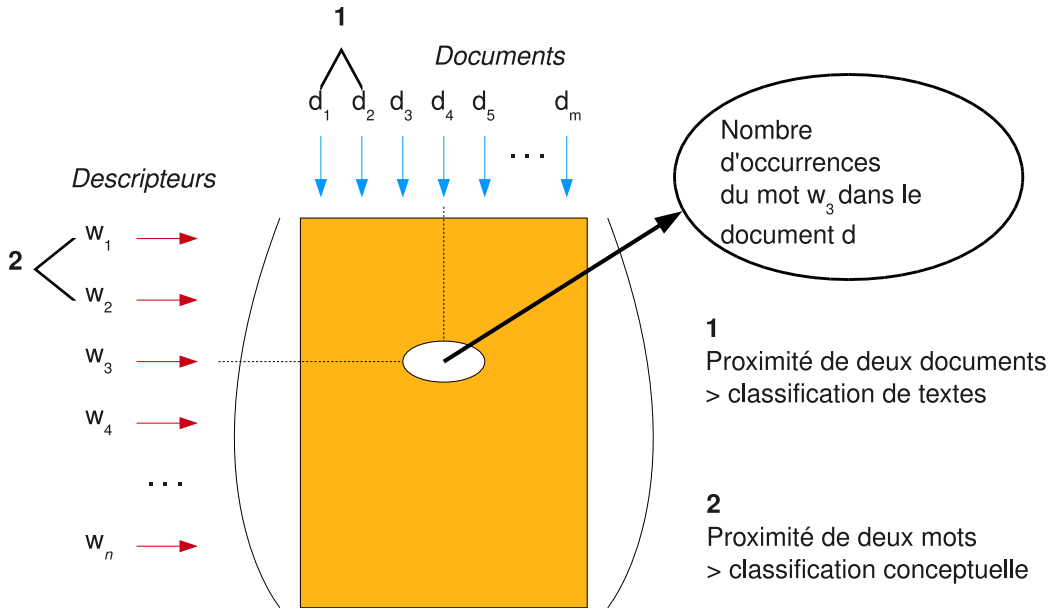

FIG. 2.6 – Classification de textes et classification conceptuelle.

2.3.1.2 La notion d'apprentissage

L'apprentissage "automatique" consiste en l'acquisition de connaissances à partir d'observation de phénomènes. Le plus souvent, l'apprentissage est associé à la construction d'un **modèle** à partir duquel de nouvelles entrées pourront être identifiées. Nous distinguons deux types d'apprentissage : le **supervisé** et le **non supervisé**. Ces deux notions sont souvent utilisées pour accomplir des tâches distinctes. L'apprentissage supervisé est exploité afin de réaliser des tâches de décision ou de prévision là où les approches non supervisées sont souvent employées à des fins exploratoires. Notons que l'apprentissage est souvent associé au procédé de classification. Rappelons cependant que ces deux notions – *apprentissage* et *classification* – se distinguent. Nous pouvons tout à fait établir une classification sans apprentissage (comme par exemple [Vernier *et al.*, 2009]) ou bien effectuer une tâche nécessitant un apprentissage sans pour autant faire de la classification. Citons par exemple l'étiqueteur grammatical TreeTagger [Schmid, 1995] qui utilise un apprentissage sous forme d'arbre de décision afin de "prédire" la catégorie grammaticale de termes.
Nous nous intéressons dans les sections suivantes aux algorithmes d'apprentissage employés dans le cadre de classification de données textuelles.

2.3.1.3 Les approches avec apprentissage supervisé

Nous présentons ci-dessous un nombre non exhaustif d'algorithmes d'apprentissage supervisé couramment utilisés dans le cadre de classification de données textuelles. Rap-

pelons par ailleurs qu'il n'est pas dans les objectifs de cette thèse de spécifiquement traiter les approches d'apprentissage. Nous nous sommes en effet focalisés sur l'étude de **la représentation des données textuelles afin d'appliquer des algorithmes "classiques" de classification**. Le lecteur intéressé par cette notion pourra par exemple se référer à [Cornuéjols & Miclet, 2002] ou encore [Mitchell, 1997].
Nous distinguons ici plusieurs types de classificateurs :

- Les "plus proches voisins"
- Les probabilistes
- Les modèles minimisant l'erreur
- Les classificateurs fondés sur des réseaux de neurones artificiels
- Les décisionnels

Nous présentons ci-dessous un exemple de chaque type de classificateurs listés ci-dessus.

Les plus proches voisins.

Les approches de type "plus proches voisins" cherchent à définir des zones propres à un concept. Ainsi, un nouvel élément à intégrer au modèle se voit attribuer la classe de ses plus proches voisins. Une approche classique de ce type est l'algorithme des k-plus proches voisins (k-ppv).
Le principe de l'algorithme des k-ppv [Cover & Hart, 1967] est de mesurer la similarité entre un nouveau document et l'ensemble des documents ayant été préalablement classés. Ces documents peuvent être considérés comme un modèle d'apprentissage, bien qu'il n'y ait pas de réelle phase d'apprentissage avec les k-ppv, point sur lequel nous reviendrons en section 2.3.1.5.

Cet algorithme revient à constituer un espace vectoriel dans lequel chaque document est modélisé par un vecteur de mots. Un tel vecteur a pour dimension le nombre de mots de la base d'apprentissage. Chaque élément de ce vecteur est en effet constitué du nombre d'occurrences d'un mot issu de la base d'apprentissage. Les documents classés sont ordonnés de manière décroissante afin que le premier document soit celui ayant obtenu le meilleur score de similarité avec le document devant être classé. Suivant la valeur de k, il est ainsi effectué un classement des k documents les plus proches. Notons que la mesure de similarité la plus couramment utilisée est le calcul du *cosinus* de l'angle formé par les deux vecteurs de documents.

Après avoir déterminé quels étaient les k plus proches voisins, il faut définir une méthodologie afin d'attribuer une classe au nouveau document. Il existe dans la littérature deux approches classiques décrites spécifiquement dans [Bergo, 2001] afin de

répondre à cette problématique :
- soit proposer de classer le document dans la même catégorie que celui ayant obtenu le meilleur score de similarité parmi le jeu d'apprentissage,
- soit, si $k > 1$ de considérer les k documents les mieux classés. Alors nous pouvons attribuer la classe suivant plusieurs options. Une première méthode peut être de calculer parmi les k documents les plus proches, pour chaque catégorie, le nombre de documents appartenant à cette catégorie (1). Une seconde propose de prendre en compte le rang des k documents (2). Il s'agit pour toutes les catégories, d'effectuer la somme des occurrences d'une catégorie multipliée par l'inverse de son rang.

Prenons par exemple un document d à classer parmi quatre classes, C1, C2, C3 et C4. Définissons $k = 6$. Considérons le classement suivant de d_{new} avec le jeu d'apprentissage D contenant les documents d_i :

documents	classe des documents	rang
d1	C2	1
d2	C2	2
d3	C4	3
d4	C4	4
d5	C1	5
d6	C4	6

En utilisant la première approche (1), nous aurions attribué la classe C4 à d_{new}. En effet la classe C4 est celle qui possède le plus de documents parmi les k plus proches voisins (trois documents). La seconde approche (2) aurait quant à elle classé d_{new} dans C2. Nous obtenons en effet avec cette mesure par exemple pour la classe C1 : un seul document dans la classe au cinquième rang soit C1 = 1/5 = 0,2. Nous obtenons pour les autres classes C2 = 1,5, C3 = 0 et C4 = 0,75.

Nous utiliserons dans les chapitres 4, 5 et 7 la première approche (1), celle-ci étant la forme la plus répandue comme décrite dans ([Yang & Liu, 1999b]) en utilisant deux paramètres :
- le seuil de classe, fixant un nombre minimal de descripteurs devant appartenir à une classe pour qu'un nouveau document soit attribué à cette classe,
- le seuil de similarité en dessous duquel, les candidats ne seront plus admis parmi les k plus proches voisins car étant jugés d'une similarité trop éloignée.

Les approches probabilistes.

Une approche probabiliste consiste à classer un nouvel élément dans une classe en fonction de sa probabilité d'appartenance. Un algorithme classique est l'approche Bayésienne naïve (NaiveBayes).
Le classificateur de type Naïve Bayes est fondé sur le théorème de [Bayes, 1763]. Considérons $v_j = (v_{j1}, ..., v_{jk}, ..., v_{jd})$ un vecteur de variables aléatoires représentant un document d_j et C un ensemble de classes.
En s'appuyant sur le théorème de Bayes, la probabilité que ce dernier appartienne à la classe $c_i \in C$ est définie par :

$$P(c_i|v_j) = \frac{P(c_i)P(v_j|c_j)}{P(v_j)}$$

La variable aléatoire v_{jk} du vecteur v_j représente l'occurrence de l'unité linguistique k retenue pour la classification dans le document d_j.
La classe c_k d'appartenance de la représentation vectorielle v_j d'un document d_j est définie comme suit :

$$c_k = arg\ max\ P(c_i \in C) \prod_k P(v_{jk}|c_j)$$

En d'autres termes, le classificateur Naïve Bayes affecte au document d_j la classe ayant obtenu la probabilité d'appartenance la plus élevée.
Alors, $p(c_i)$ est définie de la façon suivante :

$$P(c_i) = \frac{nombre\ de\ documents \in c_i}{nombre\ total\ de\ documents}$$

En faisant l'hypothèse que les v_j soient indépendantes, la probabilité conditionnelle $P(v_j|c_i)$ est définie ainsi :

$$P(v_j|c_i) = P(v_{jk}|c_i)$$

Une telle hypothèse d'indépendance des v_j peut néanmoins dégrader qualitativement les résultats obtenus avec une telle approche [Lewis, 1998].

Les approches minimisant l'erreur de classification.

De telles approches vont chercher à minimiser l'erreur de classification. L'algorithme le plus répandu est celui des machines à support vectoriel (SVM).
L'algorithme des SVM est un classificateur binaire. Il consiste dans un premier temps à projeter les données utilisées dans un espace vectoriel. Ce classificateur part du principe que les données à classifier sont linéairement séparables. Ainsi, lors de cette phase de vectorialisation des données, il est employé un *noyau*, qui n'est autre qu'une

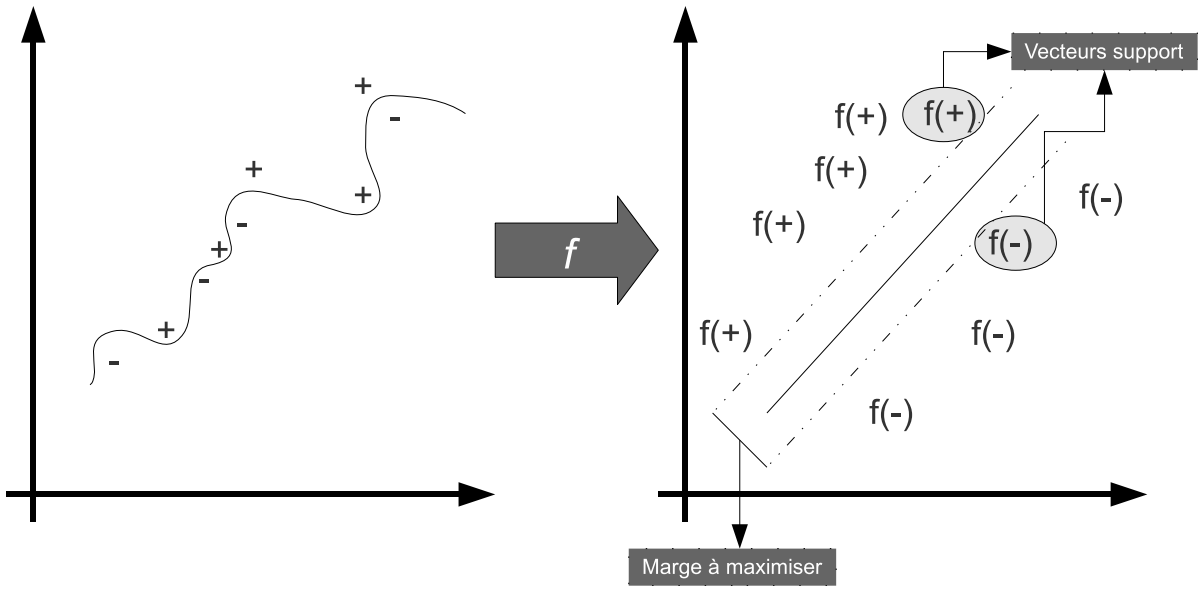

FIG. 2.7 – Transformation d'un problème non linéairement séparable en un problème linéaire via le noyau *f*.

fonction de transformation, permettant de rendre le problème linéaire. Les noyaux les plus fréquemment utilisés sont linéaires, polynomiaux ou gaussiens. Un exemple de linéarisation est présenté sur la figure 2.7.

Une fois le problème linéarisé, le classificateur va chercher un hyperplan afin de séparer au mieux les exemples positifs des exemples négatifs. L'idée est de garantir que la marge entre le plus proche des positifs et des négatifs soit maximale. Les points se situant précisément sur l'hyperplan sont appelés les vecteurs supports.
Cette approche peut facilement s'appliquer à des problèmes multi-classes (c.-à-d. supérieurs à deux) en utilisant le principe *One-against-the-Rest*. Ce dernier propose de comparer chaque classe à l'ensemble des autres afin de trouver un hyperplan séparateur. Cet algorithme est assez fréquemment utilisé dans la littérature afin de résoudre des problèmes de classification. Notons que nous utiliserons dans le chapitre 4 l'algorithme multi-classes SMO [Platt, 1999].

Les neuronales

Les approches neuronales furent dans les premières à être utilisées afin de réaliser un apprentissage. Ces approches s'inspirent du fonctionnement du système nerveux humain. Ainsi, elles se fondent sur l'utilisation de "neurones" artificiels qui vont effectuer la tâche d'apprentissage. Les automates à seuil visant à modéliser l'activité neuronale

[Mcculloch & Pitts, 1943] ainsi que les règles d'apprentissage locales introduites par [Hebb, 1961] et [Widrow, 1985] ont fortement inspiré les méthodes d'apprentissage neuronal. Les premières approches de ce type furent les "réseaux de neurones monocouche tels que le *Perceptron* [Rosenblatt, 1958] (ré-édité dans [Rosenblatt, 1988]) ou bien encore *Adaline* [Widrow & Hoff, 1960] (ré-édité dans [Widrow & Hoff, 1988]). Plus tardivement furent produits des algorithmes multicouches dont notamment le *Perceptron multicouche* [Cybenko, 1989]. Le principe général d'une approche neuronale est présenté ci-dessous.

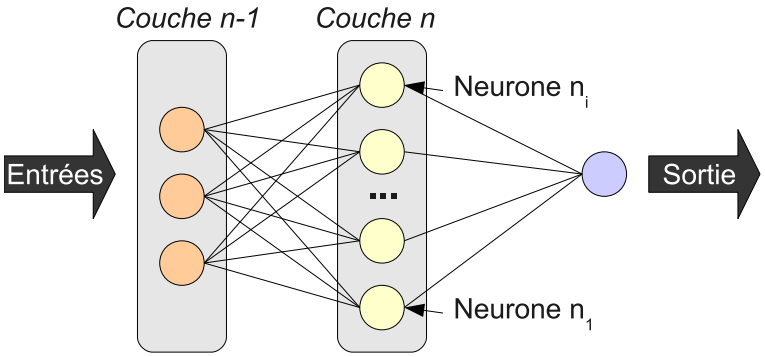

FIG. 2.8 – Architecture générale d'un réseau de neurones artificiels.

Un réseau de neurones artificiels est composé d'une ou plusieurs couches se succédant dont chaque entrée est la sortie de la couche qui la précède comme illustré sur la figure 2.8. Ainsi, une couche de rang n est composée d'un certain nombre de *neurones artificiels* dont leurs entrées sont constitués des sorties des neurones de couche *n-1*. Les neurones humains reçoivent en entrée les signaux provenant des autres neurones par des *synapses*. Appliquée aux réseaux de neurones artificiels, la connexion neuronale s'effectue par le biais de liaisons pondérées (les synapses) unidirectionnelles. Nous pouvons ainsi voir un réseau de neurones artificiels comme un réseau ou graphe orienté dont les nœuds sont les neurones artificiels. Finalement, le but va être d'attribuer des poids synaptiques à chaque neurone afin d'obtenir le résultat voulu en sortie. C'est ainsi que l'on construit une base d'apprentissage en pondérant les différents poids synaptiques d'un réseau afin de se rapprocher au mieux du résultat souhaité.

Les décisionnels

Les algorithmes décisionnels sont des systèmes à base de règles du type si A, alors B schématisé par $A \rightarrow B$, A et B étant des descripteurs ou descripteurs et classes. Dans le cas où A et B sont des jeux de données, ces règles sont nommées règles d'association. Notons que ces règles sont produites en effectuant un apprentissage non supervisée dans

le cas où A et B sont des descripteurs et supervisé si A est un descripteur et B une classe. Un exemple de construction de telles règles est donné dans [Aggarwal *et al.*, 1998] où les auteurs proposent d'associer des traits caractérisant une personne (âge, salaire, éducation, statut marital, nombre d'enfants) à leurs habitudes de consommation. Il est ainsi produit un modèle d'apprentissage pouvant prédire la consommation de nouvelles personnes.
Une extension naturelle des algorithmes décisionnels sont les systèmes à base d'arbres de décision [Paliouras *et al.*, 1999]. Un arbre de décision est un arbre tel qu'un nœud correspond à un attribut, les feuilles issues de ce nœud ont des valeurs possibles pour cet attribut. Les feuilles correspondent à une classe. Le chemin parcouru par un nouvel exemple de la racine de l'arbre jusqu'à la feuille détermine sa classe. La principale difficulté est liée à l'élaboration d'un tel arbre afin de construire une base d'apprentissage à l'instar de la définition des règles d'association dans un système éponyme. CART [Breiman *et al.*, 1984], ID3 [Quinlan, 1986] et C4.5 [Quinlan, 1993] sont des exemples d'algorithmes permettant la construction d'arbre de décisions. C4.5 vise par exemple à séparer en un ensemble le plus homogène possible des cas exemples. Cette séparation s'appuie sur l'entropie [Cornuéjols & Miclet, 2002] qui mesure la quantité d'information.

Un certain nombre d'algorithmes et autres méthodes permettent également d'effectuer une classification avec un apprentissage supervisé. Citons par exemple la régression logistique, les modèles de Makov cachés [Baum & Petrie, 1966] ou encore des techniques hybrides comme le boosting [Kearns, 1988]. Rappelons cependant que notre but n'est pas d'être exhaustif mais de présenter une vue d'ensemble de ces algorithmes avec apprentissage supervisé.

2.3.1.4 Les approches avec apprentissage non supervisé

Cette section présente quelques algorithmes de classification non supervisé. Notons que nous n'avons pas utilisé ce type d'algorithme au cours de cette thèse. Ainsi, ils seront sommairement abordés.
Une classification avec apprentissage non supervisé se différencie de l'approche supervisé avec la connaissance ou non des classes à "prédire". De plus, aucune donnée étiquetée n'est disponible. Il n'est alors pas possible de générer un modèle d'apprentissage. Ainsi, la tâche ne consiste plus seulement à attribuer une classe à un nouvel élément (approche supervisée) mais également à définir les classes et leurs nombres. Notons que ces classes sont souvent appelées des partitions (ou *clusters*). Déjà abordée dans le cadre d'apprentissage supervisé, la notion de plus proche voisin est très souvent utilisée en apprentissage non supervisé, notamment avec l'approche des *k-moyennes* [Macqueen, 1967] décrite suc-

cinctement ci-dessous.

Les k-moyennes

La première étape de cette approche est la sélection arbitraire de k centres autour desquels sont regroupés les éléments les plus proches de ces centres. Il est alors calculé le centre de gravité de chaque classe ainsi définie au fur et à mesure du regroupement des éléments. Ces derniers vont définir les nouveaux centres des classes. Cette opération est répétée jusqu'à ce que la dispersion des membres de chaque classe soit minimale. Les classes deviennent en effet à chaque itération plus compactes, permettant une convergence de l'algorithme. Notons qu'un certain nombre d'évolutions de cet algorithme ont vu le jour dont les x-moyennes [Pelleg & Moore, 2000] ou c-moyennes [Bezdek, 1981]. Cette dernière approche permet une classification floue. Ainsi, chaque document n'est plus associé à une seule classe mais à plusieurs, en fournissant différents degrés d'appartenances. En effet, l'approche originale des k-moyennes conduit souvent à un manque de précision dans le cas où des classes se chevauchent. Un des algorithmes d'apprentissage non supervisé flou les plus répandu est l'Espérance-Maximisation (EM) dont le principe est résumé ci-dessous.

L'Espérance-Maximisation

Le principe de cet algorithme proposé par [Dempster *et al.*, 1977] est d'alterner itérativement les phases dites d'espérance et de maximisation. L'espérance consiste à calculer l'espérance de vraisemblance en tenant compte des variables observées. La maximisation estime le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape d'espérance. Les paramètres résultant de la maximisation sont alors utilisés afin de réitérer l'opération. Nous pouvons trouver un certain nombre d'extensions proposées pour cet algorithme dans [Mclachlan & Krishnan, 2007].

Apprentissage par renforcement

Le principe de l'apprentissage par renforcement est de trouver le meilleur choix possible pour une action avec un processus d'*essais* et d'*erreurs*. Ainsi, à chaque action le système d'apprentissage par renforcement effectue un certain nombre d'essais dont il évalue la pertinence par le biais d'une fonction de récompense. Ce type d'apprentissage est non supervisé car la récompense donne juste un indice de qualité et non pas le résultat optimal. Parmi les premiers algorithmes, nous pouvons citer le *Q-learning* [Watkins, 1989] et le TD-learning [Sutton, 1988].

Notons que nous n'avons pas abordé les approches dites d'apprentissage **semi-supervisé**. Ces approches proposent d'apprendre à effectuer des tâches d'apprentissage supervisé en utilisant une faible quantité de données étiquetées (dont les classes sont connues et définies) et une quantité importante de données brutes. Le lecteur intéressé par ces approches pourra consulter [Zhu, 2007] qui propose un état de l'art assez complet.

2.3.1.5 Les approches sans apprentissage

Notons que la *classification sans apprentissage* n'est pas de la *classification non supervisée*. En effet cette dernière est souvent employée à tort afin de désigner une tâche de classification avec apprentissage non supervisé. La distinction entre les approches avec et sans apprentissage est assez triviale. La première utilise les données qu'elle manipule afin d'apprendre leur contenu (constituant une base d'apprentissage en supervisé et permettant de converger vers un résultat en non supervisé). Rappelons que le supervisé ou non supervisé se distinguent par le fait d'apprendre sur des données étiquetées ou non, soit en d'autres termes sur la connaissance ou non des classes. Ainsi, le simple fait de relancer un algorithme en ayant reconstitué des classes via une précédente itération est une forme d'apprentissage. A l'inverse, le fait de ne pas utiliser d'apprentissage suppose que les données à classifier ne sont aucunement utilisées afin de prédire la classe (dans le cas d'une tâche de classification mais cette notion se généralise). Parmi les approches non supervisées, nous pouvons par exemple utiliser des ressources extérieures comme des thésaurus ou ontologies afin de déterminer l'appartenance d'un terme à une classe (dans le cadre d'une classification conceptuelle). Nous pouvons citer l'approche de [Hignette *et al.*, 2007] qui utilise une ontologie afin de classer automatiquement des tableaux par types sémantiques. Notons que les auteurs se comparent à une approche de classification par apprentissage supervisé (SVM multi-classe SMO) et qu'ils obtiennent des résultats pertinents. Notons que les approches sans apprentissage sont plus fréquemment employées dans le cadre de construction de classes conceptuelles que dans le cadre de classification de documents.
Revenons par ailleurs sur l'approche des k-ppv. Pouvons-nous en effet la considérer comme une approche d'apprentissage ? Cette dernière effectue plutôt un calcul de similarité sur des exemples étiquetés. Ainsi, nous pourrions la qualifier d'approche sans apprentissage mais supervisée. Nous employons en effet des exemples étiquetés ce qui est caractéristique des approches dites supervisées mais ne construisons pas de modèle d'apprentissage avec les *k-ppv*. En effet, à chaque nouvel élément, il est nécessaire de recalculer la proximité avec l'ensemble des éléments étiquetés. Nous avons néanmoins présenté les *k-ppv* dans les approches de classification avec apprentissage supervisé, le non apprentissage supervisé n'étant pas clairement défini dans la littérature.

2.3.1.6 Type de descripteurs utilisés en classification

Les applications de classification de données textuelles sont abondantes dans la littérature. Pratiquement tous les types de descripteurs présentés dans cet état de l'art sont utilisés pour cette tâche. Les flexions et les lemmes sont les plus utilisés. Le choix du lemme est assez discuté dans la littérature (par exemple dans [Brunet, 2002] ou [Mayaffre, 2005]). Les résultats obtenus lors de tâches de classification restent cependant influencés par la lemmatisation ou non d'un corpus [Riloff, 1995]. [Gonçalves & Quaresma, 2005] évaluent l'impact de la lemmatisation en appliquant divers pré-traitements comme la suppression de mots outils. Les travaux de [Sjöblom, 2002] appliquent différentes méthodes fondées sur les formes graphiques ou sur les lemmes d'un même corpus pour pouvoir comparer leurs apports dans l'analyse des données textuelles. Par ailleurs, [Liao *et al.*, 2007] montrent que l'utilisation de radicaux améliore également les résultats de classifications, notamment avec une sélection de type *tf-idf*.
Les n-grammes de caractères semblent être les plus adaptés aux tâches de classification comme le montrent [Jalam & Chauchat, 2002]. Ce descripteur, outre les résultats pertinents qu'il obtient avec des données "classiques" est également robuste au manque d'informations en conservant des résultats de bonne qualité de classification. Ces descripteurs sont particulièrement adaptés aux langues complexes. Citons [Mansur *et al.*, 2006] qui réalise une tâche de classification d'un corpus d'actualités du langage *Bangali*[9] ou encore [Vardhan *et al.*, 2007] qui montrent également de bonnes performances pour la classification de textes du langage *Telugu*[10] en se fondant sur les 3-grammes de caractères. Les n-grammes de mots sont également performants appliqués à la classification comme [Tan *et al.*, 2002] qui améliorent les résultats des descripteurs "flexions" avec des bigrammes de mots. [Paradis & Nie, 2005] proposent également d'utiliser des n-grammes de mots pour une tâche de classification de textes. La méthode consiste à classer les documents bruités, que nous définirons dans le chapitre 5, en se fondant sur le filtrage du contenu avec les n-grammes de mots et les entités nommées sur des documents de type "appels d'offres" Pour finir, [Peng *et al.*, 2003] montrent comment le descripteur "n-gramme de mots" peut améliorer le classificateur NaiveBayes.

La sélection de descripteurs est quant à elle assez dépendante de l'algorithme de classification utilisé. Les SVM sont par exemple assez robustes et la sélection de descripteurs se révèle être moins discriminante qu'avec les k-ppv [Lewis *et al.*, 2004]. La sélection de descripteurs statistiques est assez fréquente pour des tâches de classification avec apprentissage supervisé comme dans les travaux de [Rogati & Yang, 2002] qui

[9] Langue parlée au Bengale et au Bangladesh
[10] Langue parlée en Inde

proposent une sélection via la mesure du χ^2. Les auteurs préconisent par ailleurs l'emploi du χ^2 en le combinant à d'autres approches comme le gain d'information. L'hybridation d'approches de sélection statistique est fréquemment rencontrée dans la littérature comme dans [Guo & Murphey, 2000]. Ce dernier se fonde notamment sur l'algorithme EM de classification avec apprentissage non supervisé. [Basili *et al.*, 2001] proposent également une hybridation de descripteurs statistiques et linguistiques utilisant notamment une sélection par catégorie grammaticale, ainsi que les étiquettes grammaticales associées. Leurs travaux sont une extension de [Basili *et al.*, 2000] dans lequel ils montraient l'intérêt d'une sélection de descripteurs fondée sur les catégories grammaticales et l'utilisation des étiquettes. Les descripteurs morphosyntaxiques sont particulièrement employés dans le cadre de classification de données d'opinions comme le montrent les travaux de [Genereux & Santini, 2007]. De plus, [Benamara *et al.*, 2007] montrent que les adjectifs sont particulièrement adaptés aux données d'opinions. Par ailleurs, nous trouvons dans les approches de [Greevy & Smeaton, 2004] une classification de documents racistes issus du web. Cette étude montre que les documents racistes contiennent beaucoup plus d'adjectifs que les documents non racistes. Citons pour finir [Pisetta *et al.*, 2006] qui présentent une approche de réduction par identification de contexte. Les auteurs effectuent en premier lieu une extraction terminologique. Les documents sont alors représentés avec les descripteurs préalablement extraits. La tâche de classification est alors réalisée avec l'aide de ressources externes de type WordNet. Cette approche s'inspire de [Kumps *et al.*, 2004] qui ont montré l'intérêt d'une telle réduction dans le cadre de mesures de similarité entre descripteurs.

Les sélections de descripteurs avec des méthodes statistiques et linguistiques ont été de nombreuses fois abordées dans la littérature. Ainsi, un type de sélection qui se popularise se fonde sur l'utilisation de connaissances extérieures afin de générer de nouveaux descripteurs ("Feature Generation"). Une raison simple à cet engouement est l'essor toujours grandissant des ressources de l'internet. Citons par exemple l'encyclopédie communautaire Wikipédia, avec l'approche de [Gabrilovich & Markovitch, 2006]. Ces derniers extraient des descripteurs de Wikipédia afin d'enrichir leur base d'apprentissage, améliorant par ce biais les résultats de classification. Ce dernier article est une extension de [Gabrilovich & Markovitch, 2005] dans lequel les auteurs utilisaient notamment la ressource Open Directory Project (ODP) afin d'enrichir une base d'apprentissage. [Wang & Domeniconi, 2008] vont plus loin dans l'utilisation de Wikipédia en décrivant et expérimentant une méthode visant à construire un noyau sémantique fondé sur les ressources de Wikipédia. Ainsi, ces ressources sont utilisées comme un thésaurus afin d'améliorer la classification automatique de documents.
De nombreux travaux de recherche portent également sur l'apport de ressources sé-

mantiques comme celles fournies par *WordNet*. Citons [Paolo *et al.*, 2004] qui utilisent ces ressources afin de fournir au classificateur (les k-ppv) des informations sémantiques supplémentaires. [Zhang *et al.*, 2005] proposent également d'utiliser WordNet afin de sélectionner des descripteurs pertinents. Ils obtiennent de meilleurs résultats que le gain d'information et le *tf*.

Ainsi, un nombre conséquent d'approches visant à sélectionner des descripteurs a été expérimenté dans la littérature ainsi que de multiples approches hybrides. Un type de descripteurs reste cependant peu utilisé et constitue pour nous une information essentielle, les relations syntaxiques, sélection que nous proposerons d'expérimenter dans le chapitre 4 de ce mémoire. Citons par exemple [Wiemer-Hastings & Zipitria, 2001] qui proposent de segmenter un texte en trois sous-matrices de type Salton. Ces matrices représentent les sujets, verbes et compléments. Les auteurs appliquent alors la méthode LSA sur ces trois matrices afin de réaliser des calculs de proximité de descripteurs. Notons que les approches visant à utiliser des informations syntaxiques avec LSA sont l'objet de la section 4.1.4.

Les modèles vectoriels sont couramment utilisés pour des tâches de classification. L'approche vectorielle de Salton est la plus employée comme le montre [Dongbo, 2001] mais de nombreux travaux emploient également l'approche de réduction LSA comme ceux de [Aseervatham, 2008], [Wan & Tong, 2008] (évalué sur des données d'opinions) et [Li & Park, 2007] (modèle évalué sur des corpus de *Reuter*), ou encore à base de vecteurs d'idées comme dans [Lafourcade, 2006].

2.3.2 Extraction d'information

Nous pouvons définir l'extraction d'information comme une discipline visant à analyser de manière précise le contenu d'un document. Appliquée aux documents textuels, cette discipline consiste à interpréter le contenu d'un texte (écrit en langage naturel) afin de répondre à un besoin d'informations structurées et complexes. Un processus générique d'extraction d'information peut se décomposer comme suit [Hobbs, 1993] et [Cardie, 1997].

– *La mise en forme du texte*. Il est effectué une série d'analyses lexicales et morphologiques afin d'identifier les constituants d'un texte et leurs relations syntaxiques. Nous pouvons utiliser pour ce faire des dictionnaires, des étiqueteurs grammaticaux ou encore effectuer une analyse syntaxique.

– *L'extraction de faits*. Souvent réalisée par l'application d'expressions régulières respectant des patrons d'extraction syntaxique, elle permet d'extraire l'information contenue

dans les groupes précédemment mis en forme. Ces patrons sont la plupart du temps des connaissances extérieures.

– *La génération d'un formulaire.* La dernière étape consiste en la réalisation d'un formulaire caractérisant chaque événement du corpus étudié. Notons que ce formulaire est la plupart du temps représenté par une ontologie du domaine décrite par le corpus. Finalement, l'extraction d'information est effectuée en respectant le formulaire ainsi créé. En effet, les champs constituant le formulaire représentent l'information à extraire. Le choix du descripteur est alors crucial et doit dépendre de l'information recherchée. L'extraction de la terminologie, pouvant être vue comme l'étude des mots techniques propres à un domaine et de leurs significations, est un sous-domaine de l'extraction d'information ou le choix du descripteur est primordial. Le système LEXTER proposé par [Bourigault, 1994] utilise par exemple les syntagmes nominaux. Un certain nombre d'approches se fondent sur l'utilisation de patrons syntaxiques, pouvant être créés à partir d'outils comme INTEX [Silberztein, 1993]. Citons par exemple les travaux de [Ibekwe-sanjuan & sanjuan, 2003]. Nous trouvons également des méthodes statistiques utilisant notamment l'information mutuelle afin de sélectionner des descripteurs à extraire. Citons par exemple [Church *et al.*, 1991]. Nous reviendrons sur la tâche d'extraction de la terminologie en section 7.1.2.

L'extraction d'information possède de nombreuses applications. Nous pouvons citer (de manière non exhaustive) l'information médicale, la veille technologique, l'indexation automatique d'articles ou encore le résumé automatique. Historiquement, l'extraction d'information peut être apparentée au début de l'intelligence artificielle. Nous cherchions à cette époque à créer des systèmes structurant toutes les données d'un texte écrit en langage naturel. Parmi les premiers travaux proposant réellement de l'extraction d'information, nous trouvons des systèmes d'extraction automatique de noms propres [Borkowski, 1969] ou encore l'extraction d'information issus de rapports de radiologie [Sager, 1981]. Dans les années 90, la recherche dans ce domaine s'est intensifiée. Cette période fut entre autres celle de l'essor grandissant de la communauté de compréhension des messages. Ainsi, la conférence américaine MUC (*Messages Understanding Conferences*) a été créée en 1993. Cette dernière propose d'effectuer une compétition entre divers systèmes d'extraction d'information. Parmi les tâches proposées, nous trouvons la reconnaissance d'entités nommées, la construction d'éléments de formulaire, la résolution de co-références ou encore la construction de formulaires de scénario. Notons que cette compétition est la seule du domaine d'extraction d'information ce qui en

fait une référence. Cependant, d'autres types de défis comme TREC ou DEFT[11] (DÉfi Fouille de Textes) – pour les données en français – proposent des épreuves d'extraction d'information mais ne se limitent pas à ce domaine.

2.3.3 Recherche documentaire (RD)

La recherche documentaire (RD) est, à l'instar de la classification de textes, une sous-discipline de la recherche d'information [Denoyer, 2004]. Cette tâche consiste à interroger un base de connaissance par le bais de requêtes écrites en langues naturelles ou bien sous forme de mots clefs (nommées requêtes *ad hoc*). La popularisation d'internet montre tout l'enjeu de cette discipline avec notamment les moteurs de recherche comme *Yahoo !* ou *Google*. Bien que nous ne nous soyons pas spécifiquement tournés vers cette tâche aux cours de nos travaux, cette discipline fait néanmoins intervenir la notion de descripteurs. L'objectif des nouveaux travaux menés sur le sujet est de produire le meilleur résultat possible en réponse à une requête émise. Afin d'améliorer ces systèmes, de nombreuses approches proposent d'effectuer un enrichissement de requêtes. L'objectif est évident, plus la requête émise est riche et ciblée, et plus la réponse fournie par l'outil de RD est susceptible d'être pertinente. Plusieurs types d'enrichissement sont proposés dans la littérature. Citons par exemple [Kießling, 2002] qui présente une méthode d'extension de langage de requête (ici le langage SQL et la syntaxe XPATH dédié aux documents XML). Par ailleurs, [Koutrika & Ioannidis, 2004] présentent une méthode d'enrichissement de requêtes de type langue naturelle en se fondant sur des ressources propres à l'utilisateur. Les descripteurs sont alors essentiels afin de décrire le contenu de la requête. Néanmoins, le type d'approche proposé dans ce manuscrit se fonde sur des informations syntaxiques issues de documents textuels. Ainsi, il n'est pas envisageable d'enrichir le contenu de requêtes *ad hoc* du fait de leur manque de structure syntaxique. Nous présentons en effet en section 4.1 une approche d'enrichissement de corpus se fondant sur le contexte syntaxique de ce dernier. Notons par ailleurs que nous approfondirons la tâche de sélection de descripteurs avec des données dépourvues de syntaxe dans le chapitre consacré aux données textuelles complexes (chapitre 5).
Outre l'importance des descripteurs pour la formulation de la requête, ces derniers sont également indispensables afin de décrire les connaissances contenues dans le modèle de RD. Cette dernière tâche s'effectue la plupart du temps par une étape d'indexation des documents, discipline à part entière de la recherche d'information. Cette étape consiste à sélectionner les mots clés les plus discriminants permettant de décrire au mieux des documents afin de répondre à des besoins de recherche documentaire. Le lien avec la sélection de descripteurs est ici implicite. De manière plus formelle, [Moulinier, 1996] définit la

[11] http ://deft.limsi.fr/

tâche d'indexation documentaire comme le fait "d'attribuer à un document des marques distinctives renseignant sur son contenu, en vue de le classer".
Les types de descripteurs les plus employés sont les lemmes et les radicaux [Bacchin *et al.*, 2005] mais ces derniers restent plus performants avec la langue anglaise [Braschler & Ripplinger, 2004]. Les auteurs de ce papier proposent cependant une approche permettant d'améliorer les performances de recherche documentaire avec l'allemand.
Les sélections de descripteurs les plus couramment réalisées sont de nature statistique. Un des systèmes de RD parmi les plus connus est sans doute SMART [Salton & Lesk, 1965] puis [Salton, 1971]. Les auteurs proposent dans ce cas une sélection statistique de type *tf* et *tf-idf* de descripteurs. Citons également le PageRank proposé par Larry Page pour *Google* qui est un algorithme statistique visant à faire ressortir les pages les plus pertinentes suite à une requête émise par un utilisateur du moteur de recherche *Google*. Les descripteurs sont notamment focalisés sur le nombre de liens hypertextes pointant vers des pages Web et le nombre de liens hypertextes vers lesquels ces pages pointent. Il existe également des approches utilisant des ressources externes. Citons par exemple [Spiteri, 2005] qui propose la construction d'un thésaurus en faisant intervenir des experts. Ces derniers doivent participer à des évaluations qui permettront de sélectionner les entrées du thésaurus afin d'améliorer les systèmes de recherche documentaire. Citons également [Maisonnasse *et al.*, 2008] qui motivent l'utilisation de l'expressivité et de sa modélisation afin d'améliorer les résultats de systèmes de recherche d'information. Ils expérimentent plusieurs types de sélection de descripteurs avec différents modèles de représentation de connaissances. Ils concluent sur le fait que la simple fréquence (*tf*) reste la sélection la plus satisfaisante. Comme nous allons le montrer pour d'autres tâches décrites dans ce manuscrit, la fréquence a souvent un bon comportement (cf. chapitre 7). Finalement, nous trouvons également des articles de la littérature utilisant une sélection morphosyntaxique comme [Song *et al.*, 2008] où les auteurs présentent une méthode utilisant des syntagmes de lemmes.

2.4 Discussion

Nous proposons dans cette section une synthèse des descripteurs présentés dans ce chapitre. Le tableau ci-dessous récapitule les descripteurs en fonction de leur type et de leur sélection. Nous montrons ainsi quel type de descripteurs peut bénéficier de telle ou telle sélection.

Nous montrons dans ce tableau que tous les descripteurs peuvent bénéficier d'une

Type \ Sélection	Statistique	Morphosyntaxique	Ressource externe
Flexion	Oui [Yang & Liu, 1999a]	Oui [Song *et al.*, 2008]	Oui/Non
Radical	Oui [Liao *et al.*, 2007]	Non	Non
Lemme	Oui [Sjöblom, 2002]	Oui/Non	Oui [Paolo *et al.*, 2004]
N-gramme de mots	Oui [Tan *et al.*, 2002]	Oui/Non	Oui/Non
N-gramme de caractères	Oui [Junker & Hoch, 1997]	Non	Non

TAB. 2.5 – Type de descripteurs en fonction de la sélection

sélection statistique. Notons cependant que la sélection statistique de descripteurs avec les n-grammes doit être adaptée à ces derniers.

La sélection morphosyntaxique ne peut être appliquée sur des descripteurs de type n-grammes, lemme ou radical. Il n'est pas possible en effet d'extraire des informations syntaxiques à partir de ces descripteurs. Notons cependant que nous pouvons extraire des informations syntaxiques avant la construction des n-grammes de caractères [Laroum *et al.*, 2009] comme nous le montrerons dans le chapitre 5. Nous montrerons également dans ce chapitre que la sélection morphosyntaxique ne peut pas être appliquée sur tous types de corpus, comme ceux dépourvus de structures syntaxiques. Ainsi, la sélection morphosyntaxique n'est possible qu'avec un corpus rédigé en langue naturelle construit avec des flexions. En effet, les outils tels que des analyseurs syntaxiques ne peuvent identifier de manière pertinente la structure syntaxique d'un corpus utilisant d'autres types de descripteurs. Cependant, l'extraction de structures simples (par exemple, la terminologie Nominale) peut être adaptée sur des corpus complexes ne respectant pas toujours les contraintes de la langue naturelle comme des logs [Saneifar *et al.*, 2009]. Précisons que certains étiqueteurs grammaticaux, tel que le TreeTagger [Schmid, 1995] décrit en section 5.2.2, associent les catégories lexicales aux lemmes, d'où l'entrée Oui/Non dans le tableau 2.5. Néanmoins, bien que la sélection de descripteurs soit possible dans certains cas à partir de lemmes, celle-ci nous semble peu pertinente.

La sélection de descripteurs se fondant sur des ressources externes impose un respect de format. La ressource utilisée doit en effet respecter le même format que celui du corpus étudié. La plupart du temps, les ressources externes sont construites à partir de lemmes. Nous pouvons cependant réaliser une lemmatisation avec un corpus contenant des formes flexionnelles. Alors, les ressources externes décrites par des lemmes pourront être utilisées. Par ailleurs, les n-grammes de mots sont constitués d'une succession de

flexions. Ainsi, après une étape de lemmatisation, nous pouvons utiliser ces ressources en sélectionnant par exemple les n-grammes de mots composés des lemmes pertinents tels que définis dans une ressource externe.

Comme nous l'avons montré dans ce chapitre, les descripteurs fondés sur la syntaxe sont assez peu utilisés dans la littérature. Nous proposons dans le chapitre suivant un modèle de sélection de descripteurs pertinents se fondant sur les propriétés syntaxiques d'un corpus. Les descripteurs ainsi extraits peuvent alors être employés pour les diverses tâches de fouille de textes présentées dans ce chapitre. Nous évaluerons la qualité des descripteurs dans le chapitre 4 pour des tâches de classification.

Chapitre 3

SelDe : identification de descripteurs fondée sur les connaissances syntaxiques

Sommaire

Nous présentons dans ce chapitre les modèles théoriques de diverses approches ayant été expérimentées dans le chapitre 4. Ces travaux ont été publiés dans [6 - ICDIM'08], [8 - CICLing'08], [9 - JADT'08], [11 - CIR'07], [15 - EGC'08] et [16 - INFORSID'07].

3.1 Introduction

Ce chapitre a pour objectif de présenter un modèle d'extraction de descripteurs nommé SelDe (**Sél**ection de **De**scripteurs). Ce modèle propose d'extraire des descripteurs "hybrides" dans le sens où ils sont obtenus par le biais de connaissances syntaxiques et sémantiques. Nous avons en effet présenté dans le précédent chapitre un certain nombre d'approches permettant de sélectionner des descripteurs en utilisant des méthodes syntaxiques et sémantiques. Notre motivation est de sélectionner des descripteurs de meilleure qualité par le biais d'une hybridation. L'architecture globale de ce modèle, qui sera détaillée dans la section 3.4 est la suivante :

1. Extraction de connaissances syntaxiques, des relations syntaxiques de type Verbe-Objet.
2. Utilisation d'une mesure statistique afin de déterminer la proximité sémantique des verbes provenant des relations précédemment extraites.
3. Sélection des verbes sémantiquement proches.
4. Sélection de certains objets de ces verbes comme descripteurs.

Les deux premières sections de ce chapitre seront consacrées aux outils sur lesquels nous nous sommes appuyés afin d'extraire nos descripteurs. Ainsi, nous présenterons dans un premier temps dans la section 3.2 l'analyseur syntaxique SYGFRAN développé par Jacques Chauché (premier point de l'énumération ci-dessus). Cet analyseur nous permet d'extraire d'un corpus écrit en français les relations syntaxiques qu'il contient. Nous consacrerons ensuite la section 3.3 à l'étude de la proximité sémantique de termes, en présentant par la même le système et la mesure d'ASIUM proposés par David Faure (deuxième et troisième point de l'énumération ci-dessus). Nous avons en effet sélectionné cette mesure statistique afin de mesurer la proximité sémantique de verbes.
Enfin, nous présenterons une vision plus globale du modèle SELDE et reviendrons de manière plus précise sur la sélection des descripteurs dans la section 3.4 (quatrième point de l'énumération ci-dessus).

3.2 L'analyse syntaxique

Nous définirons dans un premier temps "l'analyse syntaxique" puis présenterons alors une liste non exhaustive des analyseurs syntaxiques traitant le français pour finalement présenter l'analyseur syntaxique SYGFRAN, utilisé dans notre modèle d'extraction de descripteurs.

3.2.1 Définition

3.2.1.1 Approche générale

En considérant un *langage* comme un moyen de communication, l'*analyse syntaxique* peut être vue comme une procédure qui va décider si une phrase donnée appartient bien à un langage. La figure 3.1 illustre le contexte d'une analyse syntaxique. L'émetteur est ici l'entité qui va produire la phrase conformément à une *grammaire* propre à un langage et le récepteur celui qui doit reconnaître la phrase. Une grammaire peut-être définie comme un ensemble de règles qui vont caractériser la structure syntaxique d'un langage. Par exemple avec le *langage de programmation C*, une règle de grammaire peut permettre de définir un

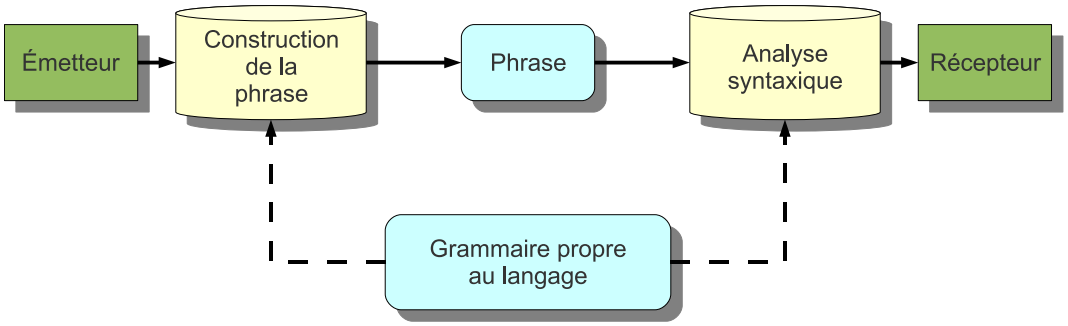

FIG. 3.1 – L'analyse syntaxique

caractère comme étant séparé par des simples quotes comme le caractère *'a'*. Finalement, la tâche de l'analyseur syntaxique consiste à produire une structure syntaxique permettant de vérifier si une phrase a bien respecté la grammaire d'un langage.

3.2.1.2 L'analyse syntaxique de données textuelles

L'analyse syntaxique de données textuelles, si elle vérifie également l'appartenance d'une phrase à un langage (par exemple une phrase formulée dans une langue naturelle comme le français), va également fournir d'autres informations concernant cette phrase. La valeur ajoutée d'un tel analyseur va être de décrire la structure syntaxique d'une phrase donnée en entrée. En d'autres termes, sa tâche va être de déterminer pour chaque terme de la phrase leur fonction syntaxique. Alors, une hiérarchie peut être établie en fonction des relations de dépendance syntaxique des éléments de la phrase (appelées des relations syntaxiques, comme "Sujet-Verbe" ou "Verbe-Objet"). Souvent, une telle hiérarchie est représentée sous forme d'arbre syntaxique.

3.2.2 Différents systèmes d'analyse syntaxique

Nous présentons dans cette section une liste non exhaustive d'analyseurs syntaxiques. Cette recherche bibliographique porte sur les analyseurs syntaxiques dits robustes ([Aït-Mokhtar & Chanod, 1997], [Grefenstette, 1999]), dont certains ont participé à la campagne d'évaluation de EASy[12]. Nous proposons ainsi dans un premier temps de décrire cette campagne d'évaluation, sur laquelle nous nous appuierons afin de montrer l'intérêt du choix de l'analyseur que nous avons sélectionné pour l'application de notre modèle de sélection de descripteurs.

[12] *http : //www.limsi.fr/Recherche/CORVAL/easy/*

3.2.2.1 La campagne d'évaluation Easy et le projet PASSAGE

La campagne d'évaluation organisée dans le cadre du projet *PASSAGE*[13] (ANR-06-MDCA-013) s'effectue dans la continuité de la campagne d'évaluation *EASy* du projet *EVALDA* (programme TECHNOLANGUE). Son but est la création d'une méthodologie d'évaluation des analyseurs syntaxiques du français et son application dans une campagne d'évaluation. De plus, cette campagne va également permettre de produire une ressource linguistique en combinant de manière automatique les données annotées.

Afin de comparer de manière détaillée les analyseurs syntaxiques en fonction de différents types de corpus et en fonction des différentes relations, des outils de mesure et des corpus spécifiques de domaines divers ont été créés (par exemple littérature, transcription de conversations, discours parlementaires, questions pour des moteurs de recherche).

À terme, les motivations de la proposition de l'ANR *PASSAGE* sont les suivantes :
– amélioration de la précision et de la robustesse des analyseurs syntaxiques de la langue française,
– exploitation des annotations syntaxiques résultantes des analyses afin de créer de nouvelles ressources linguistiques plus riches et plus extensives.

Afin de mener à bien ce projet, la méthodologie adoptée consiste en une "*boucle de rétroaction* (feedback)" entre analyseurs syntaxiques :
– création d'annotations syntaxiques via l'analyse syntaxique,
– utilisation de ces annotations afin de créer ou d'enrichir des ressources linguistiques (lexiques, grammaires ou corpus annotés),
– intégration des ressources créées ou enrichies sur la base des annotations dans les systèmes d'analyse,
– utilisation de ces analyseurs enrichis afin de créer des ressources encore plus riches (par exemple des ressources syntactico-sémantiques).
Finalement, le projet *PASSAGE* devrait aussi aider à faire émerger des chaînes de traitement linguistique exploitant des informations lexicales plus riches, en particulier sémantiques.

Une dizaine de systèmes d'analyse syntaxique a participé à ce projet et à la campagne d'évaluation *EASy* lancée en 2007. Nous présentons dans la section suivante des analyseurs ne traitant pas spécifiquement le français puis certains des analyseurs ayant

[13] *http* : *//atoll.inria.fr/passage/eval*1.*fr.html*

participé à cette campagne.

3.2.2.2 Les analyseurs syntaxiques

Les analyseurs dits "historiques"

Historiquement, l'analyseur syntaxique pouvait être décrit par un ensemble de règles visant à désambiguïser les termes polysémiques. Nous pouvons par exemple citer [Kelly & Stone, 1975] et [Small, 1980].

Le premier analyseur syntaxique présenté dans cette section fut conçu et développé entre 1958 et 1959 dans le cadre du projet *Transformations and Discourse Analysis Project* (TADP). Cet analyseur fut développé par L. Gleitman, A. Joshi, B. Kauffman et N. Sager puis C. Chomsky. Il a été réécrit dans les années 1990 par A. Joshi et P. Hopely [Joshi & Hopely, 1996]. Il possède la particularité d'être implémenté comme une *cascade de transducteurs* et fut, selon les auteurs qui l'ont ré-implémenté, le premier analyseur de ce type.

FULCRUM est un analyseur syntaxique proposé par Paul S. Garvin [Garvin, 1967] pour le russe. Une particularité de cet analyseur est qu'il est constitué d'un dictionnaire et d'un algorithme ce qui fut novateur à l'époque. En effet, les règles de grammaire étaient habituellement séparées des algorithmes d'analyse. Cet analyseur fonctionne par un système de passes pendant lesquelles est identifié un certain nombre de relations syntaxiques via la reconnaissance de patrons grammaticaux. La notion de *fulcrum* signifie les mots pivots à partir desquels est lancé une analyse locale sur les mots voisins dans la phrase. Cela permet de localiser une relation de dépendance syntaxique et ceci pour une fonction de recherche donnée (comme l'attachement d'adverbes à des adjectifs). Ainsi, le système ne traite pas à chaque passe linéairement les mots de la phrase analysée mais "se déplace" de mots pivots en mots pivots, faisant une analyse locale.

L'analyseur syntaxique *Cascaded Analysis of Syntactic Structure* (CASS) de Steven Abney est fondé sur la notion d'analyse en *cascade* de *Chunks*. Ce concept de chunks fut présenté dans la thèse de Steven Abney [Larson *et al.*, 1987]. Ils peuvent être définis comme les têtes sémantiques des types de groupes syntaxiques (NP, VP, PP, AP, AdvP). Une tête est un mot plein qui n'est pas situé entre un mot fonctionnel (déterminant, préposition, etc.) et le mot plein sélectionné par ce mot fonctionnel. Ainsi, un chunk est constitué de la séquence de mots situés entre le mot fonctionnel et le mot tête sélectionné en incluant ceux-ci. L'idée de l'analyseur est de reconnaître

dans un premier temps les chunks, puis de délimiter les propositions qui définissent les relations entre les chunks. Finalement, l'établissement des liens entre les chunks reflète le principe de l'analyse en cascade, consistant en une succession de passes. L'analyseur est décrit précisément dans [Abney & Abney, 1990], [Abney, 1991] et [Abney, 1996].

Ces premiers systèmes d'analyse syntaxique mettent en évidence la multiplicité des approches utilisées. En effet, plusieurs techniques peuvent être appliquées : emploi de transducteurs, notion de cascades, utilisation de patrons grammaticaux ou bien encore utilisation de *Chunks*. Nombre de ces approches restent utilisées de nos jours dans le cadre d'analyse syntaxique comme le montrent les approches présentées ci-dessous.

Les analyseurs de la campagne EASy

Nous présentons dans ce paragraphe divers analyseurs plus récents, ayant la particularité d'avoir pris part à la campagne EASy [Paroubek *et al.*, 2005], à l'instar de l'analyseur SYGFRAN que nous utilisons dans notre modèle de sélection de descripteurs.

L'analyseur de Jacques Vergne, *Vergne 98* est décomposé en deux étapes. Une première étape consiste en un étiquetage morphologique ou *tagging*. La seconde se fonde sur la notion de relations de dépendance. Elle consiste à placer des relations de dépendances syntaxiques entre les syntagmes non récursifs précédemment identifiés par l'étiquetage. Cet analyseur est décrit par son auteur dans [Vergne, 1990].

L'analyseur syntaxique *SYNTEX* développé par Bourigault et Farbre [Bourigault & Fabre, 2000] a dans un premier temps été produit pour enrichir l'outil *LEXTER* [Bourigault, 1994] dont l'objectif était la découverte de syntagmes nominaux terminologiques dans les corpus spécialisés. *SYNTEX* est un analyseur procédural à cascade dans le sens où il traite chaque séquence en plusieurs passes successives. La séquence donnée en entrée est étiquetée (à chaque mot est associée une catégorie grammaticale). Les liens syntaxiques sont alors placés par des heuristiques décrivant l'algorithme de parcours de la phrase étiquetée. Notons que cet analyseur a obtenu d'excellents résultats lors de la campagne d'évaluation *EASy*.

Le système *LIMA* (LIc2m Multilingual Analyzer) réalisé au laboratoire du LIC2M reprend le principe des dictionnaires *full-form* [Fluhr, 1997]. Ainsi, l'analyse est réalisée par des modules indépendants traitant la segmentation, les expressions figées, l'analyse morphologique, la désambiguïsation syntaxique, la reconnaissance des entités nommées, l'analyse syntaxique et la création de termes composés. L'analyseur pro-

prement dit implémente une grammaire de dépendance telle que les analyses soient exclusivement représentées par des relations de dépendance entre deux mots. L'analyse est effectuée en utilisant des automates à états finis. Cet analyseur est décrit dans [Besançon & Chalendar, 2005].

L'analyseur *LLP2* du *LORIA* (Laboratoire Lorrain de Recherche en Informatique et ses Applications), décrit dans [Roussanaly *et al.*, 2005] est fondé sur des grammaires d'arbres adjoints lexicalisés [Joshi *et al.*, 1975]. Il suit l'algorithme décrit dans [Lopez, 1999] traitant de l'analyse par connexité. Il intègre également un module de traitement de structures de traits d'unification. Ainsi, le corpus est dans un premier temps étiqueté puis, traité par un *lemmatizer* qui va relier les segments aux arbres élémentaires associés.

Ces différents analyseurs proposent des méthodes assez distinctes utilisant notamment l'étiquetage grammatical et la notion de cascade. La modélisation des grammaires de dépendance utilisées est effectuée par des arbres ou transducteurs. L'outil SYGMART présenté ci-dessous se fonde également sur la notion de grammaire. Cependant, à la différence d'autres systèmes d'analyse syntaxique, il permet d'analyser tout langage modélisable sous forme de transducteurs d'arbres. Nous présentons ci-dessous SYGFRAN, définissant entre autres des grammaires pour le programme SYGMART.

3.2.3 Le système SYGMART

3.2.3.1 SYGMART et SYGFRAN

Le système SYGMART (Système Grammatical de Manipulation Algorithmique et Récursive de Texte) est un système transformationnel prenant en entrée une chaîne de caractères et propose en sortie une structure arborescente. Il est à la base de l'analyseur syntaxique SYGFRAN que nous avons sélectionné afin de mettre en œuvre notre modèle. Nous présentons ainsi dans un premier temps le système SYGMART et ses différents sous-systèmes puis présentons l'analyseur syntaxique SYGFRAN, qui n'est autre qu'un ensemble de règles pour SYGMART.

3.2.3.2 Principe de SYGMART

SYGMART est un système qui se fonde sur les algorithmes de Markov, étendus aux arbres. Il permet d'analyser tout langage dont la grammaire pourrait être écrite sous forme de transducteurs d'arbres. Afin d'analyser une chaîne de caractères, le système SYGMART se fonde sur trois sous-systèmes OPALE, TELESI et AGATE. La forme générale de l'application du système SYGMART est représentée dans la figure 3.2.

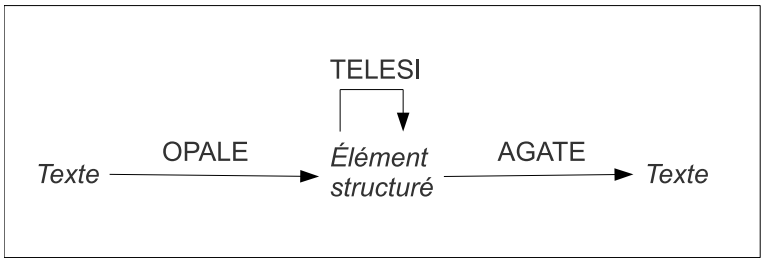

FIG. 3.2 – Application du système SYGMART

Un *élément structuré* est défini comme un quadruplet (P, S, E, F) tel que :
– P soit un ensemble fini de points
– S soit un ensemble fini de structures arborescentes sur les points P tel que chaque point de P appartienne à au moins une structure de S
– E soit un ensemble fini de multi-étiquettes
– F soit une application surjective de P sur E
La figure 3.3 présente un exemple d'élément structuré. Appliqué au problème de l'analyse

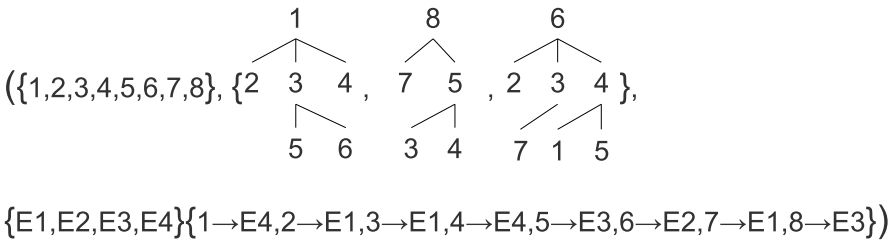

FIG. 3.3 – Exemple d'élément structuré

textuelle, le fait de définir une grammaire syntagmatique engendrant un langage peut se traduire par le fait de "*définir pour chaque élément du langage un élément structuré associé*". Alors, un élément structuré va être (en simplifiant) défini comme un quadruplet (P, S, E, F) tel que :
– P soit un ensemble fini de termes
– S soit un ensemble fini de structures arborescentes définissant des règles de grammaire
– E soit un ensemble fini de catégories ou fonctions syntaxiques
– F soit une application surjective liant les mots à des catégories ou fonctions syntaxiques
Ainsi, la multiplicité des structures associées au même ensemble de points va permettre de définir une association plus complexe [Chauché, 1984].

3.2.3.3 OPALE : le sous-système de décomposition morphologique

Le sous-système Opale permet la définition d'une transition entre un texte d'entrée et un élément structuré. Un transducteur d'états finis, construit sur la consultation d'un dictionnaire et la segmentation de la chaîne d'entrée, est ainsi utilisé afin d'effectuer la transition : *Texte d'entrée* → *Élément structuré*.
L'élément initial du sous-système Opale est donc formé d'un couple (élément structuré initial, chaîne) et propose en sortie un élément structuré dont la forme dépend de l'analyse de la chaîne. En utilisant les règles définies par l'analyseur syntaxique Sygfran pour le sous-système Opale de Sygmart, l'élément en entrée est la phrase à analyser et en sortie l'arbre produit par les règles Opale de Sygfran. La figure 3.4 fournit un exemple simplifié du rendu du sous-système Opale dans le cadre de l'analyse morphologique de la phrase *"La fête continue."*. Nous utilisons dans cette figure la représentation graphique

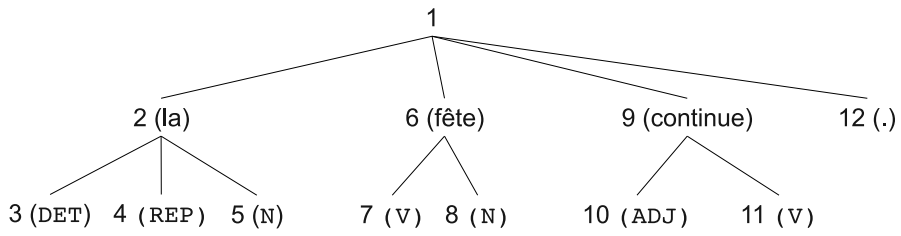

Fig. 3.4 – Exemple simplifié de sortie du sous-système Opale de Sygmart

de Sygfran afin de reporter le résultat de l'analyse. L'étiquetage des nœuds suit une numérotation en profondeur afin de lister les valeurs associées à chaque nœud. Afin de simplifier la lecture de l'exemple, nous avons reporté dans la figure 3.4 les valeurs des nœuds directement sur l'arbre.
Une ambiguïté sur la catégorie lexicale d'un mot conduit à créer autant de fils au nœud le représentant qu'il y a de catégories lexicales possibles pour ce mot. Par exemple, le mot "la" peut être un déterminant (DET), un pronom (REP pour représentant) ou bien encore un nom (N, ici la note de musique). L'ambiguïté n'est donc pas levée par le sous-système Opale qui se contente d'effectuer une analyse morphologique ne considérant pas les relations entre les mots. Cette tâche est dévolue au sous-système Telesi dans le cadre d'une analyse syntaxique. D'autres ambiguïtés comme le temps et/ou la personne des verbes sont également résolues par le sous-système Telesi (par exemple, le mot "fête" peut être à la première ou la troisième personne du présent de l'indicatif s'il est considéré comme un verbe).

3.2.3.4 TELESI : le sous-système de transformation d'éléments structurés

Principe.

Le système Telesi a pour but de définir une transition entre des éléments structurés. Telesi peut prendre en entrée la sortie d'une analyse faite par le sous-système Opale mais il peut prendre également la sortie d'un autre traitement pour finalement transformer l'entrée en la structure arborescente voulue tel que le montre la figure 3.2.

La transition entre les éléments structurés ainsi définis est effectuée par un transducteur à pile, simulant une grammaire transformationnelle qui tente de caractériser la connaissance de la langue permettant l'acte effectif du locuteur-auditeur. Le système Telesi est composé d'un système de réseau conditionnel de grammaires élémentaires. L'application sur un élément structuré est ainsi définie par un cheminement dans ce réseau conditionnel avec en chaque point une application d'une grammaire élémentaire. Une grammaire élémentaire est définie par un ensemble de règles et un mode d'application. Chaque grammaire décrit une transformation d'éléments structurés et le résultat de cette grammaire décrit le parcours du réseau. Le cheminement qui est conditionnel dépend de l'élément structuré initial. Ainsi, les grammaires du réseau qui ont été définies comme initiales vont donc servir de point de départ au cheminement. Les sorties du réseau sont repérées par des marqueurs et le système Telesi définira une transition comme terminée seulement après avoir atteint un de ces marqueurs.

Exemple de sortie de Telesi

La figure 3.5 montre la sortie du sous-système Telesi à qui on a fourni en entrée le

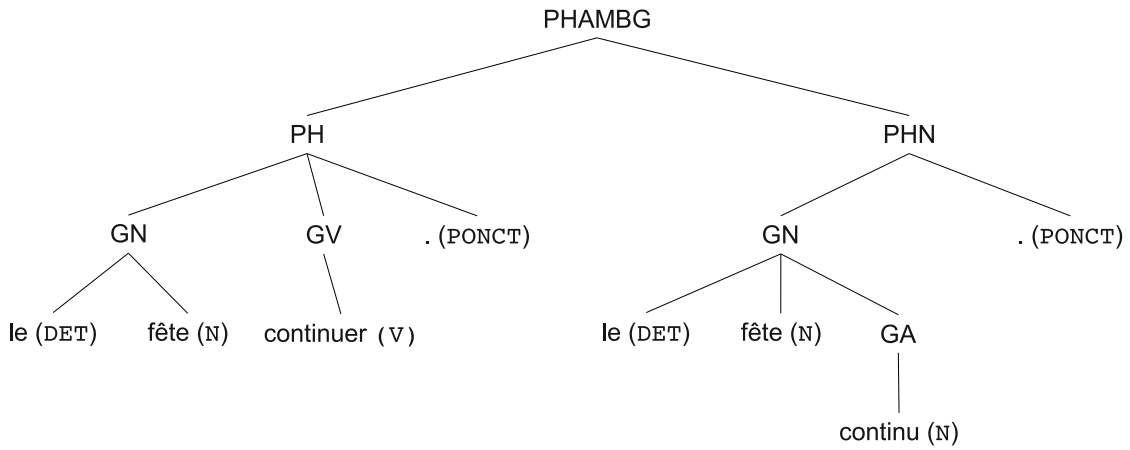

Fig. 3.5 – Exemple simplifié de sortie du sous-système Telesi de Sygmart

résultat de l'analyse du sous-système Opale pour la phrase *"La fête continue."*. À titre d'exemple, nous avons utilisé ici l'analyseur syntaxique Sygfran afin d'illustrer la sortie

produite par le sous-système Telesi de Sygmart. Notons que les instances des mots constituant la phrase ont été lemmatisés.
Le sous-système Telesi a par ailleurs levé les ambiguïtés sur les mots "la" et "fête" en les identifiant respectivement comme *déterminant* et *nom*. L'analyse du sous-système Telesi fait également apparaître les catégories lexicales des nœuds internes. Nous avons par exemple dans la figure 3.5 les mots "le" et "fête" qui appartiennent à la catégorie lexicale *GN* pour *Groupe Nominal*, lui même appartenant à la catégorie *PH* pour *PHrase*. Notons que dans cet exemple apparaissent les catégories *GV*, *GA*, *PHN* et *PHAMBG* qui correspondent respectivement aux catégories *Groupe Verbal*, *Groupe Adjectival*, *PHrase Nominale* et pour finir *PHrase AMBiGue*. En effet, nous remarquons sur la figure que deux analyses de la phrase sont proposées, venant de l'ambiguïté du mot "continue" qui n'a pu être levée par l'analyseur. Cette ambiguïté qui montre les limites de l'analyseur n'est cependant pas résolvable par un être humain qui sera incapable de dire, sans avoir connaissance du contexte de la phrase dans un texte, quelle serait l'analyse syntaxique la plus pertinente.

3.2.3.5 AGATE : le sous-système de linéarisation d'éléments structurés

Le but du sous-système Agate est de définir une transition entre une multi-étiquette associée à un point d'un élément structuré et une chaîne de caractères. Cette transition s'effectue par le parcours canonique d'une arborescence d'un champ déterminé. Ce sous-système est, comme le sous-système Opale, constitué d'un transducteur d'états finis non déterministes dont chaque étiquette associée à un point de ce parcours va définir un mot. À la différence d'Opale, le sous-système Agate fournit uniquement la première solution possible dans la recherche générale non déterministe.

3.2.4 L'analyseur morpho-syntaxique SYGFRAN

Sygfran est un programme pour le système opérationnel Sygmart. Il est constitué d'un ensemble de règles (grammaires et entrée de dictionnaires pour le sous-système Telesi et entrée de dictionnaires uniquement pour les sous-systèmes Opale et Agate) dans le but de produire une analyse morpho-syntaxique de la langue française (plus de 11 000 règles à ce jour). La figure 3.6 schématise l'interaction des règles de Sygfran pour les différents sous-systèmes de Sygmart. Ainsi, le sous-système Opale produit une analyse morphologique. Il peut par exemple identifier les formes contractées d'articles définis. L'article "*du*" est alors transformé en sa forme décomposée par exemple "*de le*". Le sous-système Telesi effectue quant à lui l'analyse syntaxique proprement dite. Ainsi, les règles décrites pour le sous-système Telesi visent à reproduire le plus fidèlement possible la grammaire française. Une (ou des) règle(s) va par exemple permettre d'établir

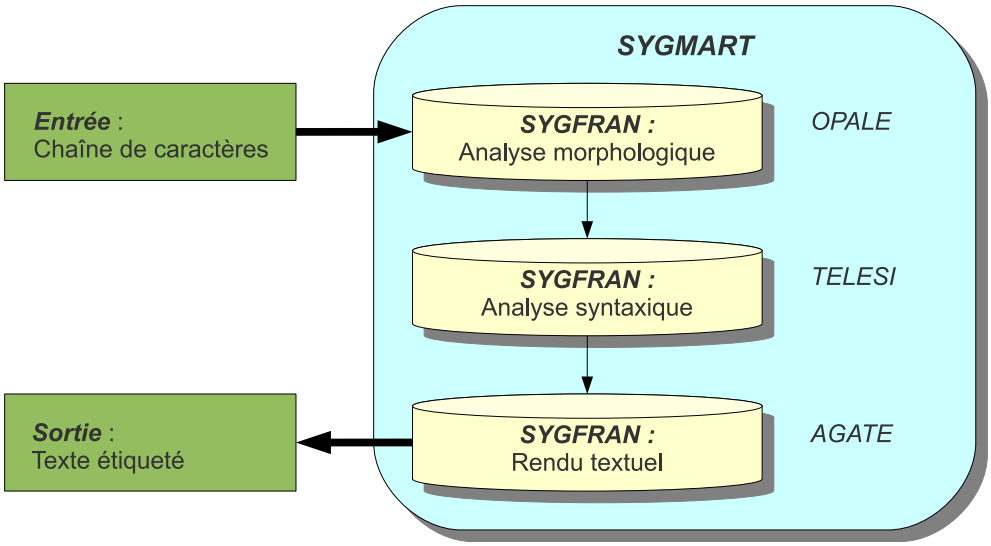

FIG. 3.6 – Le programme SYGFRAN utilisant le système SYGMART

les fonctions syntaxiques des constituants. Finalement le sous-système AGATE permet de récupérer un extrait ou la totalité de l'analyse effectuée sous forme textuelle. Par exemple, un ensemble de règles du sous-système AGATE va permettre de parcourir les feuilles de l'arbre résultant de la sortie du sous-système TELESI en extrayant uniquement les lemmes. Nous ne décrivons pas dans ce mémoire les grammaires et dictionnaires du programme SYGFRAN qui sont détaillées dans [Chauché, 2007].

Pourquoi utiliser Sygfran ?

Parmi les avantages de SYGFRAN notons :

– *La robustesse de l'analyse syntaxique.* Un analyseur syntaxique peut être qualifié de robuste s'il est relativement stable, même en présence de données erronées. SYGFRAN propose en effet, lors d'ambiguïtés sur l'analyse syntaxique d'une phrase, les différentes variantes de son analyse.

– *La vitesse d'exécution de l'analyse syntaxique.* La complexité de l'analyseur syntaxique, avec n la taille de la donnée et m le nombre de règles, en $O(m \times n \times log_2(n))$ est en effet relativement faible.

Notons finalement que SYGFRAN a obtenu en moyenne une précision[14] de $0,7$ lors de la campagne d'évaluation *EASy*, dont le principe et les objectifs sont présentés dans la section 3.2.2.1, ce qui est adapté à notre problématique d'extraction de descripteurs pertinents. En effet, notre objectif avec cette analyse syntaxique est d'obtenir un maximum de relations syntaxiques pertinentes en favorisant la précision au détriment

[14]Le protocole d'évaluation de la campagne d'évaluation *EASy* ainsi que la définition des mesures de rappel, de précision et de f-score sont explicités dans ce fichier téléchargeable à l'adresse suivante (au 21 juillet 2009) – *http : //www.limsi.fr/Recherche/CORVAL/easy/easy_eval_measures.ps*

du rappel. Notons finalement que pour des raisons de confidentialité imposées par les organisateurs de la campagne EASy, nous ne pouvons divulguer la position de SYGFRAN par rapport aux autres analyseurs.

Après avoir présenté l'analyseur syntaxique que nous utilisons afin d'extraire les relations syntaxiques d'un corpus, nous devons alors sélectionner uniquement les couples de verbes (provenant des relations de type Verbe-Objet extraites avec SYGFRAN) sémantiquement proches. Afin de sélectionner ces verbes, nous devons effectuer une étude de la proximité sémantique de termes, avec une mesure de similarité définie dans la section suivante.

3.3 L'étude de la proximité sémantique de termes

3.3.1 De la syntaxe aux connaissances sémantiques

3.3.1.1 Comment utiliser la syntaxe ?

L'acquisition de connaissances sémantiques est une importante problématique en Traitement Automatique des Langues (TAL). Ces connaissances peuvent par exemple être utilisées pour extraire des informations dans les textes ou pour la classification de documents. Nous proposons dans ce mémoire d'utiliser ces connaissances afin de sélectionner des descripteurs de textes, essentiels pour différentes tâches de fouille de textes comme montré dans le chapitre 2. De telles connaissances sémantiques peuvent être obtenues par des informations syntaxiques [Fabre & Bourigault, 2006]. Comme nous allons le montrer dans ce mémoire (en section 7.1.3), les connaissances sémantiques acquises via la syntaxe peuvent permettre de constituer des classes conceptuelles (regroupement de mots ou de termes sous forme de concepts). Par exemple, les mots *hangar*, *maison* et *mas* sont regroupés dans un concept *bâtiment*. Outre la construction de classes conceptuelles, ces connaissances sémantiques peuvent également permettre d'effectuer une expansion de corpus par exemple pour des tâches de classification de documents. Une telle expansion peut consister à enrichir un corpus avec de nouveaux termes.
Afin d'acquérir de telles connaissances sémantiques par la syntaxe, nous utilisons les relations syntaxiques précédemment extraites avec SYGFRAN.
Deux types de relations syntaxiques peuvent être utilisés pour construire les classes sémantiques : les relations issues d'une analyse syntaxique ([Lin, 1998] et [Wermter & Hahn, 2004]) et d'autres types de relations non originalement présentes dans le texte tel que l'introduit David Faure [Faure, 2000] en proposant le système ASIUM. Cette notion sera explicitée dans la section 3.3.2 puis approfondie dans le chapitre 6.

La méthode de David Faure dans le cadre du système ASIUM consiste à regrouper les objets des verbes déterminés comme proches par une mesure de qualité. D'autres approches utilisent également ce principe, comme le système UPERY [Bourigault, 2002] qui regroupe les termes par des mesures de *proximité distributionnelle*. Nous proposons dans la section suivante notre définition de la "*proximité sémantique*" et le lien établi avec l'*analyse distributionnelle*.

3.3.1.2 La notion de proximité sémantique liée à l'analyse distributionnelle

Nous définissons la "*proximité sémantique*" entre deux mots comme étant l'étude de la similarité du contexte syntaxique des mots. Ainsi, nous émettons l'hypothèse distributionnaliste de [Harris, 1951], [Harris, 1968] à savoir que *le repérage de structures syntaxiques régulières permet de mettre en évidence des familles de mots apparaissant dans des contextes communs*. L'analyse distributionnelle automatique se fondant sur cette hypothèse a donné lieu à de nombreux travaux dans le domaine du traitement automatique des langues naturelles sur des corpus spécialisés. Ces travaux ont principalement permis la construction de ressources terminologiques comme les travaux de [Habert & Nazarenko, 1996] en proposant un outil d'analyse distributionnelle automatique ZELLIG ou bien ceux de [Bourigault, 2002] et du système UPERY précédemment évoqués. Il existe de nombreux autres travaux effectuant une telle analyse pour l'acquisition de ressources terminologiques ou ontologiques à partir de textes. Citons par exemple [Bourigault & Lame, 2002] dans le domaine du droit et [Nazarenko *et al.*, 2001] dans le domaine biomédical. [Aussenac-Gilles & Jacques, 2006] proposent par ailleurs de produire et d'évaluer l'impact de patrons grammaticaux pour l'enrichissement d'ontologies. Certains travaux de la littérature proposent également d'effectuer des analyses distributionnelles de corpus non spécialisés comme [Galy & Bourigault, 2005].
Les travaux présentés dans ce mémoire s'inscrivent dans une optique d'extraction de connaissances, pouvant aboutir à l'enrichissement de contextes ou bien à la construction de classes conceptuelles. Nous présentons dans la section suivante le système ASIUM dans sa globalité pour finalement détailler la mesure d'ASIUM, que nous avons utilisée afin d'évaluer la proximité sémantique de nos relations syntaxiques.

3.3.2 Présentation générale du système ASIUM

Le système ASIUM (Aquisition of SemantIc knowledge Using Machine learning methods) fournit une méthode d'apprentissage automatique, symbolique, coopératif, empirique et non supervisé dans le but d'acquérir des connaissances à partir de données textuelles. Le principe du système ASIUM est d'utiliser des relations syntaxiques entre les verbes et les différentes têtes de leurs compléments afin d'extraire les exemples qui vont ainsi être

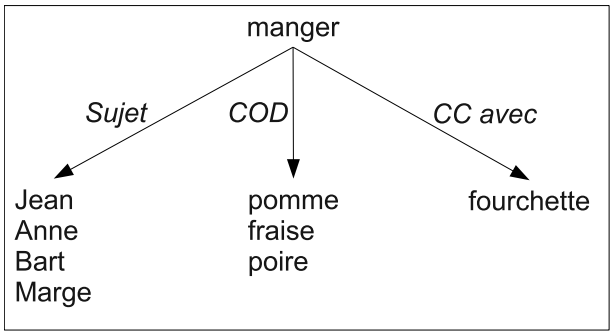

FIG. 3.7 – Exemple de graphe syntaxique

fournis au système d'apprentissage. L'objectif est alors de construire un graphe syntaxique qui sera composé d'un ensemble de parties d'arbres syntaxiques. Chaque proposition de relation syntaxique est représentée sous forme d'un arbre syntaxique comportant des arcs étiquetés avec les relations. Le graphe global regroupe finalement les différents arbres syntaxiques. Un exemple de graphe syntaxique (extrait de la thèse de D. Faure [Faure, 2000]) est donné dans la figure 3.7. Il peut être vu comme un ensemble d'arêtes liant des verbes et des noms. Ces arêtes sont étiquetées avec une fonction grammaticale et une préposition (si elle existe).
Vient alors la phase d'apprentissage du système qui prend en entrée les exemples tels que décrits précédemment. Alors, le système de classification rassemble tous les noms décrivant le même objet, ayant ainsi le même contexte syntaxique. Sont alors définies les classes conceptuelles dites de base par les auteurs du système. Elles sont constituées des noms issus des relations syntaxiques préalablement extraites d'un corpus. Finalement, il est agrégé deux ensembles de noms (donc deux classes de base) s'ils sont jugés proches par une mesure de similarité décrite dans la section suivante. Une généralisation des deux classes est alors effectuée avec l'ajout de connaissances nouvelles de types inductives qui seront au préalable validées par un expert humain. Nous reviendrons sur ces connaissances dans les sections 3.4.4 et dans le chapitre 6, traitant précisément ces connaissances. La phase d'apprentissage est ainsi répétée tant qu'il reste des nouvelles classes (classes formées par l'agrégation de deux classes).

3.3.3 La mesure d'ASIUM

3.3.3.1 Définition générale

La mesure d'ASIUM propose de calculer la proximité sémantique entre deux objets selon les descripteurs qui les décrivent. Le principe est de considérer comme proches deux objets ainsi définis s'ils partagent un nombre important de descripteurs en communs.

L'intérêt principal de la mesure d'Asium est d'éviter le *phénomène d'attraction* qui peut se présenter dans le cas d'utilisation de distances basées sur celle de Hamming[15]. En effet si une classe comporte un descripteur très fréquent et un nombre important de descripteurs peu fréquents, et qu'une autre classe dispose du même descripteur de manière très fréquente, la mesure d'Asium ne les considèrera pas comme proches en prenant en compte les descripteurs peu fréquents. La mesure d'Asium mesurant la "distance[16]" séparant les objets $O1$ et $O2$ est définie par l'équation suivante telle que décrite dans [Faure, 2000].

$$Asium_{log}(O1, O2) = 1 - \frac{log_{Asium}(\sum FOC_{O1}) + log_{Asium}(\sum FOC_{O2})}{log_{Asium}(\sum_{i=1}^{card(O1)} f(desc_{iO1})) + log_{Asium}(\sum_{i=1}^{card(02)} f(desc_{iO2}))}$$

$f(desc_{iOj})$ représente la fréquence d'apparition du descripteur $desc_i$ pour l'objet O_j. FOC_{O1} (respectivement FOC_{O2}) représente la somme des fréquences des descripteurs de $O1$ (resp. $O2$) décrivant aussi $O2$ (resp. $O1$).
Avec $log_{Asium}(x)$ valant :

- pour $x = 0$, $log_{Asium}(x) = 0$
- sinon $log_{Asium}(x) = log(x) + 1$

La fonction log_{Asium} est égale à la fonction $log()$ à un décalage près.
Cette mesure se focalise, comme d'autres mesures couramment utilisées dans la littérature (cosinus, information mutuelle, etc.) sur le fait de pondérer le nombre de descripteurs *communs* des objets par le nombre de descripteurs qui leur est propre. Ainsi, un couple d'objets possédant une grande quantité de descripteurs en commun, n'aura qu'un poids réduit si l'un ou les deux objets considérés possèdent un nombre encore plus important de descripteurs isolés (non commun aux deux objets). À l'inverse, si deux objets possèdent un faible nombre de descripteurs communs, mais que ces derniers sont tous ou presque communs aux deux objets, alors le score de proximité sera très élevé entre les deux objets. Notons cependant que cette mesure se distingue des autres de par son numérateur, traitant les descripteurs communs. En effet, les descripteurs communs d'un objet1 ne sont pas les mêmes que les descripteurs communs d'un objet2. Ainsi, cette mesure prend en compte toutes les occurrences des descripteurs *communs* des deux objets mais également séparément les occurrences des descripteurs propres à chaque objet. Les auteurs de la mesure évaluent alors la proximité de tous les couples d'objets d'un corpus. C'est alors

[15]Richard Hamming, principalement connu pour la mise en place de son code éponyme, a également proposé une mesure qui peut être définie comme le nombre distinct de bits dans une comparaison bit-wise, servant à la correction de certaines erreurs de transmission.

[16]La mesure d'Asium n'est pas une distance au sens mathématique du terme, les trois propriétés définissants une telle distance n'ayant pas été démontrées.

avec les informations peu fréquentes que le système ASIUM produit un maximum d'informations[17] nouvelles (de type inductives). Nous reviendrons sur ces informations en section 3.4.3. Nous présentons dès lors dans la section suivante la mesure d'ASIUM, telle que nous l'avons adaptée pour notre modèle de sélection de descripteurs SELDE en justifiant dans un premier temps notre choix d'utiliser des relations syntaxiques de type Verbe-Objet.

3.3.3.2 Le choix des relations syntaxiques de type Verbe-Objet

Cette section s'intéresse au choix des relations syntaxiques de type *Verbe-Objet* dans notre modèle de sélection de descripteurs SELDE.
La terminologie, pouvant être définie comme la description du vocabulaire spécialisé d'un domaine, s'appuie sur les noms ou les formes nominales d'un corpus de spécialité afin de définir des concepts décrivant un domaine spécifique [Sager, 1990]. Notons cependant que la simple information nominale n'est pas toujours adaptée afin de décrire un domaine. Prenons par exemple le domaine de l'informatique où des noms tels que "*disquette*", "*clavier*", etc. sont très spécifiques au domaine, mais il existe également des noms comme *disque*, qui, considérés seuls (c'est-à-dire hors de tout contexte), ne peuvent permettre d'identifier ce domaine.
Par ailleurs, [L'Homme, 1998] a montré que les informations verbales issues des corpus de spécialité peuvent également permettre de le décrire. Citons par exemple les verbes "*cliquer*", "*implémenter*", etc. qui sont assez caractéristiques du domaine de l'informatique. Notons également que certains de ces verbes ne peuvent être à eux seuls représentatifs du domaine de l'informatique, à l'instar des informations nominales. Citons le verbe "*formater*" par exemple.
Nous nous sommes ainsi intéressés à l'apport des relations syntaxiques afin de décrire un domaine de spécialité. Ces relations sont en effet largement utilisées dans la littérature comme dans les travaux de [Kim, 2008] ou encore [Shen *et al.*, 2005], précédemment évoqués en section 2.1.2.2. Nous distinguons trois entités principales propres aux relations syntaxiques : les *sujets*, les *verbes* et les *objets*. Les informations relatives au sujet d'un groupe de mots ou d'une phrase sont assez limitées, informations étant souvent formulées sous la forme d'*entités nommées* ou bien de *pronoms*. Notons que les pronoms sont en effet très largement utilisés dans des relations de type Sujet-Verbe comme par exemple dans le corpus lié aux Ressources Humaines que nous avons étudié au cours de ces travaux de thèse. Dans ce cas, l'identification du sujet peut être effectuée par une tâche supplémentaire de recherche d'anaphores [Cornish *et al.*, 2005], qui peut être perfectible.
Ainsi, il nous a semblé plus pertinent d'utiliser des relations syntaxiques de type Verbe-

[17] La notion d'induction d'information consiste dans ce cas à associer un verbe avec des objets auxquels il n'était pas préalablement associé.

Objet. Ces relations peuvent en effet désambiguïser certains noms ou verbes ambigus du domaine. Citons par exemple la relation Verbe-Objet : *Verbe* : "Formater", *Objet* : "Disque". L'intérêt de l'utilisation des ce type de relations est double. D'une part, elle ne nécessite pas de traitements supplémentaires comme les relations de type Sujet-Verbe, et d'autre part, elle peuvent permettre d'acquérir des informations sémantiques supplémentaires.

3.3.3.3 La mesure d'ASIUM appliquée à notre problématique

Ainsi, une fois adaptée à notre modèle (c'est à dire en utilisant des relations syntaxiques de type Verbe-Objet), nous présentons dans la figure 3.8 la mesure d'Asium permettant de mesurer la proximité sémantique de deux verbes.
Soient p et q, deux verbes avec leurs objets respectifs $p_1, ..., p_n$ et $q_1, ..., q_n$ illustrés sur la figure 3.8. $NbOccCom_p(q_i)$ représente le nombre d'occurrences des objets q_i en relation

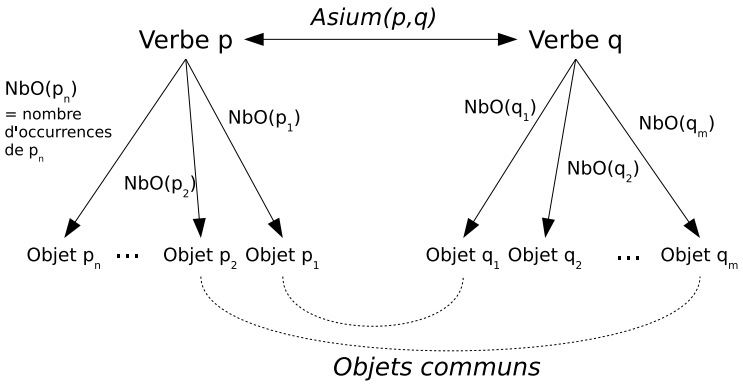

Fig. 3.8 – Score d'Asium entre les verbes p et q

avec le verbe q qui sont aussi des objets du verbe p, $NbOcc(q_i)$ représente le nombre d'occurrences des objets q_i. La mesure d'Asium est définie de la manière suivante :

$$Asium(p,q) = \frac{log_{Asium}(\sum NbOccCom_q(p_i)) + log_{Asium}(\sum NbOccCom_p(q_i))}{log_{Asium}(\sum NbOcc(p_i)) + log_{Asium}(\sum NbOcc(q_i))}$$

Une telle mesure fournit en sortie un score de proximité appartenant à l'ensemble $[0, 1]$.
– Un score de 1 signifie que chaque objet d'un verbe est également objet de l'autre verbe.
– Un score de 0 signifie que les deux verbes du couple considéré ne partagent aucun objet en commun. Nous présentons ci-dessous un exemple de calcul de proximité en utilisant la mesure d'Asium.

Exemple illustrant la mesure d'ASIUM

Afin de présenter un exemple de calcul utilisant la mesure d'ASIUM, nous nous appuyons sur la figure 3.9 illustrant le calcul de proximité de deux verbes : *écouter* et *convaincre*. Notons que le nombre d'occurrences de chaque objet de verbes y est spécifié. Le calcul de la mesure d'ASIUM entre ces deux verbes est alors le suivant.

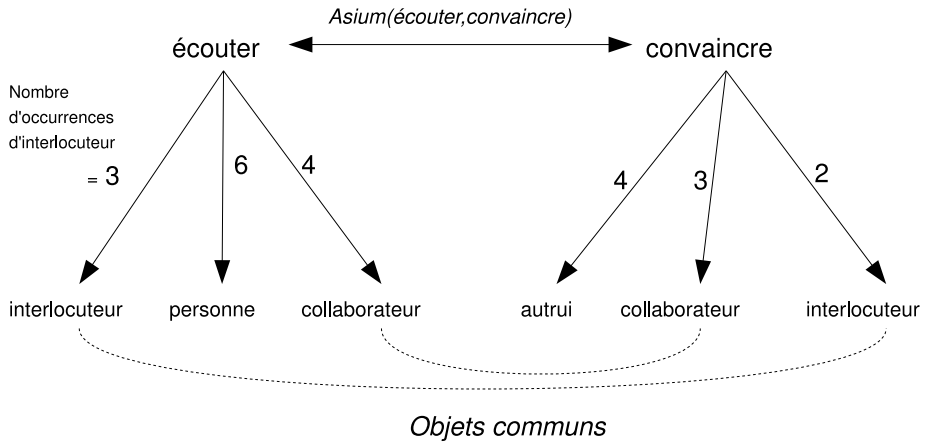

FIG. 3.9 – Mesure d'Asium entre les verbes écouter et convaincre

$$Asium(écouter, convaincre) = \frac{log_{Asium}(4+3) + log_{Asium}(3+2)}{log_{Asium}(4+6+3) + log_{Asium}(3+4+2)} = 0,87$$

Ces deux verbes sont alors dans cet exemple considérés comme assez proches par la mesure d'ASIUM.
Nous proposons dans la section suivante de discuter de la qualité de cette mesure de proximité sémantique en la comparant à diverses mesures, couramment utilisées dans la littérature.

3.3.4 Discussions sur le comportement de la mesure ASIUM

L'objectif de cette section est de comparer la mesure d'ASIUM avec des mesures couramment employées dans la littérature. Nous motiverons ainsi notre choix pour la mesure d'ASIUM en conclusion de cette section (section 3.3.4.3).
Nous utilisons la mesure d'ASIUM afin d'obtenir un classement des couples de verbes extraits avec l'analyseur syntaxique SYGFRAN en termes de proximité sémantique. Notre objectif est alors de sélectionner les couples de verbes les plus proches sémantiquement. Nous proposons alors de comparer le classement de ces couples effectué avec la mesure d'ASIUM avec le classement obtenu avec d'autres mesures de proxim-

ité fréquemment utilisées dans la littérature (notamment en recherche d'information [Frakes & Baeza-Yates, 1992]) :
– Le contexte partagé
– Le cosinus
– Le coefficient de Jaccard
– Le coefficient de Dice
– L'Information Mutuelle (IM)
– L'Information Mutuelle au cube (IM^3)

Notons que nous présentons ici uniquement des mesures pouvant être adaptées aux données textuelles que nous manipulons. En effet, ces données textuelles n'ont pas d'opposé au sens mathématique du terme. Par exemple, nous ne pouvons définir l'absence d'un verbe comme "*fournir*" dans un corpus donné. Ainsi, des mesures statistiques nécessitant des négations comme la *J-mesure*, pouvant être utilisée comme mesure de qualité pour des règles d'associations (comme par exemple dans [Lallich & Teytaud, 2004]), ne peuvent être appliquées avec notre approche et nos données textuelles. Par ailleurs, certaines de ces mesures ont également été décrites dans le chapitre 2 dans un contexte différent.

3.3.4.1 Définition des mesures de proximités

Notons en préambule de ce paragraphe que ces mesures peuvent être utilisées en prenant en compte la **fréquence** des descripteurs (ici les mots), c'est-à-dire leur nombre d'occurrences dans une relation syntaxique donnée. En effet, un objet peut apparaître plusieurs fois avec le même verbe, formant plusieurs occurrences d'une même relation syntaxique. Nous noterons alors une mesure utilisant la fréquence des relations avec l'indice "Freq" par opposition à "Bin" pour une mesure **binaire** (ou booléenne). Dans ce cas binaire, les occurrences des relations syntaxiques ne seront pas considérées. Ainsi, une même relation syntaxique ayant un grand nombre d'occurrences dans un corpus ne sera comptée qu'une seule fois. Finalement, suivant le type de mesures utilisées, le contexte propre à un verbe et le contexte partagé de deux verbes tels que présentés ci-dessous tiendront compte du nombre d'occurrences des relations syntaxiques ou du nombre distinct de relations. Rappelons que notre objectif est de situer la qualité du classement des relations syntaxiques avec la mesure d'Asium par rapport aux autres mesures de la littérature. Pour cela, nous allons comparer le classement des couples de verbes effectué avec la mesure d'Asium aux classements obtenus en utilisant d'autres mesures. Le nombre de couples communs parmi les n premiers nous donnera alors un

score de correspondance.

Ainsi, pour la suite de cette section, nous allons nous appuyer sur les notations ci-dessous pour définir les mesures de qualités avec lesquels nous allons comparer la mesure d'ASIUM.
– a : contexte syntaxique partagé par les deux verbes d'un couple
– $c1$ et $c2$: contexte propre à un premier mot et respectivement à un second

Dans le cas binaire, la quantité a est donc le nombre d'objets distincts partagés par les deux verbes d'un couple donné. $c1$ et $c2$ vont quant à elles représenter le nombre d'objets distincts d'un premier et d'un second verbe.
Dans le cas fréquentiel, a est le nombre d'occurrences d'objets partagés par deux verbes. Par exemple, si un verbe a deux occurrences d'un objet et un autre verbe en possède trois, le résultat donné par a entre ces deux verbes sera égal à deux. Par ailleurs, les notations $c1$ et $c2$, représentent le nombre d'occurrences d'objets d'un premier et d'un second verbe.

Une fois ces quantités définies, nous présentons ci-dessous les mesures expérimentées dans cette section. Nous nous appuierons alors sur l'exemple de la figure 3.9 (section 3.3.3.3) présentant les verbes "*écouter*" et "*convaincre*" avec leurs objets respectifs afin de montrer un exemple de calcul employant ces mesures dans la section 3.3.4.2.

Le contexte partagé

Nous entendons par contexte partagé le nombre d'objets partagés par les deux verbes d'un couple donné. Ainsi, plus un couple de verbes a d'objets en communs et plus ils seront considérés comme sémantiquement proches. Cette mesure est représentée par la notation "a" telle que précédemment définie :

$$Sim_{SimpleMatching} = Nombre\ d'objets\ partagés\ par\ deux\ verbes \tag{3.1}$$

Le coefficient de Jaccard

Rappelons que cette mesure présentée précédemment dans le chapitre 2 est définie comme le rapport entre la taille de l'intersection des deux ensembles et la taille de l'union de deux ensembles. Adaptée à la mesure de la proximité sémantique pour notre approche, la mesure est définie ainsi :

$$Sim_{Jaccard} = \frac{a}{c1 + c2 - a} \tag{3.2}$$

Notons que le coefficient de Jaccard n'est pas employé traditionnellement de manière fréquentielle. La plupart du temps, cette mesure est utilisée afin de mesurer la proximité de vecteurs binaires en termes de correspondance de bits.

Le coefficient de Dice

Le coefficient de Dice [Smadja *et al.*, 1996] évoqué dans le chapitre 2 est assez proche du coefficient de Jaccard. Il propose de mesurer le rapport entre deux fois la taille de l'intersection de deux ensembles et la somme de la taille respective des deux ensembles. Adaptée à la mesure de la proximité sémantique pour notre approche, la mesure est définie comme suit :

$$Sim_{Dice} = \frac{2 \times a}{c1 + c2} \tag{3.3}$$

Notons que le coefficient de Dice a tendance à moins pénaliser les cas où peu de mots sont communs aux deux termes (*i.e.* peu d'objets sont propres aux deux verbes du couple considéré).

Le cosinus

Le cosinus, que nous avons détaillé section 2.2.5.2 est une mesure fréquemment employée afin de mesurer la proximité sémantique de termes. Elle est assez similaire à celle de Dice. Notons que le cosinus a tendance à moins pénaliser les ensembles de tailles très différentes. Par exemple, un couple de verbes formé d'un verbe possédant beaucoup d'objets et d'un second en possédant très peu ne sera pas pénalisé par cet écart, réduit par la racine carrée. Le cosinus est défini ainsi :

$$Sim_{Cosine} = \frac{a}{\sqrt{c1 \times c2}} \tag{3.4}$$

L'Information Mutuelle

La "*Pointwise Mutual Information (PMI)*" [Church & Hanks, 1989], [Church *et al.*, 1991] est une mesure dérivée de l'information mutuelle déjà évoquée en section 2.1.2.1 dans le cadre de sélection statistique de descripteurs. Celle-ci propose d'évaluer la dépendance statistique de deux variables aléatoires.

$$IM(x, y) \quad = \quad \log_2 \frac{P(x, y)}{P(x)P(y)} \tag{3.5}$$

Partant de l'information mutuelle "classique" décrite dans l'équation 3.5, cette mesure est définie ainsi, en l'adaptant pour notre approche :

$$Sim_{IM} = \frac{a}{c1 \times c2} \tag{3.6}$$

L'idée de l'information mutuelle (*c'est-à-dire* de *PMI*) est de faire ressortir les co-occurrences les plus rares et les plus spécifiques [Daille, 1996], [Thanopoulos *et al.*, 2002]. Notons que le logarithme a été supprimé dans l'équation 3.6 par rapport à l'information mutuelle originale. Cette fonction étant strictement croissante, la suppression du logarithme n'a pas d'incidence sur le classement obtenu avec cette mesure.

L'Information Mutuelle au cube

L'information mutuelle au cube est une information empirique fondée sur l'information mutuelle, qui accentue l'impact des co-occurrences fréquentes, ce qui n'est pas le cas avec l'information mutuelle originale [Daille, 1994]. Cette mesure, une fois adaptée à notre problématique, est définie ainsi :

$$Sim_{IM^3} = \frac{a^3}{c1 \times c2} \tag{3.7}$$

3.3.4.2 Exemple de calcul des mesures

Après avoir présenté les différentes mesures de similarité que nous souhaitons comparer avec celle d'Asium, nous proposons dans cette section, avant la comparaison proprement dite, un exemple de calcul de ces scores de similarité. Nous nous appuyons sur l'exemple des verbes "*écouter*" et "*convaincre*", (cf. figure 3.3.3.3 dans la section 3.3.4.2).

Nous proposons dans un premier temps de calculer les quantités "a", "$c1$" et "$c2$" précédemment définies.

En binaire :

– $a(écouter, convaincre) = 2$

– $c1(écouter) = 3$

– $c2(convaincre) = 3$

Nous avons effectivement seulement deux objets en commun entre ces deux verbes. Les deux verbes possèdent par ailleurs chacun trois objets.

En fréquentiel :

– $a(écouter, convaincre) = 2 + 3 = 5$

– $c1(écouter) = 3 + 6 + 4 = 13$

– $c2(convaincre) = 4 + 3 + 2 = 9$

En effet, deux occurrences d'*interlocuteur* et trois de *collaborateur* sont à prendre en compte pour le calcul de "a". Par ailleurs, le nombre d'occurrences d'objets de "*écouter*" et "*convaincre*" sont respectivement de 13 et 9.

Il devient alors trivial de calculer les scores résultant des mesures de similarité. Par exemple avec le coefficient de Dice :

En binaire :

$Dice(écouter, convaincre) = \frac{2 \times 2}{3 + 3} \approx 0,67$

Nous avons effectivement seulement deux objets en communs entre ces deux verbes. Les deux verbes possèdent par ailleurs chacun trois objets.

En fréquentiel :

$Dice(écouter, convaincre) = \frac{2 \times 5}{13 + 9} \approx 0,45$

Nous proposons maintenant de comparer ces mesures de similarité avec celle d'Asium.

3.3.4.3 Comparaison des mesures avec la mesure d'Asium

L'objectif fixé dans cette section et de savoir si la mesure d'Asium a le même comportement que certaines des mesures décrites ci-dessus. Dans ce contexte, nous souhaitons savoir si des relations syntaxiques placées en tête de liste avec la mesure d'Asium sont également placées en tête avec les autres mesures de similarité. Par exemple, parmi les 10 premières relations syntaxiques classées avec la mesure d'Asium, combien de celles-ci se retrouvent classées dans les 10 premières avec une autre mesure de similarité ?

Afin de comparer la mesure d'Asium avec ces mesures, nous allons suivre le protocole expérimental suivant. Nous disposons d'un corpus de dépêches d'actualités provenant du site d'information de *Yahoo !* écrit en français. Nous avons extrait de ce corpus avec notre analyseur syntaxique Sygfran *47 097* relations syntaxiques de type Verbe-Objet. Les verbes composant ce corpus ont été comparés deux à deux avec les différentes mesures présentées précédemment, dont celle d'Asium. *130 992* couples ont ainsi été ordonnées via ces mesures.

Les différents calculs de similarité entre nos couples de verbes ont conduit à des résultats identiques pour une même mesure. Par exemple, avec le cosinus, plusieurs couples de verbes ont obtenu le même score de 1. Ces égalités sont problématiques car elles faussent la notion de rang, une fois les couples de verbes classés par ordre décroissant de proximité. Ainsi, pour ne pas biaiser les résultats présentés dans cette

section, nous considérerons les couples ayant obtenu un même score de similarité pour une même mesure comme de même rang. Avec aucun score identique, nous devrions avoir un classement de *130 992* couples, soit *130 992* rangs distincts. Avec par exemple l'information mutuelle, en comptant de même rang les scores identiques, nous avons obtenu seulement *17 058* rangs.

Afin d'évaluer le comportement d'une mesure par rapport à la mesure d'Asium, nous proposons de mesurer le recouvrement des n premiers couples de verbes classés par une mesure autre qu'Asium par rapport aux n premiers couples classés avec la mesure d'Asium. Par exemple avec un $n = 100$, les 100 premiers couples ordonnés par la mesure d'Asium, en incluant les couples ayant obtenu des scores de similarité identiques. Alors nous comptons le nombre de couples ayant également été classés parmi les 100 premiers couples triés par l'autre mesure. Finalement nous divisons ce score par le nombre de couples considérés avec la mesure d'Asium afin d'obtenir un *Pourcentage de Recouvrement des deux mesures. (PR).*

Comparaison avec la mesure d'Asium

Nous présentons ci-dessous les résultats obtenus en suivant ce protocole en comparant à la mesure d'Asium les six autres mesures de proximité sémantique présentées section 3.3.4.1, d'un point de vue binaire (figure 3.10) et fréquentiel (figure 3.11). Les résultats

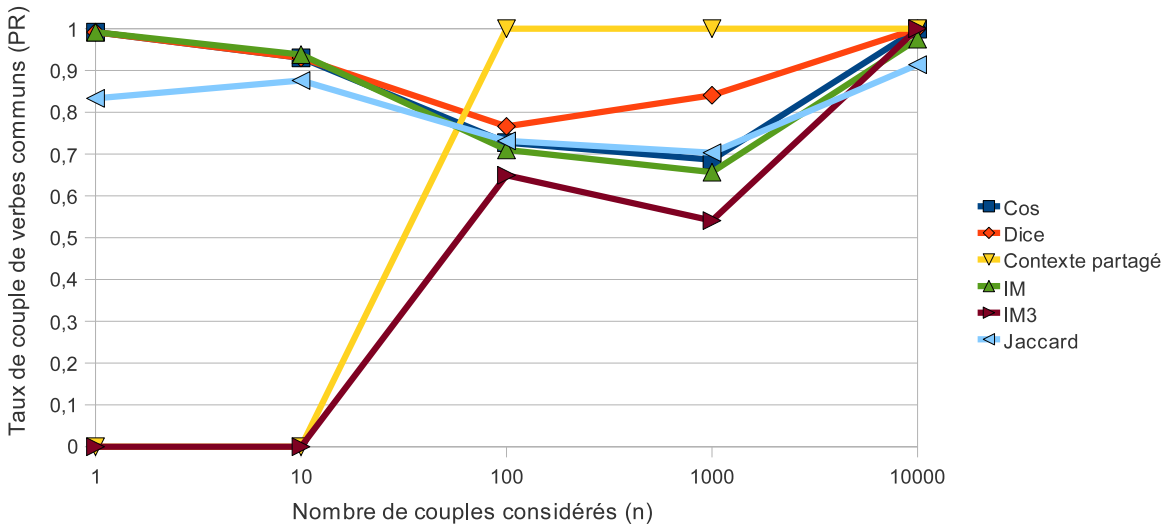

Fig. 3.10 – Comparaison des différentes mesures de proximité sémantique (version binaire)

obtenus montrent que pour le type binaire que les mesures *Cosinus*, *IM* et *Jaccard* ont un comportement assez identique par rapport à la mesure d'Asium. Notons également que le coefficient de *Dice* est assez similaire mais se rapproche d'avantage de la mesure

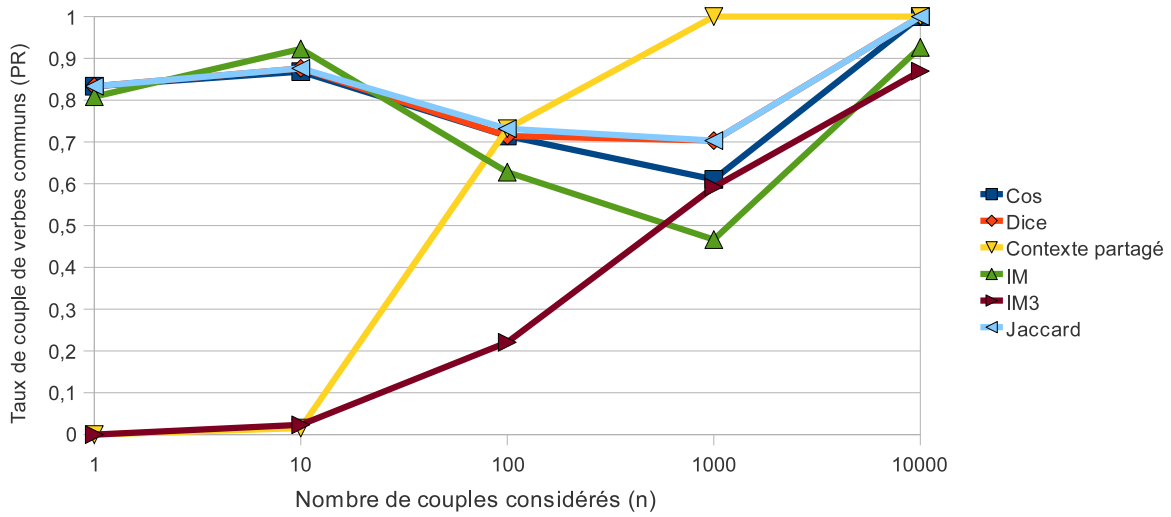

FIG. 3.11 – Comparaison des différentes mesures de proximité sémantique (version fréquentielle)

d'ASIUM que *Cosinus*, *IM* et *Jaccard* au delà des 1000 couples considérés. Par ailleurs, pour ces quatre mesures, en considérant les 100 premiers couples classés avec la mesure d'ASIUM, entre 70 et 80% des couples de verbes sont également présents dans les 100 premiers classés avec ces quatre mesures.

Pour les mesures fréquentielles, nous observons un comportement assez identique pour les quatre mesures *Cosinus*, *IM*, *Jaccard* et *Dice*. Cependant, seuls les résultats des mesures *Dice* et *Jaccard* sont vraiment similaires. Par rapport à la mesure d'ASIUM, les mêmes 70 et 80% des couples de verbes sont retrouvés pour les 100 premiers couples classés avec ASIUM comparativement aux mesures *Cosinus*, *Jaccard* et *Dice*. Néanmoins, la mesure *IM* se voit réduite à 62% de correspondances.

Les quatre mesures (d'un point de vue binaire et fréquentiel) ont un comportement assez proche de la mesure d'ASIUM du fait de leur nature similaire. En effet, la mesure d'ASIUM, comme les quatre mesures *Cosinus*, *IM*, *Jaccard* et *Dice* prend en considération deux facteurs : le nombre d'objets communs des verbes et le nombre d'objets propres à chaque verbe. Certes, les quatre mesures nuancent l'importance de l'un ou l'autre de ces facteurs comme l'information mutuelle, donnant plus d'importance au nombre d'objets propres à chaque verbe. Notons également que pour la mesure d'ASIUM, le nombre d'objets communs entre deux verbes n'est pas symétrique. En effet, dans cet exemple, $nb\_commun_{Asium}(écouter)$, noté $\sum NbOcc(p_i)$ dans la figure 3.8, se distingue de $nb\_commun_{Asium}(convaincre)$, noté $\sum NbOcc(q_i)$ dans la figure 3.8, ces deux calculs valant respectivement 7 et 5.

Les mesures *contexte partagé* et IM^3 ont un comportement différent des autres mesures de similarités, qu'elles soient fréquentielles ou binaires. Elles présentent en effet très peu de concordance avec la mesure d'Asium. La mesure *contexte partagé* se contente de se focaliser sur les objets communs des couples de verbes. La mesure IM^3 élève au cube cette même quantité (son numérateur), réduisant ainsi le poids donné aux objets propres à chaque verbe (son dénominateur). Ainsi, ces deux mesures sont assez proches. Ceci explique le manque de concordance avec la mesure d'Asium, celle-ci donnant plus d'importance aux objets propres à chaque verbe d'un couple.

Nous remarquons finalement que chaque résultat de concordance (c'est-à-dire pour chaque mesure) converge vers 1. En d'autres termes, les n premiers couples triés par la mesure d'Asium se voient tous, à partir d'un certain rang, classés dans les n premiers couples triés avec les autres mesures. Cela semble trivial car tous les couples de verbes présents dans le corpus se voient classés au final, avec un n égal au nombre total de couples de verbes distinctement classés. Notons cependant que ce nombre total de couples de verbes diffère suivant la mesure. Rappelons en effet qu'un nombre important de couples de verbes ont obtenu des scores de similarité identiques pour une même mesure. Cela explique également les résultats de la mesure *contexte partagé*. En effet, cette mesure se voit attribuer un score de 100% de concordance avec les 100 premiers couples (en binaire). Ainsi, les 100 premiers couples triés avec Asium sont retrouvés dans les 100 premiers couples de verbes en terme de rang triés avec la mesure *contexte partagé*. Ces résultats s'expliquent non pas parce que cette mesure se comporte comme la mesure d'Asium, mais parce qu'elle ne comporte que 45 rangs distincts. En effet, cette mesure considérée de manière binaire se voit très souvent attribuer des scores identiques. Une telle mesure a donc le défaut de ne pas être assez discriminante. Elle ne se focalise en effet que sur les objets communs des couples de verbes, quantité qui peut souvent être très faible, sans prendre en compte le nombre d'occurrences.

Le choix de la mesure d'Asium

Nous avons choisi d'utiliser la mesure d'Asium dans notre modèle de sélection de descripteurs SelDe, qui sera présenté dans la section suivante. Plusieurs motivations nous ont amené à faire ce choix.

– *Comportement similaire aux autres mesures de la littérature.*

Comme nous venons de le montrer dans le paragraphe précédent, la mesure d'Asium a un comportement assez proche des mesures usuelles telles que le *cosinus* ou l'*information mutuelle*. De plus la mesure d'Asium, comme celles précédemment citées évite le phénomène d'attraction se présentant dans le cadre d'utilisation de distances fondées sur celle de Hamming (que nous avons déjà évoqué en section 3.3.3).

– *La valeur ajoutée par les objets communs.*
La mesure d'Asium est cependant plus discriminante que les mesures dont elle s'inspire, en prenant en compte les objets communs de chaque verbe comme nous l'avons explicité dans la section précédente.
– *Adaptée à la classification conceptuelle*
Un dernier point nous ayant encouragé à utiliser cette mesure est son contexte original, le système Asium. La mesure est en effet adaptée à la construction de classes conceptuelles. Nous reviendrons sur ce point dans la section 7.1.3.
Après avoir décrit l'analyse syntaxique et la mesure de proximité que nous avons retenues, nous présentons dans la section suivante notre modèle de sélection de descripteurs SelDe.

3.4 Le modèle SELDE

Cette section présente notre approche d'extraction de descripteurs pertinents d'un document, le modèle Selde (**Sél**ection des **De**scripteurs). La figure 3.12, sur laquelle nous nous appuierons dans cette section, synthétise le modèle d'extraction des descripteurs.
Nous présentons dans un premier temps l'architecture générale de SelDe (section 3.4.1). Puis nous détaillerons un certain nombre de points pour la réalisation d'un tel modèle. Nous présenterons en premier lieu les différents post-traitements que nous avons appliqués aux relations syntaxiques extraites par l'analyseur Sygfran (section 3.4.2). Nous reviendrons ensuite sur une étape majeure du modèle SelDe, la sélection des objets pouvant servir de descripteurs (section 3.4.3). Alors, nous présenterons dans la section 3.4.4 quel rôle peuvent jouer les objets complémentaires dans le modèle SelDe pour finir sur la valeur ajoutée de nos descripteurs (section 3.4.5).

3.4.1 Les différentes étapes

La première étape (**étape 1** de la figure 3.12) consiste à extraire les relations syntaxiques Verbe-Objet d'un corpus. Nous utilisons pour cela l'analyseur morpho-syntaxique Sygfran précédemment détaillé. Ainsi, avec cette analyse syntaxique nous avons par exemple extrait de la phrase "L'accompagnement nécessite des professionnels." la relation syntaxique "verbe : nécessiter, COD : professionnels".

Une fois l'ensemble des relations syntaxiques Verbe-Objet du corpus extraites, nous mesurons la proximité sémantique des verbes en nous appuyant sur la mesure d'Asium (**étape 2** de la figure 3.12). Ainsi, nous mesurons la proximité sémantique entre les verbes en évaluant chaque verbe du corpus avec tous les autres. Nous obtenons alors une liste de couples de verbes avec pour chacun, un score de proximité associé. Notons

que nous excluons, lors du calcul de la proximité sémantique des verbes, les verbes trop fréquents comme "*avoir*" ou "*faire*". En effet, de tels verbes peuvent apporter du bruit car ils peuvent être reliés syntaxiquement à un nombre considérable d'objets ayant en fait peu de points communs sémantiques.

L'étape suivante (**étape 3** de la figure 3.12) de notre modèle consiste dans un premier temps à ordonner les couples de verbes par proximité sémantique. Alors, nous ne conservons que les couples de verbes dont le score de similarité est supérieur à un certain seuil, appelé seuil d'ASIUM (noté SA, sur lequel nous reviendrons dans la section 3.4.3.3). De plus, si un verbe apparaît dans plusieurs couples, nous ne conservons que le couple de verbes ayant obtenu un score d'ASIUM maximum.
Par exemple, considérons le couple de verbes précédemment cité : *requérir* et *nécessiter* (1). Le verbe *nécessiter* a pu être jugé également proche du verbe *réclamer* pour le couple *réclamer* et *nécessiter* (2). Si le score de similarité obtenu avec la mesure d'ASIUM est plus faible pour le couple (2), seul le couple (1) est retenu pour le verbe.

La dernière étape du modèle SELDE est le regroupement d'objets communs des verbes, ces objets devenant ainsi les descripteurs extraits (**étape 4** de la figure 3.12). En effet, nous regroupons pour chaque couple de verbes distinct tous les objets de ces deux verbes. Ces objets sont finalement sélectionnés pour définir nos descripteurs pertinents.
Après avoir décrit les différentes étapes du modèle SELDE, nous détaillons dans la section suivante différents traitements qui ont été appliqués après avoir extrait les relations syntaxiques d'un corpus (*étape 1* de la figure 3.12).

3.4.2 Les post-traitements apportés à l'analyse de Sygfran

Bien que la précision de l'analyseur syntaxique SYGFRAN soit de bonne qualité, nous nous devons de réduire encore le bruit dans les résultats produits. Ainsi, dans le cadre des travaux menés au cours de cette thèse, nous appliquons un post-traitement aux relations syntaxiques résultantes de l'analyse en appliquant un filtrage supplémentaire. Ces traitements peuvent être situés dans la figure 3.12 entre les étapes 1 et 2.

Les noms propres.
L'analyse syntaxique de SYGFRAN, bien que robuste et fiable, ne peut parfois lever les ambiguïtés syntaxiques de certaines phrases tel qu'explicité dans la section 3.2.3.4 (rôle du sous-système TELESI de SYGMART). De telles ambiguïtés peuvent survenir sur l'attribution de la fonction syntaxique d'un groupe nominal, pouvant contenir des noms propres. Par exemple, avec la phrase "*Sors, ordonne Xavier.*", l'analyseur propose pour le nom

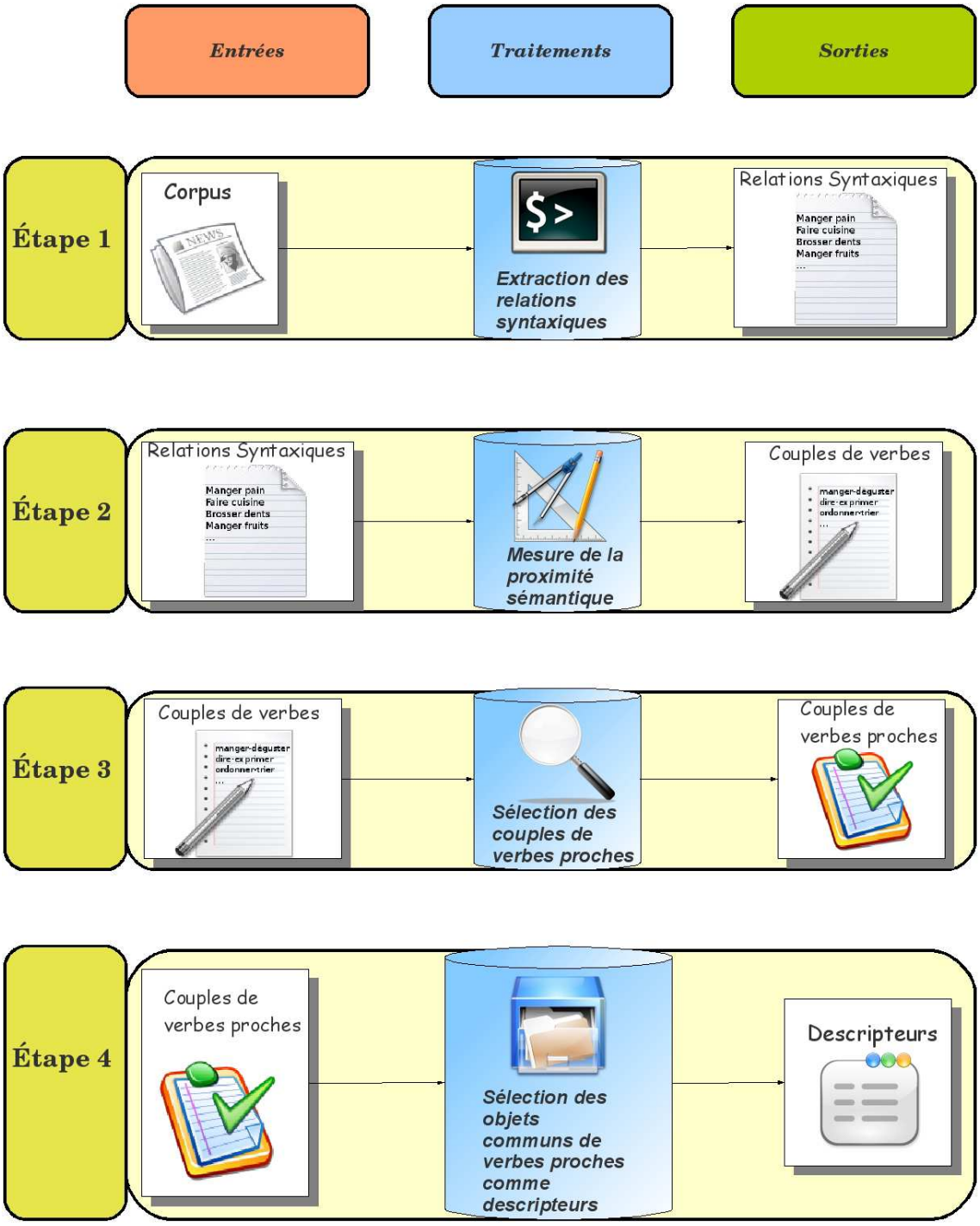

FIG. 3.12 – Modèle d'extraction des descripteurs

propre "*Xavier*" la fonction de Sujet ou d'Objet tel que montré sur le figure 3.13. Dans le

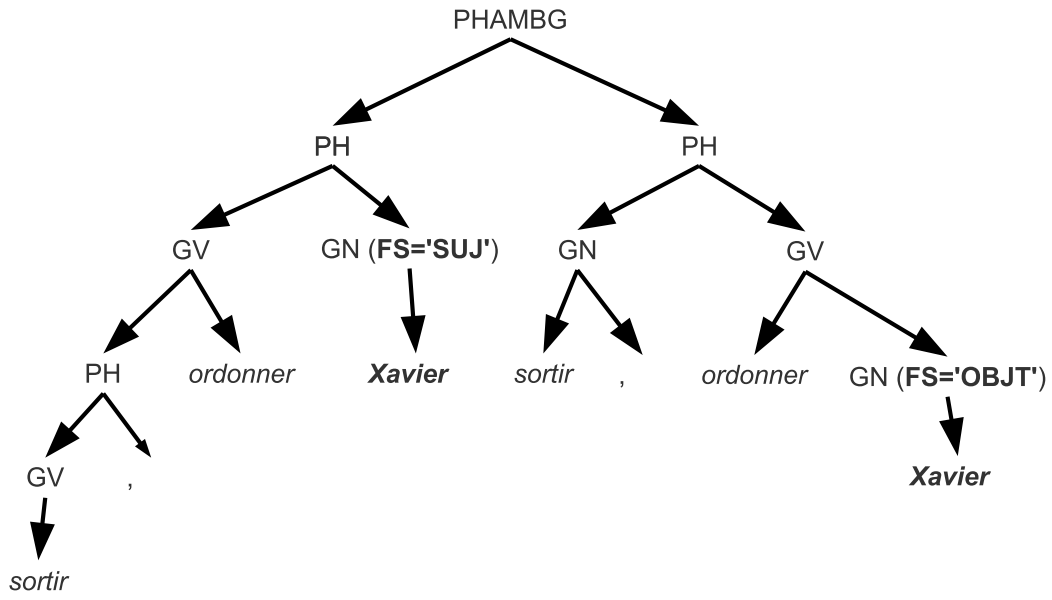

FIG. 3.13 – Analyse simplifiée de la phrase "Sors, ordonne Xavier."

premier cas, ce nom propre est sujet de la phrase "*Sors, ordonne Xavier.*" et dans le second cas, ce nom est objet du verbe "*ordonner*". SYGFRAN nous impose alors de faire un choix entre les deux propositions d'analyse. Ainsi, pour limiter le bruit lié à ces ambiguïtés, nous avons supprimé les relations syntaxiques extraites contenant des noms propres, favorisant alors la précision des relations syntaxiques. Néanmoins, l'ambiguïté relevée ici peut également survenir lors de l'analyse de la phrase "*Sors, ordonne le directeur.*". Notons finalement que ce type d'ambiguïté a été levé dans la dernière version de l'analyseur syntaxique en évolution constante.

Les groupes nominaux.

L'utilisation des groupes nominaux dans notre approche en tant qu'objet est certes séduisante car garantissant une meilleure qualité sémantique des objets, mais n'est pas envisageable. En effet, l'étude de la proximité des verbes présentés dans la section suivante implique le partage d'un nombre important d'objets d'un couple de verbe. Or, en considérant une relation syntaxique comme formée d'un verbe et d'un groupe nominal, le nombre de couples résultant se voit diminué. Avec l'utilisation d'un corpus de grande taille, le nombre de relation pourrait encore être suffisant afin d'acquérir suffisamment de connaissances mais cette approche n'est pas applicable pour des corpus moins conséquents. Ainsi, nous ne privilégierons pas cette approche, afin de pouvoir appliquer SELDE avec toutes tailles de corpus. Nous ne conservons que les *têtes* des groupes nominaux objets

(ou *gouverneurs*) au sens de SYGFRAN avec le verbe afin de construire les relations syntaxiques. Par exemple dans le groupe nominal "*le ministre de l'agriculture*", nous ne conservons que le terme "*ministre*", gouverneur de ce groupe.

Les verbes.
Un dernier post-traitement consiste en la suppression des verbes trop fréquents comme *avoir*, *faire*, *dire*, etc. par le biais d'une liste manuellement définie. En effet, ces verbes peuvent apporter du bruit lors de la sélection des couples de verbes sémantiquement proches. Le verbe "*avoir*" par exemple va être jugé proche de *posséder*, *détenir*, *éprouver*, etc.

Nous présentons dans la section suivante l'*étape 4* de SELDE, décrivant de manière plus précise la sélection des objets de verbes.

3.4.3 La sélection des objets en tant que descripteurs

3.4.3.1 Le choix du type d'objet

Une fois les verbes d'un corpus jugés sémantiquement proches avec la mesure d'ASIUM et les objets de ces verbes regroupés, nous devons **sélectionner certains de ces objets constituant nos descripteurs** (*étape 5* de la figure 3.12). Dans un premier temps, nous nous intéressons à la provenance des objets. Sont-ils objets des deux verbes du couple ou d'un seul ?
Ainsi, nous distinguons deux types d'objets à ces verbes :
– Les objets communs
– Les objets complémentaires
Considérons deux verbes sémantiquement proches V_1 et V_2.
Soit $Obj_1^{V_1}...Obj_n^{V_1}$ et $Obj_1^{V_2}...Obj_m^{V_2}$ les objets des verbes V_1 et V_2.
$Obj_i^{V_1}$ $(i \in [0, n])$ est appelé un objet **commun** si $\exists j \in [1, m]$ tel que $Obj_i^{V_1} = Obj_j^{V_2}$.
Si $Obj_k^{V_1}$ (et respectivement $Obj_k^{V_2}$) n'est pas un objet commun, alors la relation syntaxique V_2-$Obj_k^{V_1}$ (et respectivement V_1-$Obj_k^{V_2}$) est appelée **relation syntaxique induite** et l'objet $Obj_k^{V_1}$ est appelé objet **complémentaire**.

Par exemple, dans la figure 3.14 où les verbes *consommer* et *manger* sont jugés proches, nous identifions les objets de ces deux verbes : *essence*, *légume*, *nourriture* et *fruit*. Les objets *légume* et *nourriture* sont des objets **communs** aux deux verbes *consommer* et *manger*. L'objet *fruit* qui est objet du verbe *manger* est un objet **complémentaire** du verbe *consommer*, tout comme l'objet *essence* pour le verbe *manger*. Les relations syntaxiques induites sont donc sur l'exemple de la figure 3.14

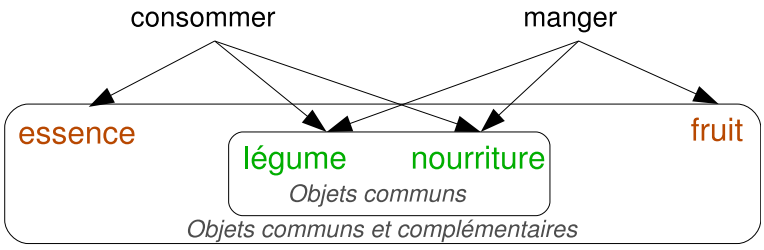

FIG. 3.14 – Objets communs et complémentaires des verbes “consommer” et “manger”.

les relations *manger essence* et *consommer fruit*. Notons que ces relations syntaxiques induites sont des connaissances nouvelles “apprises” à partir des corpus car elles ne sont pas explicitement présentes dans les données textuelles.

Nous utilisons dans le modèle SELDE uniquement les objets communs des verbes. Mais ces objets complémentaires ont un intérêt essentiel qui sera développé dans le chapitre 6 traitant du modèle SELDEF (**Sél**ection de **De**scripteurs avec **F**iltrage). La section 3.4.4 discutera à la fin de ce chapitre de ces objets complémentaires et montrera qu'une simple sélection statistique de ceux-ci ne peut être pertinente.

3.4.3.2 Le Seuil d'Asium – *SA*

Afin de considérer deux couples de verbes comme proches avec la mesure d'ASIUM, nous devons fixer un seuil. Celui-ci, que nous noterons dans la suite de ce mémoire ***SA*** pour **S**euil d'**A**sium, doit en effet déterminer quelle valeur obtenue avec le score d'ASIUM doit être atteinte afin de considérer deux verbes comme proches. Rappelons que les scores obtenus avec la mesure d'ASIUM appartiennent à l'intervalle $[0, 1]$.
Ainsi, un seuil SA fixé à $0,9$ signifie un nombre réduit de descripteurs sélectionnés car peu de verbes vont obtenir un score supérieur à $0,9$. Mais ceci implique un regroupement de descripteurs (objets) de meilleure qualité car le score est assez proche de 1, score signifiant que les deux verbes partagent exactement les mêmes objets. À l'inverse, un score de $0,6$ signifie un nombre important de descripteurs mais de moins bonne qualité. Notons fixer des scores SA à 0,6 ou bien à 0,9 revient à favoriser respectivement le rappel ou la précision avec notre approche.

Par exemple, avec un corpus d'une taille moyenne d'environ 600 000 mots :
– 3132 couples de verbes ont obtenu un score d'ASIUM supérieur ou égal à $0,6$
– 255 couples de verbes ont obtenus un score d'ASIUM supérieur ou égal à $0,9$

Un exemple de couples de verbes ayant respectivement obtenu un score supérieur à 0,9 et un score supérieur à 0,6 (mais inférieur à 0,7) est donné ci-dessous. Les objets communs de ses couples, définissant nos descripteurs avec SelDe, sont également reportés.
Couple de verbes = *analyser-traiter*, objets communs : "*cas(6 occurrences), problème(18), situation(3), sujet(2), question(3)*"
Couple de verbes = *sentir-manifester*, objets communs : "*présence(2), autonomie(2), inquiétude(2), intérêt(2), attention(5), résistance(2)*"

Nous remarquons avec les exemples ci-dessus que le nombre de couples de verbes est en effet bien supérieur pour un score d'Asium au delà de $0,6$ par rapport à $0,9$. Notons par ailleurs que la proximité sémantique des verbes obtenus avec le couple *analyser-traiter* est avérée. Cette même proximité est plus discutable pour le couple *sentir-manifester*. En effet, rappelons que plus le score d'Asium est proche de 1 et plus la proximité établie entre deux verbes sera avérée.

Rappelons pour finir qu'avec l'utilisation de SelDe, seuls les objets communs vont être utilisés comme descripteurs et non les objets complémentaires. L'utilisation de tels objets sera abordée avec le modèle de sélection de descripteurs filtrés SelDeF (section 6.2).

3.4.3.3 Les différents paramètres pour la sélection de descripteurs

Nous présentons dans cette section les paramètres que nous avons définis afin d'améliorer la sélection effectuée dans l'approche d'Asium originale.

Le nombre d'occurrences d'un objet : *NbOccMin* et *NbOccMax*.
Dans un corpus donné, le nombre d'occurrences des mots est dépendant de chaque mot. En d'autres termes, tous les mots ne reviennent pas à la même fréquence. Une loi empirique énoncée en 1936 par le sociologue américain G. Zipf (publié dans [Zipf, 1941]) indique que si nous classons les mots d'un texte donné par ordre décroissant de leur nombre d'occurrences, la fréquence du $k^{\text{ième}}$ mot est approximativement proportionnelle à $1/k$. Ainsi, si le mot le plus fréquent d'un corpus a pour nombre d'occurrences occ_{max}, alors le moins fréquent du même corpus se voit classer au rang occ_{max}, si son nombre d'occurrences est minimal.

Nous proposons dans ce paragraphe de nous concentrer sur le nombre d'occurrences d'un objet pour une relation syntaxique donnée. Par exemple, combien de fois l'objet "*viande*" a-t'il été objet du verbe "*manger*" ? En effet, il existe dans un texte un nombre très important de termes présents une seule fois dans un corpus et à l'inverse un

nombre important d'occurrences de termes peu fréquents, en suivant plus ou moins la loi de Zipf. Deux cas doivent ainsi être considérés pour les relations syntaxiques : un nombre d'occurrences *trop faible* et *trop élevé*.
Le premier cas propose de ne pas sélectionner les objets rares qui, suivant la tâche pour laquelle nous utiliserons les descripteurs, pourront s'avérer pertinents. Par exemple pour une tâche d'enrichissement de corpus, un terme trop isolé n'apportera pas d'information supplémentaire. Nous notons ce paramètre ***nbOccMin***.
Le second aura la conséquence inverse à savoir ne pas sélectionner les objets trop fréquents du corpus pour un couple de verbe donné. Par exemple, afin d'effectuer une tâche d'indexation de documents, utiliser toujours le même mot-clé afin de décrire un document ne se révèle pas très judicieux. Nous notons ce second paramètre ***nbOccMax***.
Il est donc nécessaire de trouver un compromis entre ces deux paramètres. Ainsi, pour résumer l'utilisation de ces deux paramètres, nous sélectionnerons les descripteurs en fonction de leur nombre d'occurrences, tel qu'appartenant à l'intervalle indiqué ci-dessous :

$$nbOccMin \geq \text{Descripteurs sélectionnés} \geq nbOccMax$$

Limiter le nombre d'objets résultant de chaque couple : *nbObj*.
Un autre paramètre présenté dans cette section permet de limiter le nombre d'objets résultant de chaque couple de verbes jugés sémantiquement proches. Un tel paramètre va permettre par exemple de ne pas avoir un nombre trop important d'objets par couple. Il n'est en effet pas rare d'obtenir plus d'une cinquantaine d'objets distincts pour un couple de verbes. Ainsi ce paramètre fixe un nombre maximal d'objets possibles par couple. Par exemple si *nbObj* vaut 5, seulement cinq objets par couple de verbes seront conservés. Cependant, avec ce seul paramètre, nous ne pouvons pas extraire d'objets précis. Prenons par exemple le couple de verbes précédemment cité :
analyser-traiter, objets communs : "*cas(6 occurrences), problème(18), situation(3), sujet(2), question(3)*"
En fixant *nbObj* à 2, quels objets allons nous conserver ? Afin de remédier à ce problème nous introduisons le nombre d'occurrences à cette sélection. Nous proposons en effet de ne conserver que les objets plus fréquents ou les moins fréquents en termes de nombre d'occurrences. Un terme fréquent va en effet être plus représentatif d'un corpus, cependant, les objets moins fréquents sont porteurs d'informations nouvelles. Ce choix sera déterminé par le paramètre *Order*. Celui-ci peut alors prendre les valeurs "c" dans le cas où nous sélectionnons les objets en ordre croissant en terme de nombre d'occurrences ou bien "d" dans le cas contraire. Avec notre exemple précédent, nous extrairions avec $Order = c$ les termes "sujet, question" ou "sujet, situation" et avec $Order = d$ les termes "problème,

cas". Notons qu'en cas d'égalité, comme c'est le cas dans cet exemple, nous sélectionnons aléatoirement le terme approprié (ici "question" ou "situation").
Une fois tous les paramètres permettant de sélectionner les descripteurs, nous revenons dans la section suivante sur notre choix, dans un premier temps, de ne pas intégrer les objets complémentaires à SelDe.

3.4.4 Les objets complémentaires dans le modèle SelDe

Les objets complémentaires qui forment avec un verbe des relations syntaxiques induites constituent des connaissances nouvelles, non initialement présentes dans un corpus. En effet, aucun analyseur syntaxique n'est en mesure d'extraire ces relations à partir d'un corpus. Cependant, ces relations sont dans de très nombreux cas porteuses de bruit. Par exemple citons la relation sémantiquement peu probable "manger essence" de la figure 3.14. Il parait alors évident que de telles relations doivent être filtrées, toutes n'étant pas porteuses d'informations de qualité.

Le modèle SelDe propose de filtrer les relations syntaxiques "classiques" en utilisant des paramètres prenant en compte le nombre d'occurrences des objets pour chaque verbe d'un couple (NbOccMax/NbOccMin), mais également le nombre maximum résultant de descripteurs par couple (nbObj). Ces mesures sont efficaces avec des relations syntaxiques classiques car elles en limitent le nombre. Ainsi, des relations peu fréquentes peuvent être éliminées (relations pouvant être apparentées à du bruit dans le corpus) et les trop fréquentes ne peuvent également pas être retenues du fait de leur nature trop homogène dans le corpus, n'apportant pas d'informations utiles. Mais ces paramètres supposent également le fait que ces relations syntaxiques soient existantes dans le corpus, et donc *a fortiori* plausibles.

Qu'en est-il des relations induites ? Elles n'existent pas à l'origine dans le corpus et ne peuvent donc pas être considérées comme plausibles. Les paramètres précédemment évoqués ne peuvent donc s'appliquer. Prenons par exemple le couple de verbes "sortir-traiter" ayant obtenu un score approximatif de $0,6$ avec la mesure d'Asium. Nous listons ci-dessous les objets résultant de ce couples :
Objets communs : cas(6 occurrences), affaire(4), sujet(2)
Objets complémentaires : vision(1 occurrence), embarras(4), norme(3), commun(13), frontière(1), impasse(4), fléchissement(1), ordinaire(11), dilemme(2), attribution(1), ennui(2), environnement(1), interlocuteur(1), suggestion(1), accident(2), événement(1), ordre(1), divergence(1), demande(1), difficulté(3), problème(17), chiffre(1), dossier(4), aléa(1), imprévu(3), crise(1), situation(2), question(2)

Cet exemple montre tout d'abord qu'un couple de verbe génère beaucoup plus d'objets complémentaires que d'objets communs. Mais il montre également que les objets complémentaires sont de moins bonne qualité. En effet, les termes "*ordinaire*", "*dossier*" et "*accident*" ne semblent pas avoir de cohésion sémantique particulière. Finalement nous montrons par cet exemple que les simples paramètres de SelDe sont insuffisants pour sélectionner des objets complémentaires ayant un sens commun.

La méthode d'Asium [Faure, 2000] suppose une sélection manuelle de ces objets complémentaires, permettant ainsi de les valider. Notre objectif est de produire un modèle de sélection de descripteurs fonctionnant de manière autonome. Ainsi, nous n'avons pas intégré ces relations induites à notre modèle SelDe, mais ces relations sont considérées dans le modèle SelDeF. Ceci est l'objet du chapitre 6. Nous proposons dans ce second modèle de traiter les objets complémentaires (et donc les relations induites) avec des méthodes de filtrage plus élaborées que celles de SelDe, utilisant notamment des ressources Web et linguistiques.

Nous conclurons ce chapitre par le paragraphe suivant en motivant l'utilisation de descripteurs hybrides proposées par SelDe. En effet, ces descripteurs sont extraits en utilisant des informations statistiques et linguistiques.

3.4.5 Les apports des descripteurs hybrides

Les descripteurs sélectionnés par notre approche SelDe sont fondés sur l'utilisation d'informations morpho-syntaxiques fournies par l'analyseur Sygfran et sur une mesure statistique *Asium*. Ainsi, les descripteurs sélectionnés avec cette approche sont obtenus en combinant la qualité des descripteurs sélectionnés avec des approches purement linguistiques et statistiques. En effet, pour la partie linguistique de notre approche, nous sélectionnons des termes souvent liés à des verbes par des relations syntaxiques. Ce lien s'apparente à une association en collocations au sens de [Mel'čuk, 1998] et [Hausmann, 1989] à savoir : "*co-occurrences privilégiées de deux constituants linguistiques entretenant une relation sémantique et syntaxique*". Les termes que nous sélectionnons sont des objets. Ainsi, ces collocations forment des relations syntaxiques de type Verbe-Objet. Ceci fournit une information plus riche et plus précise qu'une simple sélection de termes fondée sur des méthodes statistiques comme par exemple selon la fréquence des mots dans le corpus ou dans chaque document.

Alors, les relations syntaxiques sémantiquement proches sont regroupées deux à deux, cette tâche est effectuée par une mesure statistique, obtenant des résultats

assez proches de mesures reconnues comme étant de qualité dans la littérature comme l'information mutuelle ou bien le coefficient de Jaccard (tel que nous l'avons montré dans la section 3.3.4). Ainsi, les descripteurs sélectionnés constituent des groupes de mots, les objets des verbes, choisis en fonction de leurs caractéristiques sémantiques communes.

La qualité de la sélection de descripteurs avec le modèle SelDe a été mesurée dans le cadre d'expérimentations présentées dans le chapitre suivant. Nous nous sommes plus particulièrement intéressés aux tâches de classifications conceptuelles et textuelles, en proposant une approche permettant d'enrichir un corpus en utilisant SelDe.

Chapitre 4

Application du modèle SelDe pour l'enrichissement de contextes

Sommaire

Le chapitre précédent a présenté notre approche de sélection de descripteurs SelDe. Ce modèle permet d'extraire des termes pertinents d'un corpus afin de décrire celui-ci. Nous proposons dans ce chapitre une application de ce modèle en présentant et en évaluant une méthode d'enrichissement de corpus. Les descripteurs extraits par SelDe vont être employés afin d'effectuer une expansion de termes dans un corpus. Cette méthode est présentée dans la section 4.1. Cette approche, outre le fait de démontrer la qualité des descripteurs extraits, va permettre la construction d'un corpus plus riche sémantiquement, pouvant par exemple être utilisé pour des tâches de classification automatique. Ainsi, nous évaluerons cette méthode avec deux types d'applications : la classification conceptuelle et la classification de textes en utilisant des approches avec apprentissage supervisé. Ces expérimentations sont respectivement présentées dans les sections 4.2 et 4.3.

4.1 Un modèle d'expansion de corpus appliqué à la classification

4.1.1 Description du modèle d'expansion

Notre approche vise à enrichir un corpus initial lemmatisé en faisant une expansion des phrases fondée sur la méthode SelDe présentée dans le chapitre 3. Une telle approche va permettre de produire un contexte plus riche. Celui-ci est construit en complétant les mots d'un corpus par les descripteurs extraits avec SelDe. Un exemple d'enrichissement est donné ci-dessous :
Soit la phrase : *"Vos interlocuteurs seront donc bien inspirés de placer les échanges ..."*.
Nous la transformons tout d'abord en phrase lemmatisée via le système Sygmart : *"Votre interlocuteur être donc bien inspiré de placer le échange ..."*.
Enfin, elle va être enrichie avec nos descripteurs en devenant la phrase (en s'appuyant sur la figure 3.14 du chapitre 3) *"Votre (interlocuteur collaborateur) être donc bien inspiré de placer le échange ..."*.
L'expansion d'un corpus s'effectue en différentes étapes résumées dans la figure 4.1.

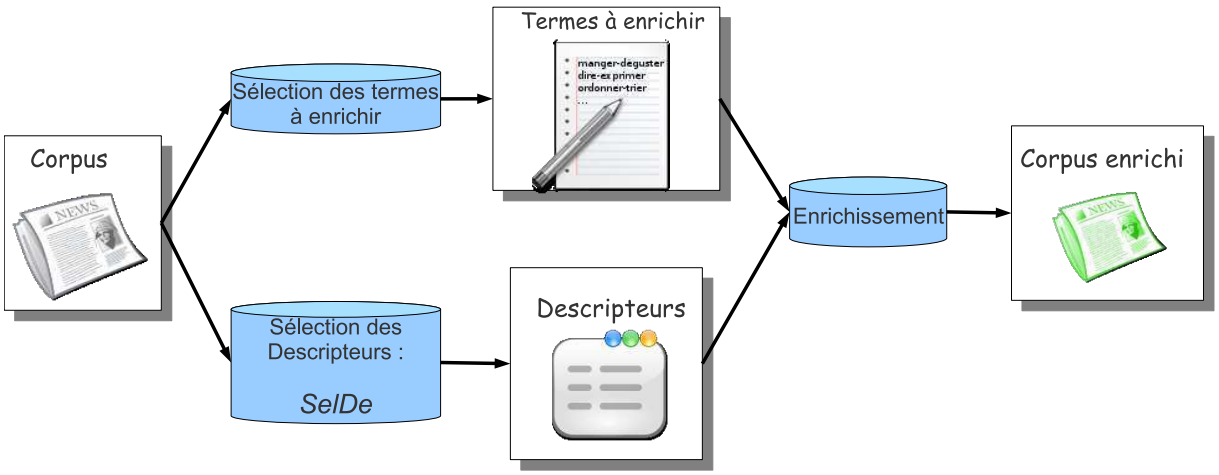

Fig. 4.1 – Modèle d'expansion de corpus

Sélection des termes à enrichir et des descripteurs
La sélection des descripteurs s'effectue selon le modèle décrit dans le chapitre 3. Ces descripteurs vont alors être utilisés afin d'enrichir les termes sélectionnés pour l'enrichissement. Ces termes sont tous les noms du corpus. En effet, tous les noms du corpus peuvent potentiellement être enrichis par cette approche. Les termes à enrichir dépendent directement des descripteurs extraits avec SelDe, qui ne sont autres que les objets communs

des couples de verbes jugés sémantiquement proches. Cette nuance est explicitée dans le paragraphe suivant, traitant de l'enrichissement.

L'enrichissement

L'étape d'enrichissement consiste de manière triviale à enrichir chaque terme candidat à l'expansion par les descripteurs extraits avec SelDe. Rappelons que lors de la sélection de descripteurs avec SelDe, nous mesurons la proximité de verbes en fonction de leurs objets. Ainsi, la mesure d'Asium nous donne un score. Nous fixons alors un seuil, noté SA pour "Seuil d'Asium", au delà duquel les verbes sont considérés comme proches. Nous cherchons alors à enrichir un terme par d'autres termes qui partagent son contexte (les descripteurs de SelDe formés par les objets des couples de verbes proches). Nous sélectionnons donc les couples dans lesquels le terme à enrichir apparait. Une question se pose néanmoins : de quel couple de verbes sémantiquement proches allons nous sélectionner les objets communs afin d'enrichir notre candidat ? Il se peut en effet que le terme à enrichir apparaisse dans plusieurs couples.

Ainsi, nous proposons deux types de sélections possibles pour le couple de verbes dans lequel le terme apparaît :

– *Sélectionner le couple de verbes ayant obtenu le score d'*Asium *le plus élevé.* Cette sélection se distingue du seuil d'Asium SA. Le choix effectué ici est la sélection des objets du couple de verbes ayant obtenu le score d'Asium le plus élevé appartenant à $[SA, 1]$. Par exemple, soit un seuil $SA = 0,8$, un terme à enrichir t_e, un ensemble de couples de termes C_t dans lesquelles apparaissent t_e. Si le score d'Asium le plus important obtenu parmi C_t est 0,87, les objets du couple ayant obtenu ce score seront sélectionnés pour l'enrichissement de t_e. Cependant, si SA valait 0,9, aucun couple ne serait sélectionné pour un enrichissement de t_e.

– *Sélectionner le couple de verbes dans lequel le nombre d'occurrences du terme à enrichir est le plus élevé.*

Chaque couple de verbes contient un certain nombre d'objets communs. Ces derniers ont également un nombre d'occurrences propre. Avec cette sélection, nous nous focalisons non plus sur le plus important score obtenu avec Asium. Nous mettons ici en valeur le terme ayant le nombre d'occurrences le plus élevé. Ainsi, pour chaque couple de verbes où le terme à enrichir apparait, nous sélectionnons celui où le terme à enrichir possède un maximum d'occurrences.

Nous noterons ce paramètre de choix du couple de verbes *ChVerb* dans la suite de ce mémoire, en lui attribuant comme valeur "Asium" (seuil d'Asium maximum) pour la première méthode et "Occurrences" (le nombre maximum d'occurrences) pour la seconde. Nous illustrons ce principe d'enrichissement avec l'exemple ci-dessous.

Exemple d'enrichissement
Soit la phrase "*Quelles sont les compétences d'un dirigeant d'entreprise ?*". Afin de simplifier cet exemple, seul le nom "compétence" de notre phrase va être utilisé comme terme candidat à l'expansion. Ainsi, notre objectif est d'enrichir le nom "compétence". Après avoir extrait les descripteurs en suivant le modèle SelDe, nous disposons d'un certain nombre de couples de verbes dans lesquels le nom "compétence" apparait (nous n'en reporterons que trois) :

couple 1 : "stimuler-mettre", *score d'*Asium = 0.67, *Objets communs* = capacité(5 occurrences), possibilité(2), imagination(2), **compétence**(12)

couple 2 : "utiliser-disposer", *score d'*Asium = 0.85, *Objets communs* = énergie(5 occurrences), ressource(6), facilité(3), atout(18), **compétence**(3)

couple 3 : "douter-dépasser", *score d'*Asium = 0.79, *Objets communs* = capacité(6 occurrences), **compétence**(6)

Alors, avec $ChVerb = Asium$, nous sélectionnerions le *couple 2* (score d'Asium à 0,85) et avec $ChVerb = Occurrences$, nous sélectionnerions le *couple 1* (12 occurrences de "compétence").

Finalement, la phrase initiale :
- Quelles sont les compétences d'un dirigeant d'entreprise ?

devient après lemmatisation :
- Quel être le compétence de un dirigeant de entreprise ?

puis, après enrichissement avec la première méthode de sélection (seuil d'Asium maximum) devient :
- Quel être le (***compétence*** **, énergie , ressource , facilité , atout**) de un dirigeant de entreprise ?

et finalement enrichie avec la seconde méthode (le nombre maximum d'occurrences), la phrase devient :
- Quel être le (***compétence*** **, capacité , possibilité , imagination**) de un dirigeant de entreprise ?

4.1.2 Corpus enrichi et classification

La tâche de classification consiste à regrouper des contextes dans différentes classes qui correspondent à des catégories thématiques (par exemple, les thèmes "politique",

"sport", "technologies", etc). Nous distinguons deux types de classification : (1) la classification conceptuelle et (2) la classification de textes.
L'objectif de (1) est de regrouper des termes dans des classes. Le contexte est ici le ou les documents pouvant être de granularités diverses comme une phrase, un paragraphe, un texte, etc., dans lesquels ce terme apparait.
A contrario, (2) consiste à classer des documents dans des classes. Le contexte est dans ce cas formé des termes étant contenus dans le document à classifier.

Un contexte contient un certain nombre de descripteurs qui peuvent être insuffisants quantitativement pour la réalisation d'une classification automatique. Prenons par exemple les phrases suivantes, en considérant une tâche de classification de documents textuels, le contexte étant la phrase.

- P1 : *Le député s'adresse aux consuls*
- P2 : *Le sénateur s'exprime devant les ambassadeurs*

Nous constatons que les phrases P1 et P2 n'ont aucun mot en commun ce qui classerait ces phrases dans deux catégories différentes en nous appuyant sur des méthodes statistiques. Après expansion avec la méthode qui sera décrite dans la section 4.1, il est possible d'enrichir ces phrases de la manière suivante :

- E1 : *Le (député sénateur parlementaire) s'adresse aux (consuls ambassadeurs diplomates)*
- E2 : *Le (sénateur député parlementaire) s'exprime devant les (ambassadeurs consuls diplomates)*

Dans ce cas, les phrases E1 et E2 possèdent six mots communs signifiant une proximité thématique. Nous montrons par cet exemple que deux phrases proches sémantiquement peuvent s'avérer difficiles à classer sans utiliser de connaissances sémantiques. Avec notre enrichissement, l'information apportée peut remédier à cette contrainte. Rappelons que nous plaçons nos travaux dans le cadre du traitement de textes qui peut se révéler plus ou moins spécialisé sans utiliser de connaissances sémantiques tels que des dictionnaires du domaine. Après avoir motivé notre démarche visant à enrichir un corpus, nous présentons ci-dessous un modèle d'enrichissement de corpus, adapté à une tâche de classification.

4.1.3 Un modèle d'enrichissement de corpus pour une tâche de classification

Une fois le modèle d'expansion d'un corpus décrit, nous proposons un modèle adapté à la classification de données textuelles. La figure 4.2 présente l'architecture globale de notre approche.

Après avoir enrichi le corpus original, nous représentons vectoriellement le corpus avec

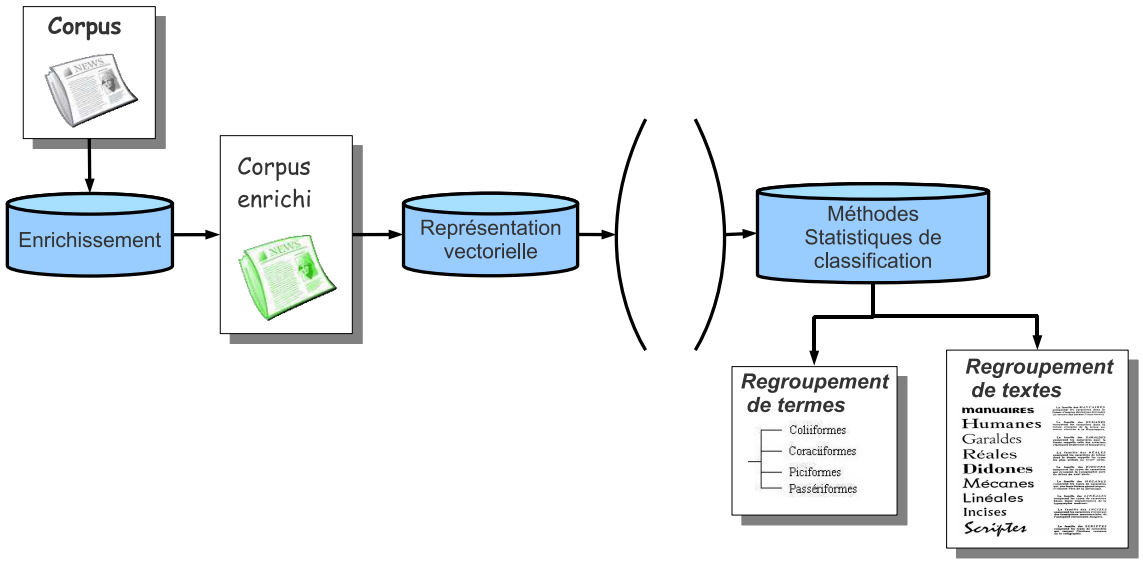

FIG. 4.2 – Modèle d'expansion de corpus

une méthode de type *Salton* ou *LSA* (tel que décrit dans la section 2.2.4). Finalement, nous regroupons les termes du corpus enrichi, première étape d'une classification conceptuelle ou bien nous regroupons les textes afin d'établir une classification de textes. Nous présentons dans les sections suivantes les différentes expérimentations que nous avons menées avec cette méthode d'enrichissement dans un contexte de classification.

Nous utilisons uniquement pour cette tâche la vectorialisation de type *LSA*. Notre approche d'**Exp**ansion de corpus avec **LSA** sera maintenant nommée **ExpLSA**. Notons que l'enrichissement proposé par *ExpLSA* a deux objectifs : (1) proposer un corpus plus riche contenant de nouvelles informations, (2) combler les lacunes de LSA. Ainsi, nous proposons dans la section suivante, en préambule de nos expérimentations, un état de l'art sur les différentes approches de la littérature proposant d'améliorer la méthode LSA.

4.1.4 LSA et la syntaxe

L'approche LSA présentée dans la section 2.2.4 présente des avantages parmi lesquelles la notion d'indépendance par rapport à la langue du corpus étudié, le fait de se dispenser de connaissances linguistiques ainsi que de celles du domaine tels que des thésaurus.
Bien que cette approche soit prometteuse, il n'en demeure pas moins que son utilisation soulève des contraintes. Notons tout d'abord l'importance de la taille des contextes choisis. [Rehder *et al.*, 1998] ont montré lors de leurs expérimentations que si les contextes possèdent moins de 60 mots, les résultats s'avèrent être décevants. Par ailleurs,

[Landauer *et al.*, 1997] posent le problème du manque d'informations syntaxiques dans LSA en comparant cette méthode à une évaluation humaine. Il est question de proposer à des experts humains d'attribuer des notes à des essais sur le cœur humain de 250 mots rédigés par des étudiants. Un espace sémantique a été créé à partir de 27 articles écrits en anglais traitant du cœur humain "appris" par LSA. Les tests effectués concluent que la méthode LSA obtient des résultats satisfaisants comparativement à l'expertise humaine. Il en ressort que les mauvais résultats étaient dus à une absence de connaissances syntaxiques dans l'approche utilisée. Ainsi les travaux décrits ci-dessous montrent de quelle manière de telles connaissances peuvent être ajoutées à LSA. L'une des solutions peut consister à ajouter des connaissances syntaxiques à un corpus avant l'application de LSA.

La première approche de [Wiemer-Hastings & Zipitria, 2001] utilise des étiquettes grammaticales [Brill, 1994] appliquées à l'ensemble du corpus étudié (corpus de textes d'étudiants). Les étiquettes étant rattachées à chaque mot avec un blanc souligné ("_"), l'analyse qui s'en suit via LSA considère le mot associé à son étiquette comme un seul terme. Les résultats de calculs de similarités obtenus avec une telle méthode restent décevants. Notons que de telles informations grammaticales ne sont pas des connaissances syntaxiques proprement dites contrairement à la seconde approche de [Wiemer-Hastings & Zipitria, 2001] décrite ci-dessous. Cette seconde approche se traduit par l'utilisation d'un analyseur syntaxique afin de segmenter le texte avant d'appliquer l'analyse sémantique latente. Cette approche est appelée "LSA structurée" (SLSA). Une décomposition syntaxique des phrases en différents composants (sujet, verbe, objet) est tout d'abord effectuée. La similarité est ensuite calculée en traitant séparément par LSA les trois ensembles décrits précédemment. Les similarités (calcul du cosinus) entre les vecteurs des trois matrices formées sont alors évaluées. La moyenne des similarités est enfin calculée. Cette méthode a donné des résultats satisfaisants par rapport à "LSA classique" en augmentant la corrélation des scores obtenus avec les experts pour une tâche d'évaluation de réponses données par des étudiants à un test d'informatique.

[Kanejiya *et al.*, 2003] proposent un modèle appelé SELSA. Au lieu de générer une matrice de co-occurrences mot/document, il est proposé une matrice dans laquelle chaque ligne contient toutes les combinaisons mot_étiquette et en colonne les documents. L'étiquette "préfixe" renseigne sur le type grammatical du voisinage du mot traité. Le sens d'un mot est en effet donné par le voisinage grammatical duquel il est issu. Cette approche est assez similaire à l'utilisation des étiquettes de [Brill, 1994] présentée dans les travaux de [Wiemer-Hastings & Zipitria, 2001]. Mais SELSA étend ce travail vers un cadre plus général où un mot avec un contexte syntaxique spécifié par ses mots adjacents

est considéré comme une unité de représentation de connaissances.

Rappelons pour finir que notre approche *ExpLSA* utilise des informations syntaxiques et une mesure statistique afin d'enrichir un corpus. L'apport pour la méthode LSA est assez intuitif. Enrichir un corpus en ayant pour objectif de le vectorialiser avec LSA permet d'ajouter de l'information aux contextes. De plus, cet enrichissement se fait par l'utilisation d'informations syntaxiques qui manquent à LSA.

4.2 Première expérimentation évaluant SelDe : la classification conceptuelle

L'étude menée dans cette section porte sur la classification automatique de termes extraits grâce à des systèmes tels que ACABIT [Daille, 1994], LEXTER [Bourigault, 1993], SYNTEX [Bourigault & Fabre, 2000], EXIT [Roche *et al.*, 2004]. L'objectif est de regrouper les termes nominaux extraits avec EXIT. Ceux-ci sont des groupes de mots respectant des patrons syntaxiques (nom-préposition-nom, adjectif-nom, nom-adjectif, etc.). Nous décrivons de manière plus détaillée dans la section suivante le corpus d'où nous avons extrait ces termes ainsi que le protocole expérimental que nous avons suivi afin de mener nos expérimentations.
Notons que cette étude fut la première menée utilisant les descripteurs fournis par le modèle SELDE en vue d'enrichir un corpus [Béchet *et al.*, 2007].

4.2.1 Protocole expérimental

4.2.1.1 Description et caractéristiques du corpus étudié

Pour nos expérimentations nous nous appuyons sur un corpus qui traite d'un ensemble de textes écrits en français et qui sont issus du domaine des Ressources Humaines. Ce corpus a été rédigé par un psychologue de la société PerformanSe[18]. Les textes écrits correspondent à des commentaires de tests de psychologie de 378 individus. Le corpus comporte plus de 616 000 mots et 23 000 phrases pour une taille de 3,6 Mo une fois nettoyé. Un extrait du corpus est donné ci-dessous.

```
ATTENTION : Les relations humaines constituent peut-être un domaine
où vous pourriez être déçu. Puisque vous savez écouter, il semble
prudent de ne pas prendre de décision en la matière sans recueillir
```

[18] http ://www.performanse.fr/

```
de conseils. Sans jamais vous montrer envahissant, en gardant
toujours une part de réserve, vous savez être coopératif, voire
même conciliant. Les propos d'autrui, le comportement des groupes,
sont des sujets qui vous intéressent, et sur lesquels vous pouvez
travailler sans subir l'influence de l'un ou de l'autre; lorsque
vous donnez des conseils, c'est avec une certaine fermeté, sans
douter de vous.
```

Ce corpus a fait l'objet d'une expertise manuelle effectuée par Yves Kofratoff (LRI) en collaboration avec Serge Baquedano (Société PerformanSe) nous permettant ainsi de valider nos expérimentations. De cette expertise ressort une classification conceptuelle de l'ensemble des termes extraits par Exit, les concepts ayant été définis par l'expert.
L'expertise menée a abouti à la construction d'une classification conceptuelle en trois niveaux : 18 concepts pour le premier niveau, 3 pour le deuxième et 1 pour le troisième. Nous nous intéressons dans nos travaux à regrouper les termes dans les concepts de premier niveau. Notons que la répartition des termes dans les classes définies par cette classification du premier niveau est très hétérogène. Le tableau 4.1 montre cette répartition. Nous présentons dans le tableau 4.2 un exemple d'instances de concepts définies

Concept	Nombre d'instances
Relationnel	358
Environnement	329
Comportement&Attitude	266
Activité	263
Vous-Même	184
Rôle	154
Stress	57
Indépendance	40
Influence	38
Implication	28
Hiérarchie	27
Communication	18
Activité_Gestion&Administration	17
Savoir	7
Erreur	4
Expansion	2
JugementdeValeur	1

Tab. 4.1 – Répartition des termes extraits par Exit dans les concepts

par l'expert. Par exemple, l'expert a défini le concept de premier niveau "Relationnel"

Activité	Comportement et Attitude	Environnement
approches-rigoureuses	pleine-efficacité	grand-risque
domaines-déjà-explorés	voltes-faces	besoins-de-temps
efforts-nécessaire	brusque-revirement	grand-attrait
domaines-du-travail	saine-gestion	exigences-de-rigueur
résolutions-des-cas	périodes-de-lassitude	bon-climat
travail-de-conception-de-projets	mûres-réflexions	règle-de-base
point-d-appui	sans-gêne	milieux-agressifs
qualité-des-résultats	énergie-peu-commune	situation-risquée
aspects-purement-techniques	réaction-vive	cas-extrême

TAB. 4.2 – Exemple d'instances de concepts

dont les termes *confrontation ouverte*, *contact superficiel* et *entourage compréhensif* sont des instances.
Une caractéristique essentielle de ce corpus est qu'il utilise un vocabulaire spécialisé du domaine des ressources humaines. Par ailleurs, il contient des tournures de phrases revenant souvent, ce qui peut influencer positivement le traitement avec LSA comme l'ont montré [Roche & Chauché, 2006]. Citons par exemple les phrases :
– "*ATTENTION : Une telle orientation pourrait probablement s'avérer risquée pour votre évolution.*" et
– "*ATTENTION : Une telle orientation semble de nature à vous faire courir, le cas échéant, des risques importants.*".

Finalement, la nature syntagmatique des termes que nous souhaitons classifier impose un prétraitement du corpus. Nous lemmatisons dans un premier temps le corpus, en incluant les termes à classifier. Alors, ces termes sont identifiés dans le corpus. Cette identification consiste à représenter le terme par un seul mot (par exemple, le terme *attitude profondément participative* issu du corpus des Ressources Humaines devient *nom234* qui représente le 234ème terme parmi une liste extraite par EXIT).
Nous pouvons alors effectuer les expérimentations sur notre corpus en suivant le protocole présenté dans la section suivante.

4.2.1.2 Démarche expérimentale

La classification conceptuelle

Afin d'effectuer une classification conceptuelle à partir d'un corpus donné, nous représentons dans un premier temps ce corpus sous forme matricielle avec LSA. Dès lors, les lignes de la matrice LSA relatives aux termes extraits par EXIT servent d'entrée à

des algorithmes usuels de classification avec apprentissage supervisé[19] : les *k plus proches voisins* (k-ppv), les *machines à support vectoriel* (SVM) et l'approche *bayésienne naïve* (NaiveBayes). Ces algorithmes sont décrits dans la section 2.3.1.3.

Afin d'estimer la fiabilité de ces algorithmes avec les différentes approches, nous appliquons un processus de validation croisée[20] en segmentant les données en dix sous-ensembles. Cette méthode permet en effet de considérer alternativement les sous-ensembles comme des ensembles d'apprentissage ou comme des jeux de test dans un processus de classification avec apprentissage supervisé. Ces derniers sont évalués en utilisant les mesures de *rappel* et de *précision* pour lesquelles nous calculons la *micro* et la *macro-moyenne*. La différence majeure entre ces deux approches réside dans la prise en compte ou non de la quantité de données par classe. En effet, la micro-moyenne calcule une moyenne sur l'ensemble des candidats (nos termes à classer) là où la macro-moyenne va faire une moyenne par classe. Ainsi, une classe composée de 4 termes et une autre composée de 200 termes auront le même poids avec une macro-moyenne. Nous définissons ci-dessous ces deux moyennes.

La **micro-moyenne** consiste dans notre cas à calculer le rappel et la précision pour chaque article, ce qui revient finalement au calcul du taux d'exactitude [Slonim *et al.*, 2002]. Ainsi, dans notre contexte expérimental où un document est assigné à une seule classe, la micro-moyenne du rappel et de la précision sont égaux et peuvent être définis ainsi :

$$MicroMoyenne(rappel, précision) = \frac{termes\ correctement\ attribués\ à\ la\ classe}{nombre\ total\ de\ termes} \quad (4.1)$$

La **macro-moyenne** nécessite quant à elle de calculer le rappel et la précision pour chaque classe. Ainsi, pour chaque classe i ces mesures sont définies comme suit.

$$précision_i = \frac{nombre\ de\ termes\ correctement\ attribués\ à\ la\ classe\ i}{nombre\ de\ termes\ attribués\ à\ la\ classe\ i} \quad (4.2)$$

$$rappel_i = \frac{nombre\ de\ termes\ correctement\ attribués\ à\ la\ classe\ i}{nombre\ de\ termes\ appartenant\ à\ la\ classe\ i} \quad (4.3)$$

[19] Nous utilisons l'outil Weka afin d'appliquer ces algorithmes
http : *//www.cs.waikato.ac.nz/˜ml/*
[20] Réalisée avec l'application Weka.

Alors la macro-moyenne de la précision et du rappel n'est autre que respectivement la moyenne des $precision_i$ et la moyenne des $rappel_i$.
En général il est important de déterminer un compromis entre le rappel et la précision. Pour cela, nous pouvons utiliser une mesure prenant en compte ces deux critères d'évaluation en calculant le f-score tel que défini ci-dessous.

$$fscore(\beta) = \frac{(\beta^2 + 1) \times \text{précision} \times \text{rappel}}{\beta^2 \times \text{précision} + \text{rappel}} \tag{4.4}$$

Le paramètre β permet de régler les influences respectives de la précision et du rappel. Il est très souvent fixé à 1 pour accorder le même poids à ces deux mesures d'évaluation. Dans les sections suivantes, nous nous appuierons sur la mesure de f-score avec $\beta = 1$.

Discussion sur le choix des concepts à expérimenter

La répartition des différents termes à classifier dans les concepts présentés dans le tableau 4.1 se révèle être très inégale. Ainsi, nous nous interrogeons sur la qualité d'un apprentissage supervisé effectué sur des classes d'une telle hétérogénéité. Prenons par exemple les concepts "Erreur", "Expansion" et "Jugement de Valeur" ayant respectivement dans le corpus 4, 2 et 1 instances. En effectuant une validation croisée par segment de 10, il est presque impossible qu'un algorithme "apprenne" correctement les instances de ces concepts. Dans un cas, elles vont toutes être "apprises" et donc il ne restera pas d'instance de ces concepts à classifier dans la base de test. Dans l'autre cas, aucune de ces instances de concepts ne vont être apprises par l'algorithme dédié. Ainsi, les instances présentes dans la base de test ne pourront être classifiées positivement. Notons que cette disparité, bien qu'affectant la micro-moyenne, est plutôt perceptible lors de calcul de macro-moyenne. Rappelons que le calcul de macro-moyenne donne le même poids à chacune des classes (ou concepts) d'un corpus.
Ainsi, nous présenterons dans la section suivante, consacrée aux résultats expérimentaux, uniquement une classification conceptuelle effectuée sur les quatre concepts les plus représentatifs dans le corpus étudié à savoir les concepts "Relationnel", "Environnement", "Comportement et Attitude " et "Activité".

4.2.2 Résultats expérimentaux

4.2.2.1 Plan des expérimentations

L'objectif des travaux que nous avons menés ici est d'effectuer une classification conceptuelle des termes extraits avec Exit. En d'autres termes, nous proposons de classer automatiquement ces termes dans des concepts dont des instances sont présentées dans le tableau 4.2.
Nous mesurons la qualité de notre approche d'expansion ExpLSA en effectuant plusieurs expérimentations telles que décrites ci-dessous.

1. **Choix de la valeur du paramètre** k pour LSA et de **l'algorithme le plus performant**.
2. **Choix des paramètres pour** SelDe. L'objectif est de déterminer quels sont les paramètres les plus influents sur ce type d'expérimentations à savoir la classification conceptuelle.
3. **Mesure de la valeur ajoutée par l'expansion**. Nous comparons ici la qualité d'une classification avec un corpus enrichi ou non.
4. **Comparaison avec une autre méthode de la littérature**. Cette dernière expérimentation évalue ExpLSA à une méthode utilisant un étiqueteur grammatical.

4.2.2.2 Le choix du paramètre k de LSA et de l'algorithme

Cette section a pour objectif de déterminer quelle est la valeur la plus adaptée pour le paramètre k de l'approche LSA. Elle permettra également de déterminer quel va être l'algorithme donnant les meilleurs résultats en termes de f-score.
Nous effectuons les expérimentations de cette section avec notre corpus de ressources humaine, sans effectuer d'expansion. Les valeurs de k testées sont 100, 200, 300 et 400. Les algorithmes utilisés sont les k-ppv ($k = 10$, type de distance : $1/distance$), les SVM (avec un *noyau polynomial du second degré*) et l'approche Bayésienne Naïve. Les résultats obtenus sont présentés dans le tableau 4.3. Les résultats montrent que quel que soit l'algorithme employé, la valeur du paramètre k la plus adaptée à notre expérimentation est 100. Le contexte utilisé avec LSA est la phrase pour ce corpus. Ainsi, il est cohérent qu'une faible valeur de k soit plus adaptée. Notons également que cette valeur $k = 100$ est assez fréquemment employée dans la littérature afin de réduire un contexte "court" comme par exemple dans [Cederberg & Widdows, 2003]. Les auteurs cherchent ici à réduire un total de 1000 termes à 100 dans le but d'améliorer la précision et le rappel pour une tâche d'extraction automatique d'hyperonymes.
Par ailleurs, l'algorithme SVM a obtenu les meilleurs f-scores (avec la micro et macro

		F-score	
Algorithme	*Valeur de k*	*MicroMoy*	*MacroMoy*
SVM	***100***	**43,84%**	**42,10%**
	200	43,81%	42,07%
	300	43,76%	42,00%
	400	43,59%	41,92%
NaiveBayes	***100***	**40,49%**	**39,76%**
	200	38,37%	38,46%
	300	39,18%	39,62%
	400	38,45%	39,13%
K-ppv	***100***	**36,24%**	**33,90%**
	200	29,55%	27,97%
	300	29,31%	11,33%
	400	29,31%	11,33%

TAB. 4.3 – Comparaison de différents algorithmes et de différentes valeurs de k avec LSA.

moyenne). Notons cependant que ces scores restent très faibles. Rappelons que ce corpus de Ressources Humaines a la particularité d'être d'un domaine spécialisé. Ainsi, les concepts pour lesquels nous devons retrouver les instances de manière automatique en effectuant cette classification sont sémantiquement assez proches. Ceci explique les difficultés à classifier efficacement les termes de ce corpus.

4.2.2.3 L'enrichissement avec SelDe pour différents seuils d'Asium et choix du couple de verbes

Après avoir sélectionné l'algorithme et la valeur du paramètre k, nous présentons dans cette section les résultats obtenus avec l'enrichissement ExpLSA, en utilisant comme unique paramètre le seuil de la mesure d'ASIUM noté SA (décrit dans la section 3.4.3.3). Rappelons dès lors que les expérimentations effectuées utiliseront uniquement l'algorithme SVM et la valeur du paramètre k de LSA sera fixée à 100. Notons que l'utilisation de ExpLSA avec uniquement le paramètre SA, outre le fait de sélectionner les descripteurs afin d'enrichir un corpus, se traduit par une utilisation du modèle d'ASIUM de manière classique. Les descripteurs utilisés afin d'enrichir notre corpus sont issus de couples de verbes sémantiquement proches. Dans la section 4.1.1 nous proposions de sélectionner les couples soit en fonction du score de la mesure d'ASIUM ($ChVerb = SAsium$) soit en fonction d'un nombre d'occurrences ($ChVerb = Occurrences$). Les résultats obtenus en faisant varier le paramètre SA, avec une sélection des couples en fonction du score d'ASIUM ($ChVerb$=*"Asium"*), sont présentés dans le tableau 4.4. Les résultats montrent la micro moyenne, la macro moyenne et le taux d'enrichissement du corpus. Les descripteurs sélectionnés en utilisant uniquement le paramètre SA ne semblent pas satisfaisants. Nous constatons en effet que la dégradation des résultats est proportionnelle au taux d'en-

Type de corpus	MicroMoy	MacroMoy	Taux d'enrich.
Corpus initial	**43,84%**	**42,10%**	0,00%
C. Enrichi, SA = 0,9	43,59%	42,06%	8,91%
C. Enrichi, SA = 0,8	39,84%	37,97%	38,54%
C. Enrichi, SA = 0,7	37,55%	34,87%	50,79%
C. Enrichi, SA = 0,6	36,41%	34,49%	54,44%

TAB. 4.4 – Enrichissement selon une sélection de type $ChVerb = Asium$

richissement. Ainsi, les données servant à enrichir le corpus ne semblent pas adaptées. Ces résultats montrent qu'une simple sélection avec SA n'est pas suffisante car ce paramètre ne filtre pas le nombre d'objets servant à enrichir le corpus mais filtre uniquement le nombre de couples de verbes pouvant servir à l'enrichissement (cf. section 3.4.3.3).
Nous présentons dans le tableau 4.5 les résultats obtenus en faisant varier le paramètre SA, sélection des couples en fonction du nombre d'occurrences ($ChVerb$=*"Occurrences"*). Les résultats de ces tableaux sont assez similaires à ceux du tableau 4.4 avec des scores

Type de corpus	MicroMoy	MacroMoy	Taux d'enrich.
Corpus initial	43,84%	42,10%	0,00%
C. Enrichi, SA = 0,9	**44,73%**	**43,51%**	**9,44%**
C. Enrichi, SA = 0,8	37,80%	35,82%	40,24%
C. Enrichi, SA = 0,7	39,35%	37,06%	42,02%
C. Enrichi, SA = 0,6	36,82%	33,90%	41,41%

TAB. 4.5 – Enrichissement selon une sélection de type $ChVerb = Occurrences$

qui se dégradent en fonction du taux d'enrichissement. Cependant, ce type de sélection de couples de verbes semble plus pertinent, les scores étant supérieurs à ceux obtenus dans le tableau 4.4 pour un même SA. De plus, les scores obtenus avec $SA = 0,9$ sont assez encourageants car les résultats obtenus avec le corpus d'origine sont améliorés. Ces résultats s'expliquent par le fait qu'une valeur de SA proche de "1" limite le bruit apporté au corpus car les couples de verbes sélectionnés avec un tel score sont sémantiquement très proches (selon la mesure d'ASIUM). Il en ressort des descripteurs de qualités car les objets communs des verbes sont également proches sémantiquement. Ainsi, pour $SA = 0,9$, l'enrichissement apporté à ce corpus n'a pas généré de bruit (quel que soit la méthode de sélection de couple), confirmant la qualité des descripteurs utilisés pour l'enrichissement avec un tel seuil. Notons cependant que l'enrichissement effectué avec un SA à $0,9$ reste faible (environs 9% pour chaque méthode de sélection de couples) et ainsi, suivant la tâche pour laquelle ce corpus enrichi sera utilisé, peu d'informations nouvelles y seront introduites.
Notons par ailleurs que le taux d'enrichissement des corpus diffère en fonction du choix du couple de verbes (paramètre $ChVerb$). En effet, avec une sélection des couples en

fonction du nombre d'occurrences du terme à enrichir ($ChVerb = Occurrences$), le taux d'enrichissement est moins important qu'avec une sélection utilisant le seuil d'Asium ($ChVerb = Asium$). De plus, avec un $SA = 0,6$, l'enrichissement est moindre qu'avec un SA à $0,7$. Cela signifie que les couples sélectionnés avec cette approche possèdent moins d'objets en communs (ce qui semble cohérent car un faible score d'Asium peut se traduire par un faible nombre d'objets communs en fonction du nombre total d'objets).

4.2.2.4 Choix des paramètres de SelDe

Notre objectif est maintenant de déterminer quelles sont les valeurs des paramètres les mieux adaptées du modèle SelDe afin de conserver des descripteurs de qualité, quel que soit la valeur de SA. Les corpus enrichis pour ces expérimentations ont tous un seuil d'Asium fixé à 0,8, qui est un bon compromis entre qualité et quantité d'enrichissement. Rappelons en effet qu'avec $SA = 0,9$, nous obtiendrions un enrichissement de qualité mais en faible quantité. À l'inverse, $SA = 0,6$ produirait un corpus fortement enrichi mais assez bruité. Ainsi, notre choix s'est porté sur $SA = 0,8$, notre objectif étant de fixer les paramètres de SelDe, quel que soit le type s'enrichissement effectué (c.-à-d. quel que soit la valeur du seuil d'Asium). Les différents paramètres testés ici sont ceux décrits dans la section 3.4.3.3 (page 84). Ainsi, nous évaluons l'influence :

1. Du nombre minimal et maximal d'occurrences autorisé des objets communs dans un couple de verbes (*NbOccMin* et *NbOccMax*).
2. De la fréquence autorisée de termes résultant d'un couple de verbes (*nbObj*).
3. De l'ordre des termes dans un couple de verbes, classés en fonction du nombre d'occurrences de manière croissante ou décroissante (*order*).

Le nombre minimal (NbOccMin) et maximal d'occurrences (NbOccMax)
Les expérimentations présentées dans ce paragraphe ont pour objectif de sélectionner les valeurs les plus appropriées pour les paramètres *NbOccMin* et *NbOccMax*. Pour cela, nous sélectionnons, dans chaque couple de verbes servant à enrichir le corpus, les objets ayant un minimum et/ou un maximum d'occurrences (respectivement *NbOccMin* et *NbOccMax*). Ainsi, seuls ces objets seront utilisés afin d'enrichir le corpus. Nous avons ainsi évalué plusieurs valeurs pour ces paramètres à savoir pour un minimum et/ou un maximum de 2, 4, 6, 8 occurrences de chaque terme pour un couple donné. Notons que nous avons sélectionné les couples servant à enrichir le corpus avec les deux approches possibles ($ChVerb =$ "$Asium$" et $ChVerb =$ "$Occurrences$"). Nous présentons dans le tableau 4.6 une synthèse des meilleurs résultats obtenus. L'ensemble des expérimentations effectuées sont placées en annexes. Les résultats montrent que **les paramètres suivants sont les plus adaptés** :

Sél. Couple	NbMin	NbMax	MicroMoy	MacroMoy	Taux
Asium	6	-	44,57%	42,95%	3,72%
Occ	6	-	44,00%	42,30%	3,55%
Asium	2	4	43,84%	42,46%	15,05%
Occ	2	4	43,10%	41,69%	15,41%
Corpus initial			43,84%	42,10%	0,00%

TAB. 4.6 – Evaluation des paramètres NbOccMin et NbOccMax

– $NbOccMin = 6$ avec la sélection de couples ASIUM
– $NbOccMin$ et $NbOccMax = 2$ et 4 (sélection ASIUM)
La sélection des couples, une fois ces paramètres appliqués, donnent de meilleurs résultats avec la **sélection par le score d'Asium**, ce qui va à l'encontre des résultats obtenus précédemment (tableau 4.5). Notons que le paramètre $NbOccMax$ seul a été expérimenté (c.-à-d. pas uniquement utilisé conjointement au paramètre $NbOccMin$). Les meilleurs résultats ont été obtenus avec $NbOccMax = 4$, cependant ils sont identiques à ceux obtenus avec $NbOccMin = 2$ et $NbOccMax = 4$. Ainsi, nous ne conserverons que ces derniers, obtenant sur d'autres expérimentations de meilleurs résultats (notamment sur différents concepts[21]).

La fréquence autorisée (nbObj) et l'ordre des termes dans un couple de verbes (order)
Deux derniers paramètres restent à expérimenter : *nbObj*, le nombre d'objets à sélectionner par couple et *order*, l'ordre dans lesquels les objets sont triés (en terme de nombre d'occurrences, croissant ou décroissant). Ces résultats ne seront pas présentés dans cette section car ces deux paramètres n'ont pas apporté de résultats probants (ils sont néanmoins présentés en annexes). Le paramètre *nbObj* met en valeur les objets les plus ou les moins fréquents, en fonction de l'ordre choisi (paramètre *order*). Nous notons alors que les résultats décevants obtenus avec ce paramètre pourraient se justifier par la loi de Zipf. Cette dernière rappelons le indique que les termes les plus et les moins fréquents d'un corpus ne sont pas les plus discriminants pour des tâches de classification. L'utilisation de ces paramètres peut cependant être bénéfique à d'autres tâches.

Synthèse
Nous avons sélectionné, suite aux expérimentations présentées ci-dessus un certain nombre de paramètres qui vont définir notre approche d'enrichissement *ExpLSA*. Deux variantes seront utilisées. Notons que pour celles-ci, certains paramètres prennent la même valeur :

[21] Tous ces résultats sont présentés en annexes.

- *nbObj* = non défini
- *order* = "d" (décroissant)
- *Sélection des couples* = Ceux ayant le meilleur score d'Asium

Ainsi, ces deux variantes sont :

1. *NbMin* = 6, *NbMax* = indéfini
2. *NbMin* = 2, *NbMax* = 4

Nous noterons respectivement dans les expérimentations suivantes ces approches *ExpLSA_1* et *ExpLSA_2* avec un seuil d'Asium SA associé.

4.2.2.5 ExpLSA comparé à LSA

Cette section présente les résultats expérimentaux de notre approche *ExpLSA* pour les deux variantes retenues lors de notre sélection des valeurs de paramètres. Notre objectif est double en expérimentant *ExpLSA* :

– **Obtenir un corpus enrichi de qualité**. Dans un premier temps, nous cherchons à améliorer les résultats obtenus avec l'approche d'Asium "seule", à savoir, un enrichissement du corpus, sans traiter les objets servant à enrichir (c.-à-d. sans appliquer les paramètres de SelDe). Rappelons que pour un seuil d'Asium fixé à $0,6$, l'expansion étant assez importante mais la classification effectuée avec le corpus enrichi n'était pas pertinente. Avec les paramètres de SelDe, l'objectif est de réduire la quantité de descripteurs sélectionnés pour l'enrichissement, tout en conservant une classification de qualité, par rapport à la classification obtenue avec le corpus original. Il en résulte un corpus contenant de nouvelles informations cohérentes.

– **Améliorer la classification automatique**. Un second objectif est d'améliorer les résultats de classification conceptuelle. En effet, en utilisant SA à $0,9$ avec une sélection précise des termes servant à l'enrichissement avec nos paramètres, nous pouvons, bien qu'effectuant un très faible enrichissement, améliorer les résultats de classification.

Nous présentons dans le tableau 4.7 les résultats expérimentaux obtenus en utilisant le corpus original, l'expansion avec la méthode d'Asium sans paramètre (noté *Asium*) et les deux approches *ExpLSA*. Nous évaluons ces approches enrichissant le corpus avec deux seuils d'Asium : $0,6$ et $0,9$. Les trois algorithmes SVM, NaiveBayes et *k-ppv* ont été employés.

• Les résultats de ce tableau montrent que notre premier objectif est atteint. En effet, **les résultats de *ExpLSA* améliorent toujours ceux d'Asium**, avec un taux d'enrichissement réduit ne conservant que les descripteurs les plus pertinents et ceci

Algorithme	Type de corpus	MicroMoy	MacroMoy	Taux
SVM	Asium (SA =0,9)	43,59%	42,06%	9%
	ExpLSA_1 (SA=0,9)	44,33%	42,61%	1%
	ExpLSA_2 (SA=0,9)	**44,73%**	**43,69%**	4%
	Asium (SA = 0,6)	36,41%	34,49%	54%
	ExpLSA_1 (SA=0,6)	42,94%	41,14%	4%
	ExpLSA_2 (SA=0,6)	39,59%	38,00%	24%
	Corpus initial	43,84%	42,10%	0%
K-ppv	Asium (SA =0,9)	33,63%	31,85%	9%
	ExpLSA_1 (SA=0,9)	**38,61%**	**36,75%**	1%
	ExpLSA_2 (SA=0,9)	35,84%	33,99%	4%
	Asium (SA = 0,6)	33,55%	31,70%	54%
	ExpLSA_1 (SA=0,6)	35,76%	33,86%	4%
	ExpLSA_2 (SA=0,6)	32,49%	30,78%	24%
	Corpus initial	36,24%	33,90%	0%
NaiveBayes	Asium (SA =0,9)	40,00%	39,62%	9%
	ExpLSA_1 (SA=0,9)	40,82%	40,03%	1%
	ExpLSA_2 (SA=0,9)	**41,71%**	**41,16%**	4%
	Asium (SA = 0,6)	33,31%	31,90%	54%
	ExpLSA_1 (SA=0,6)	39,59%	39,04%	4%
	ExpLSA_2 (SA=0,6)	39,43%	38,30%	24%
	Corpus initial	40,49%	39,76%	0%

TAB. 4.7 – Résultats comparatifs de LSA et *ExpLSA*

quel que soit l'algorithme de classification utilisé. Nous passons par exemple pour la micro-moyenne, $SA = 0,6$, de 36,41% avec ASIUM à **42,94**% avec *ExpLSA_1* pour les *SVM*. Nous avons ainsi produit des **corpus enrichis de qualités** avec les deux approches *ExpLSA*, les résultats de classification étant proches ou légèrement supérieurs à ceux obtenus avec le corpus d'origine (c.-à-d. sans enrichissement). Ces deux approches ont cependant des comportements qui varient en fonction de l'algorithme et du seuil d'ASIUM.

• L'atteinte de notre second objectif est quant à elle plus discutable. Certes les résultats de classification sont légèrement améliorés. Citons par exemple avec l'algorithme $k - ppv$ l'approche *ExpLSA* ($SA = 0,9$) qui améliore de plus de 2 points les résultats obtenus avec le corpus original. Cependant, ces résultats restent très faibles ($\approx 40\%$) et ne peuvent prétendre à une classification "pertinente" de nos termes dans leur quatre concepts respectifs.

Nous discuterons de ces résultats dans la section 4.2.3.

4.2.2.6 ExpLSA comparé à l'approche utilisant TreeTagger

L'approche *ExpLSA* propose d'enrichir un corpus afin d'ensuite appliquer une vectorialisation avec LSA. Un des objectifs de cet enrichissement est de combler les lacunes d'une méthode LSA ne prenant en compte aucune information syntaxique. Ainsi,

afin de mesurer la qualité de cette approche en terme de classification, nous confrontons dans cette section notre approche *ExpLSA* à une approche de la littérature [Wiemer-Hastings & Zipitria, 2001] dont l'objectif est le même : apporter des informations syntaxiques supplémentaires à LSA. Nous nommerons cette approche *TreeTagger* pour le fait qu'elle propose d'utiliser un corpus étiqueté grammaticalement afin qu'il serve de base d'apprentissage tel que décrit dans la section 4.1.4. Notons cependant que l'auteur utilise dans la version originale de l'approche l'étiqueteur de Brill, mais le principe de l'approche reste inchangé. Le tableau 4.8 montre les résultats comparant *ExpLSA* et

Algorithme	*Type de corpus*	*MicroMoy*	*MacroMoy*
SVM	TreeTagger	42,20%	40,36%
	Asium	43,59%	42,06%
	ExpLSA	**44,73%**	**43,69%**
	Corpus initial	43,84%	42,10%
K-ppv	**TreeTagger**	**36,33%**	**34,44%**
	Asium	33,63%	31,85%
	ExpLSA	35,84%	33,99%
	Corpus initial	36,24%	33,90%
NaiveBayes	**TreeTagger**	**42,12%**	**41,24%**
	Asium	40,00%	39,62%
	ExpLSA	41,71%	41,16%
	Corpus initial	40,49%	39,76%

Tab. 4.8 – Résultats comparatifs de *ExpLSA* et l'approche utilisant TreeTagger

l'approche fondée sur l'utilisation d'un étiqueteur "TreeTagger". L'approche *ExpLSA* utilisée ici est *ExpLSA_2* telle que définie dans la section précédente. Ces résultats montrent que l'approche TreeTagger semble plus adaptée à l'algorithme NaiveBayes en obtenant des résultats supérieurs à ceux d'*ExpLSA* et du corpus original. Concernant les *k-ppv*, l'approche TreeTagger obtient des résultats similaires à ceux obtenus avec le corpus original. Rappelons cependant que si l'approche TreeTagger est ici meilleure qu'*ExpLSA*, cette dernière est plus adaptée en utilisant l'approche *ExpLSA_1* pour les *k-ppv* comme montré dans la section précédente. Finalement, les SVM, qui obtiennent les meilleurs résultats lors de ces expérimentations, **sont améliorés uniquement avec l'approche *ExpLSA***. *ExpLSA* avec les SVM obtient en effet **les meilleures micro et macro moyennes, tout algorithme et méthode confondues**. Cette amélioration est cependant assez faible, de l'ordre de 1% pour la micro-moyenne par rapport au corpus initial. Nous pouvons ainsi conclure qu'*ExpLSA* donne des résultats encourageants mais pas suffisants en terme de f-score. L'approche TreeTagger quant à elle ne semble pas adaptée.

4.2.3 Synthèse et discussions

Nous avons présenté dans cette section 4.2 des expérimentations visant à évaluer la qualité de notre modèle SelDe dans le cadre d'une classification conceptuelle. Le but était de regrouper automatiquement un ensemble de termes extraits par l'outil Exit. Rappelons que ces termes sont situés au niveau des lignes dans les matrices construites avec LSA et ExpLSA. Deux objectifs étaient visés :

1. Évaluer la qualité des descripteurs extraits par SelDe
2. Mesurer l'impact d'un enrichissement avec *ExpLSA*

Ces expérimentations ont confirmé la qualité des descripteurs extraits avec SelDe. En effet, en nous appuyant sur les expérimentations menées sur un corpus enrichi par rapport au corpus initial, les résultats obtenus en termes de f-score sont du même ordre. Ceci indique que l'intégrité du corpus est maintenue et que les informations ayant servi à l'enrichir sont de qualité.
Par ailleurs, l'approche *ExpLSA* utilisant le modèle SelDe afin d'enrichir un corpus ne s'est pas toujours montrée adaptée pour une tâche de classification conceptuelle fondée sur des matrices d'occurrences. En effet, la tâche à effectuer est relativement complexe. Citons les points suivants montrant la difficulté de celle-ci.

- **Le caractère spécialisé du corpus**. Le corpus utilisé traite de Ressources Humaines, et emploie un vocabulaire très spécifique. Il en résulte une classification conceptuelle dans des classes (ou concepts) de thématiques très fines comme "comportement et attitude" et "relationnel".
- **La taille des contextes**. Rappelons que la taille des contextes est un élément important afin d'obtenir des résultats pertinents avec LSA. Cependant, notre contexte lors de ces expérimentations est la phrase, ne contenant pas assez de mots pour LSA. Notre corpus ayant en effet un nombre moyen de 25 termes par phrase. Rappelons qu'il est spécifié dans la littérature qu'un contexte minimum de 60 termes est nécessaire afin d'obtenir de bonnes performances avec LSA [Rehder *et al.*, 1998].
- **L'expertise humaine**. Les expérimentations effectuées ici sont très dépendantes de l'évaluation manuelle menée. Celle-ci est en effet très subjective. Un travail considérable a été effectué par le *LRI* (Laboratoire de Recherche en Informatique d'Orsay) et la société *PerformanSe* afin de fournir cette classification, mais celle-ci reste perfectible. En effet, un autre expert aurait probablement eu une vision différente pour certains termes.

4.3 Seconde application pour évaluer SelDe : la classification de textes

Les expérimentations menées dans cette section portent sur la classification automatique de textes. Une telle classification est une tâche où le contexte des données est primordial, notamment lors de l'emploi d'approches statistiques afin de réaliser cette tâche. Par exemple, [Salton *et al.*, 1983] produisent des résultats à partir des calculs de fréquence d'occurrences de termes extraits. Nous utilisons ici *LSA* avec notre approche d'enrichissement *ExpLSA* afin d'en évaluer la qualité et de confirmer également la qualité des descripteurs extraits avec SelDe. À l'instar d'une tâche de classification conceptuelle, l'utilisation d'un contexte plus riche produit par notre enrichissement va permettre, *a priori*, un apprentissage plus aisé.
Le contexte utilisé avec l'approche LSA, comme nous venons de l'expliciter précédemment, est primordial. Ce contexte est défini par la donnée textuelle sur laquelle la classification de textes est réalisée. Il s'agit dans notre cas de dépêches journalistiques. Ainsi, un certain nombre de facteurs peuvent influencer une classification de textes. Nous présentons ceux-ci dans la section suivante.

4.3.1 L'impact des différents types de données textuelles sur la classification de textes

4.3.1.1 Taille des documents

La taille des contextes est un aspect important pour une tâche de classification. Le contexte utilisé dans ces travaux est le document.
Une classification de textes avec un apprentissage effectué sur des documents de faible taille est une tâche difficile. En effet, le nombre trop faible de descripteurs "appris" ne peut permettre une classification optimale. Afin d'améliorer la qualité de la base d'apprentissage, une solution consiste à enrichir cette base avec de nouvelles connaissances comme dans les travaux de *Zelikovitz* [Zelikovitz & Hirsh, 2000], [Zelikovitz, 2004] qui proposent d'utiliser une base d'apprentissage de donnés étiquetées et un second jeu de données non étiqueté afin d'assister le travail de classification automatique. Notre approche se fonde sur le même principe en enrichissant un corpus initial afin d'en augmenter le nombre de descripteurs.

4.3.1.2 Taille des corpus

Nous nous intéressons également à la taille des corpus et à leur impact sur notre approche d'expansion. L'expansion est dépendante du nombre de relations syntaxiques extraites d'un corpus. En effet, d'un nombre trop faible de relations syntaxiques va résulter une expansion réduite. *A contrario*, un nombre important de relations syntaxiques va générer une expansion conséquente, qui pourra introduire un certain nombre de données bruitées. Il est alors ici nécessaire de faire varier les paramètres $nbOccMin$, $nbOccMax$ et/ou $nbObj(Order)$ définis dans la section 3.4.3.3 afin de filtrer le nombre de termes candidats à l'expansion de corpus.

4.3.1.3 Thème du corpus

La thématique d'un corpus est un facteur important lors de la classification automatique de textes. En effet, le domaine d'un corpus influence la nature et le nombre de classes. Un corpus d'un style journalistique contenant des articles d'actualité va par exemple être constitué d'un nombre important de classes. De plus, les classes d'un corpus peuvent être très éloignées et facilement séparables comme une classe *science* et une classe *sport* ou à l'inverse très proches comme les classes *société* et *politique*. Par ailleurs, la thématique générale d'un corpus va influencer la qualité des relations syntaxiques extraites pour l'expansion. Dans un corpus d'opinion par exemple, l'homogénéité du domaine du corpus va permettre d'obtenir des relations syntaxiques très pertinentes. Néanmoins, la classification automatique d'un corpus d'opinion reste une tâche difficile de par la nature binaire des classes d'un même domaine (opinion positive ou négative). Nous trouvons en effet un certain nombre d'articles de la littérature confirmant ce fait. Citons par exemple [Dini & Mazzini, 2002] qui proposent d'effectuer une extraction d'opinions de clients sur différents produits de consommation. Les auteurs confirment la difficulté d'extraire de telles informations par le biais de patrons syntaxiques, assez difficiles à identifier. Ils indiquent par ailleurs que les approches de type "sac de mots" peuvent être dans certains cas efficaces mais ne permettent pas une finesse suffisante afin notamment de résoudre des problèmes liés à la négation.

Nous présentons dans la section suivante le protocole expérimental suivi pour évaluer la qualité de l'approche *ExpLSA* dans le cadre de la classification de textes. Nous présentons dans un premier temps les corpus utilisés dans ces expérimentations. Les expérimentations détaillées dans la section 4.3.3 proposent de mesurer l'impact de l'approche *ExpLSA* sur deux types de corpus, données d'opinion et dépêches

journalistiques.

4.3.2 Protocole expérimental

4.3.2.1 Description des corpus étudiés

Type de Corpus	*Thématique*	*Tailles des Articles*	*Nombre d'articles*	*Nombre de mots*	*Taille en Mo*
Référence	Actualité	*Tous*	2 972	1 224 109	7,0
Grand corpus	Actualité	*Tous*	14 863	6 181 873	35,1
Petit corpus	Actualité	*Tous*	1 486	615 194	3,5
Longs articles	Actualité	Supérieurs à 550 mots	6 751	4 284 487	24,3
Petits articles	Actualité	Inférieurs à 550 mots	8 111	1 897 386	10,9
Grand corpus	Opinions	*Tous*	17 298	4 816 618	27,5
Petit corpus	Opinions	*Tous*	1 729	489 924	2,8
Longs articles	Opinions	Supérieurs à 450 mots	8 731	3 806 292	21,6
Petits articles	Opinions	Inférieurs à 450 mots	8 566	1 010 326	5,8

TAB. 4.9 – Caractéristiques des corpus étudiés

Afin de mesurer l'impact des différents types de corpus, tel que présenté dans la section précédente, nous proposons d'effectuer des expérimentations sur plusieurs corpus écrits en français.
Nous étudions dans un premier temps l'impact de la thématique du corpus. Ainsi, nous expérimenterons la classification de textes pour plusieurs corpus d'actualités contenant un ensemble de dépêches provenant du site d'information de *Yahoo !* (http ://fr.news.yahoo.com/) et un corpus de débats d'opinions parlementaires à l'Assemblée Nationale provenant de la campagne d'évaluation DEFT de 2007 (http ://deft07.limsi.fr). Les corpus de dépêches sont divisés en onze classes telles que "sport", "insolite" ou "France". Ceux d'opinions sont quant à eux divisés en deux classes visant à déterminer si le locuteur appartient à l'opposition ou bien à la majorité. Notons que ces corpus résultent de comptes rendus des débats à l'Assemblée Nationale au cours de la XIIe législature (2002/2007).
Pour les deux domaines précédemment explicités, nous expérimentons la taille des articles (courts ou longs) et la taille des corpus (courts ou longs). Pour finir, un corpus *de référence* de taille moyenne de style journalistique sera utilisé afin de sélectionner les paramètres d'*ExpLSA*. Ce dernier est indépendant des corpus de tests, son objectif étant de sélectionner au mieux les paramètres de SELDE afin de pouvoir appliquer cette méthode. Notons également que ce corpus possède une taille moyenne. Il est de plus d'un domaine général. Nous n'avons pas opté pour le choix d'un corpus de référence contenant des données d'opinions. En effet, la tâche de classification de données d'opinions est trop spécifique

afin de permettre une sélection adéquate de nos paramètres.
Nous présentons dans le tableau 4.9 les différentes caractéristiques des corpus étudiés. Par ailleurs, tous ces corpus sont au préalable lemmatisés.

4.3.2.2 Démarche expérimentale

L'objectif des expérimentations est de comparer la méthode LSA à notre approche *ExpLSA* en réalisant une classification automatique d'articles dans un contexte de classification supervisée. Pour cela, nous avons expérimenté les trois mêmes algorithmes de classification supervisée[22] déjà appliqués dans le but de construire une classification conceptuelle : les *k plus proches voisins* – k-ppv ($k = 10$, type de distance : $1/distance$), les *machines à support vectoriel* – SVM (avec un *noyau polynomial du second degré*) et l'approche *bayésienne naïve* – NaiveBayes. Ainsi, nous respecterons le même protocole expérimental proposé lors des expérimentations menées dans la section 4.2.
Nous appliquerons une validation croisée de 10. Nous mesurerons la *micro* et la *macro-moyenne* du rappel et de la précision pour finalement présenter les résultats sous forme de f-scores avec un paramètre $\beta = 1$. Notons finalement qu'outre la lemmatisation, aucun pré-traitement n'est appliqué aux corpus.

4.3.3 Résultats expérimentaux

4.3.3.1 Plan des expérimentations

Le plan listé comme suit pour expérimenter la classification de documents textuels est assez similaire à celui des expérimentations précédentes.

1. Choix de la valeur du paramètre k pour LSA pour le reste des expérimentations et sélection de l'algorithme le plus performant. Ainsi, nous nous fonderons sur cet algorithme afin de déterminer la valeur des paramètres.
2. Choix des paramètres pour SelDe. Les paramètres définis pour la classification conceptuelle ne sont en effet pas nécessairement adaptés à cette tâche de classification de textes.
3. Mesure de la valeur ajoutée par l'expansion. Il est ici évalué la qualité d'*ExpLSA* avec LSA et Asium (enrichissement sans paramètres) pour les différents corpus.

Notons que les choix de la valeur de k, de l'algorithme et des paramètres de SelDe s'effectuent avec le *corpus de référence* tel que présenté dans la section 4.3.2.1.

[22] Nous utilisons l'outil Weka afin d'appliquer ces algorithmes.
http : //www.cs.waikato.ac.nz/˜ml/

4.3.3.2 Le choix du paramètre k de LSA et de l'algorithme

Nous évaluons dans cette section quelle est la valeur du paramètre k la plus adaptée à nos expérimentations. Pour cela, nous comparons pour le corpus de référence quatre valeurs de k : 100, 200, 300 et 400. Nous calculons la micro et la macro moyenne du rappel et de la précision afin d'obtenir un f-score. Nous évaluons par la même les différents algorithmes déjà expérimentés. Le tableau 4.10 présente les résultats expérimentaux obtenus. Ces résultats montrent que le paramètre k est dépendant de l'algorithme de classification,

		F-score	
Algorithme	*Valeur de k*	*MicroMoy*	*MacroMoy*
SVM	*100*	72,82%	67,78%
	200	74,13%	70,40%
	300	**74,87%**	**71,23%**
	400	74,34%	71,08%
NaiveBayes	*100*	68,62%	64,81%
	200	69,79%	66,96%
	300	70,90%	68,81%
	400	**72,49%**	**70,76%**
K-ppv	*100*	71,81%	67,35%
	200	**72,49%**	**69,44%**
	300	71,51%	68,76%
	400	70,70%	67,28%

Tab. 4.10 – Choix du paramètre k et de l'algorithme

ce qui n'était pas le cas lors de classification conceptuelle (cf. section 4.2.2.2). Cependant, l'algorithme SVM reste celui obtenant de meilleurs résultats, ce qui confirme les résultats obtenus pour la classification conceptuelle. Pour cet algorithme, le paramètre k fixé à 300 donne les meilleurs scores. Le choix de cette valeur montre qu'un corpus écrit en français, dans notre cas, obtient les meilleurs performances avec LSA ce qui va dans le sens de [Landauer & Dumais, 1997]. Ces derniers montrent que la valeur de k est optimum lorsqu'elle vaut 300 pour les corpus anglais.
Ainsi, pour les expérimentations suivantes, nous emploierons ces deux critères : SVM avec $k = 300$. Notons que le choix du paramètre k n'est pas anodin. En effet, avec par exemple l'algorithme NaiveBayes, les f-scores varient de 68,62% à 72,49% pour la micro-moyenne et de 64,81% à 70,76% pour la macro-moyenne. Précisons pour finir que le choix du corpus de référence pour effectuer la sélection des paramètres de SelDe est tout à fait justifié. En effet, le domaine de ce corpus, dépêches d'actualités avec un style journalistique, est plus neutre en termes de classification de textes qu'un corpus contenant des données d'opinions par exemple. De plus, sa taille et son nombre "moyen" d'articles est un bon compromis afin de déterminer les paramètres les plus adaptés, quel que soit le corpus.

4.3.3.3 L'enrichissement avec SelDe pour différents seuils d'Asium et choix du couple de verbes

Plusieurs objectifs sont visés avec l'enrichissement via *ExpLSA*. Le premier est de montrer la **qualité de nos descripteurs vis-à-vis de la méthode d'Asium**, dont les résultats sont présentés dans cette section. La méthode d'Asium "simple" propose, rappelons le, d'enrichir un corpus *sans utiliser de paramètres*. Lors de nos précédentes expérimentations, nous avons montré que l'enrichissement d'Asium était amélioré en utilisant les paramètres de SelDe. Cependant, cette amélioration était parfois assez faible. Néanmoins, le taux d'enrichissement d'un corpus avec Asium est très dépendant de la taille initial du corpus à enrichir. Il en résulte alors un grand nombre de relations syntaxiques extraites et ainsi un fort enrichissement, qui bien sur est d'autant plus important que le seuil d'Asium est faible.
Le second objectif fixé est l'**amélioration des résultats de classifications**.
Nous présentons respectivement dans les tableau 4.11 et 4.12 les résultats obtenus avec

Type de corpus	*MicroMoy*	*MacroMoy*	*Taux d'enr.*
corpus initial	**74,87%**	71,23%	0,00%
C. enrichi, *SA = 0,6*	71,91%	67,50%	120,03%
C. enrichi, *SA = 0,7*	73,26%	69,47%	95,99%
C. enrichi, *SA = 0,8*	74,34%	**71,33%**	50,64%
C. enrichi, *SA = 0,9*	74,47%	71,20%	2,75%

Tab. 4.11 – Enrichissement avec Asium, sélection par score Asium

la méthode d'Asium en faisant varier la méthode du choix du couple (cf. section 4.1.1), choix selon SA puis choix selon le nombre d'occurrences. Les résultats montrent que quel

Type de corpus	*MicroMoy*	*MacroMoy*	*Taux d'enr.*
corpus initial	**74,87%**	**71,23%**	0,00%
C. enrichi, *SA = 0,6*	72,38%	68,14%	104,31%
C. enrichi, *SA = 0,7*	73,49%	70,06%	93,32%
C. enrichi, *SA = 0,8*	73,33%	69,59%	50,68%
C. enrichi, *SA = 0,9*	74,47%	71,20%	2,75%

Tab. 4.12 – Enrichissement avec Asium, sélection par nombre d'occurrences

que soit le choix du couple, les résultats se dégradent quand le seuil d'Asium diminue. Notons cependant une amélioration des résultats avec la macro-moyenne du f-score, pour la sélection du couple par scores d'Asium et pour $SA = 0,8$ les résultats sont légèrement améliorés. Ce résultat n'est néanmoins pas significatif.
Notons par ailleurs que le taux d'enrichissement est très important avec $SA = 0,6$ ce qui

confirme les hypothèses précédemment avancées concernant le manque de filtrage.
Remarquons finalement que le nombre trop faible de relations syntaxiques utilisables avec un score d'Asium à 0,9 explique que les résultats obtenus pour $SA = 0,9$ soient les mêmes avec $ChVerb =$ "*Occurrences*" et $ChVerb =$ "*Asium*". En effet, cela signifie qu'un seul couple de verbes n'est possible pour l'expansion. Dès lors la méthode de sélection du couple n'a pas d'importance.

4.3.3.4 Choix des paramètres de SelDe

Cette section a pour objectif de sélectionner les paramètres du modèle SelDe qui seront les plus adaptés afin d'enrichir un corpus en vue d'une classification automatique de textes. Nous utiliserons un corpus enrichi avec $SA = 0,8$ à l'instar des expérimentations de classification conceptuelle, ce seuil étant un bon compromis qualité/quantité afin de sélectionner aux mieux les valeurs de nos paramètres. Nous évaluerons ici l'influence des mêmes paramètres (présentés en section 3.4.3.3) que ceux évalués pour la tâche de classification conceptuelle (section 4.2) : *NbOccMin*, *NbOccMax*, *nbObj*, *order*. Afin de présenter les résultats obtenus pour ces paramètres, nous suivrons le protocole expérimental précédemment décrit. Nous ne présentons que les résultats les plus significatifs, les autres étant placés en annexes.

Le nombre minimal (NbOccMin) et maximal d'occurrence (NbOccMax)
Nous mesurons dans un premier temps l'impact des paramètres NbOccMin et NbOccMax. Nous évaluons pour ces deux paramètres les valeurs *2, 4, 6* et *8*. Notons que nous faisons également varier le type de sélection du couple, se faisant avec $ChVerb = Asium$ ou $ChVerb = Occurrences$. Les résultats sont présentés dans le tableau 4.13. À l'instar

Sél. Couple	*NbOccMin*	*NbOccMax*	*MicroMoy*	*MacroMoy*	*Taux*
Asium	4	-	**75,55%**	**72,47%**	9,27%
Asium	-	4	74,27%	71,13%	38,70%
Occ	3	5	**74,94%**	**71,57%**	10,25%
Corpus original			74,87%	71,23%	-

Tab. 4.13 – Évaluation des paramètres "NbOccMin" et "NbOccMax".

des résultats de classification conceptuelle, le paramètre le plus influent reste *NbOccMin*, dont les meilleurs f-scores ont été ici obtenus avec une valeur de 4. Nous expliquons l'influence de ce paramètre par le fait qu'il privilégie les termes fréquents. Rappelons le principe d'enrichissement. Les objets communs de couples de verbes jugés proches avec Asium sont sélectionnés afin d'enrichir un corpus. Avec le paramètre *NbOccMin*, nous limitons ces objets à ceux ayant ici au minimum 4 occurrences communes entre les

deux verbes du couple. Ainsi, les objets rares sont écartés du processus d'enrichissement. Ces objets peu fréquents génèrent une quantité de bruit plus importante que ceux très largement présents. En effet, pour ces derniers, même s'ils sont bruités, il est plus probable qu'ils soient également fortement présents dans le corpus et donc dans d'autres couples. L'enrichissement est alors "égalisé". Pour les objets rares, nous pouvons supposer qu'ils sont moins présents dans le corpus et ainsi, l'enrichissement effectué avec ceux-ci sera assez ponctuel et sujet au bruit.

La fréquence autorisée (nbObj) et l'ordre des termes dans un couple de verbes (order)

La fréquence autorisée (nbObj) qui, rappelons le est le nombre d'objets à sélectionner par couple et le paramètre *order*, l'ordre dans lesquels les objets sont triés (en terme de nombre d'occurrences, croissant ou décroissant), n'ont pas donné de résultats pertinents lors de la sélection des paramètres adaptés à la classification conceptuelle. Les résultats obtenus avec la classification de textes sont assez différents tels que montré dans le tableau 4.14. En effet, les scores résultant de ces paramètres améliorent les résultats, notamment

Sél. Couple	Order	nbObj	*MicroMoy*	*MacroMoy*	*Taux*
Asium	c	2	**75,55%**	72,51%	8,60%
Occ	c	2	75,31%	72,30%	8,58%
Occ	d	2	75,38%	**72,95%**	7,23%
Corpus original			74,87%	71,23%	-

TAB. 4.14 – Évaluation des paramètres "nbObj" et "Order"

la macro moyenne en passant de 71,23% à 72,95%.

Choix des paramètres pour l'approche ExpLSA

Nous avons sélectionné, suite aux expérimentations présentées ci-dessus un certain nombre de paramètres qui vont définir notre approche d'enrichissement *ExpLSA*. Rappelons que ces expérimentations vont être menées avec le corpus de *référence*. Ce dernier est en effet à distinguer des autres corpus qui serviront à évaluer la qualité de l'approche *ExpLSA*.

Une première approche nommée *ExpLSA_1*, utilisera un enrichissement fondé sur le modèle SELDE en utilisant le paramètre *nbObj* fixé à 2 et le paramètre *order* à "c". La sélection du couple de verbes est alors effectuée en fonction du score d'ASIUM.

La seconde approche nommée *ExpLSA_2* utilisera quant à elle les paramètres *NbOccMin* et *NbOccMax* aux valeurs respectives 3 et 5. Notons que l'utilisation simultanée de ces

deux seuils est plus bénéfique que l'utilisation seul de *NbOccMin* ou *NbOccMax*. Certes l'approche avec *NbOccMin* fixé à 4 donne les meilleures f-scores (dans la tableau 4.13). Cependant, le corpus de référence est de taille moyenne. Avec l'utilisation d'un corpus de plus grande taille, nous pouvons supposer que l'expansion avec $NbOccMin = 4$ serait trop importante, introduisant du bruit. Ainsi, nous avons opté pour l'utilisation de *NbOccMin* et *NbOccMax* simultanément. Ce choix de paramètres produira des descripteurs de moindre qualité avec les petits corpus, mais de meilleure qualité avec ceux de taille plus importante, constituant un bon compromis. Finalement, nous présenterons dans la section suivante les résultats des approches Asium (l'enrichissement sans filtrage avec les paramètres de SelDe), *ExpLSA_1* et *ExpLSA_2*, et les résultats obtenus avec le corpus original.

4.3.3.5 Résultats obtenus

Nous présentons dans cette section les résultats expérimentaux obtenus avec notre approche *ExpLSA* pour les deux variantes définies lors de la sélection des paramètres. Parmi les résultats expérimentaux obtenus, nous ne présenterons, pour des raisons de lisibilité, qu'uniquement ceux obtenus avec l'algorithme *SVM*. Les résultats obtenus pour les algorithmes *NaiveBayes* et *k-ppv* sont présentés en annexes. Nous utiliserons, comme lors des expérimentations avec la classification conceptuelle, les seuils d'Asium 0,6 et 0,9. Précisons également que les résultats présentés pour l'approche *ExpLSA* (1 et 2) respecteront la légende suivante.

– Scores en gras : l'approche *ExpLSA* améliore les résultats obtenus avec LSA seul (le corpus original).

– Scores en italiques : l'approche *ExpLSA* améliore les résultats obtenus avec Asium (sans paramètres).

Les objectifs fixés résultants de ces expérimentations sont les suivants.

– **Confirmer la qualité des descripteurs**. Nous avons en effet montré lors des expérimentations de classification conceptuelle que l'approche *ExpLSA* proposait un enrichissement de qualité, en ne dégradant pas ou peu les résultats de classification conceptuelle. L'objectif est donc de confirmer ces résultats.

– **Améliorer les résultats de classification de textes**. La classification de textes et la classification conceptuelle sont deux tâches distinctes, et ayant un comportement différent face à un enrichissement tel que celui proposé dans nos approches. Ainsi, enrichir un corpus afin d'effectuer une classification de textes ne rend pas uniquement le contexte plus riche, comme avec la classification conceptuelle.

– **Obtenir de meilleurs scores avec les paramètres de SelDe**. Notre dernier objectif est d'obtenir de meilleurs résultats que l'approche Asium, qui n'utilise pas les paramètres de filtrages introduits dans cette thèse. Ce dernier point sera plus visible avec un seuil d'Asium faible comme pour les résultats de classification conceptuelle.

Rappelons finalement que notre tâche première est de **mesurer la qualité des descripteurs de SelDe et de construire un corpus enrichi de qualité**. Ainsi, une amélioration des résultats de classification est un bon indicateur de qualité mais des scores équivalents le sont également comme nous allons le discuter en fin de chapitre. Par ailleurs, nous n'avons pas cherché à améliorer la qualité d'un algorithme de classification ou bien à effectuer des prétraitements avancés sur nos corpus, notre objectif étant d'étudier les descripteurs de SelDe.

Résultats selon la taille du corpus

Les résultats présentés dans cette section comparent, pour les corpus d'opinions et de dépêches, les micro et macro moyennes des f-scores des corpus de grandes (tableau 4.15) et petites tailles (4.16). Ils montrent dans un premier temps que la tâche de classification de données d'opinions est plus complexe que celle de dépêches. Ces résultats peuvent sembler contradictoires car pour les données d'opinions, seules deux classes sont distinguées là où les dépêches sont réparties en 11. Cependant, les classes des dépêches sont assez distinctes, facilitant un apprentissage supervisé de qualité. La classe "*sport*" est par exemple éloignée thématiquement de la classe "*économie*". Les deux classes du corpus de données d'opinions sont quant à elles plus difficiles à "apprendre". La négation n'est par exemple pas considérée, ou bien encore l'ironie, pouvant apparaître dans un corpus politique comme celui-ci.

Pour les corpus de grandes tailles, avec SA 0,9, les résultats de classification obtenus avec *ExpLSA_1* sont du même ordre que ceux du corpus original. Avec $SA = 0,6$, les scores de *ExpLSA_1* sont plus faibles mais restent meilleurs que Asium ou *ExpLSA_2*. Ces résultats s'expliquent par le fait qu'un grand corpus contient un nombre important d'informations. Un enrichissement est donc moins utile qu'avec un corpus de taille plus réduite. En effet, les informations apportées par l'expansion peuvent contenir du bruit, même avec le filtrage des paramètres. Par ailleurs, les approches *ExpLSA_1* et *ExpLSA_2* **améliorent systématiquement les scores de l'approche Asium**. Cela est d'autant plus visible pour les deux corpus avec $SA = 0,6$. Nous passons par exemple avec le corpus des dépêches d'une micro-moyenne de f-score de 64,74% pour Asium à 78,04% avec *ExpLSA_1*.

Seuil d'Asium	*Type*	*Dépêches*		*Opinions*	
		MicroMoy	*MacroMoy*	*MicroMoy*	*MacroMoy*
	original	**79,25%**	**76,78%**	**70,55%**	**68,62%**
SA = 0,9	Asium	75,73%	72,45%	61,66%	56,96%
	ExpLSA1	*79,04%*	*76,58%*	*70,18%*	*68,17%*
	ExpLSA2	*77,83%*	*75,31%*	*67,81%*	*65,06%*
SA = 0,6	Asium	64,74%	61,23%	60,30%	54,78%
	ExpLSA1	*78,04%*	*75,47%*	*68,55%*	*66,34%*
	ExpLSA2	*75,47%*	*72,64%*	*65,45%*	*62,14%*

TAB. 4.15 – Comparaison des f-scores pour les grand corpus

Seuil d'Asium	*Type*	*Dépêches*		*Opinions*	
		MicroMoy	*MacroMoy*	*MicroMoy*	*MacroMoy*
	original	69,74%	66,78%	65,95%	62,92%
SA = 0,9	Asium	70,14%	67,58%	**66,65%**	**63,72%**
	ExpLSA1	69,60%	67,40%	66,24%	63,21%
	ExpLSA2	69,74%	66,78%	65,95%	62,92%
SA = 0,6	Asium	70,54%	67,79%	65,09%	61,40%
	ExpLSA1	70,48%	67,59%	64,22%	60,99%
	ExpLSA2	***70,88%***	***68,55%***	*66,07%*	*63,09%*

TAB. 4.16 – Comparaison des f-scores pour les petits corpus

Les résultats avec les petits corpus sont assez différents et donnent des résultats de bonne qualité pour les deux corpus (dépêches et opinions) avec l'approche *ExpLSA_2* et un seuil d'Asium à 0,6. Ces corpus étant de faibles tailles, à l'opposé des grands corpus, l'enrichissement sera moins conséquent. Ainsi, un seuil d'Asium plus faible permet un enrichissement plus important, sans pour autant ajouter de bruit. L'approche Asium avec un seuil $SA = 0,9$ obtient également des résultats pertinents, les meilleurs sont obtenus avec le corpus d'opinions. Cependant, *ExpLSA* pour ce même seuil est en retrait. Une faible expansion, qui en plus se voit réduite par les paramètres de SelDe devient de moins bonne qualité car trop faible en terme de quantité.

Pour résumer les résultats obtenus, nous proposons d'utiliser les approches suivantes afin d'enrichir un corpus, suivant la taille et la thématique du corpus.
– Corpus de taille importante : *ExpLSA_1* $SA = 0,9$
– Corpus de faible taille : *ExpLSA_2* $SA = 0,6$ ou Asium $SA = 0,9$

Résultats selon la taille des articles

Ce paragraphe présente les résultats obtenus en fonction de la taille des articles de nos corpus d'opinions et de dépêches. Les résultats pour les grands et petits articles (définis dans le tableau 4.9) sont respectivement présentés dans les tableaux 4.17 et 4.18.
Nous confirmons dans un premier temps avec ces résultats qu'une tâche de classification de textes d'opinions est plus difficile qu'avec des dépêches.
Concernant les corpus de tailles importantes, les résultats obtenus avec *ExpLSA_1* sont assez similaires à ceux obtenus avec le corpus original. Rappelons que dans le paragraphe précédent, les mêmes résultats expérimentaux étaient obtenus avec les "grands corpus". Ces corpus ont les mêmes caractéristiques que les grands corpus. Leurs tailles sont conséquentes (cf. tableau 4.9). De plus, une base d'apprentissage importante est "apprise" avec ces "grands articles". Ainsi, à l'instar des "grands corpus" évalués dans le paragraphe précédent, l'enrichissement apporte peu d'informations nouvelles réellement "utiles" afin de mener une tâche de classification. Notons cependant que l'approche *ExpLSA* améliore systématiquement les résultats de l'approche Asium, ce qui confirme que les corpus de grandes tailles et ceux contenant de grands articles aient les mêmes caractéristiques.
Les corpus contenant des petits articles se distinguent suivant leurs thématiques. L'approche *ExpLSA* avec le corpus des dépêches améliore systématiquement les résultats d'Asium, mais jamais ceux du corpus d'origine, l'approche *ExpLSA_1* avec $SA = 0,9$ s'en rapprochant cependant, offrant alors dans ce cas la meilleure approche d'enrichissement. Ces résultats ne sont pas similaires à ceux obtenus avec les "petits corpus". Notons qu'avec de grands articles, le corpus produit est assez important et se comporte vis-à-vis

Seuil d'Asium	*Type*	*Dépêches*		*Opinions*	
		MicroMoy	*MacroMoy*	*MicroMoy*	*MacroMoy*
	original	81,03%	75,43%	75,05%	**72,36%**
SA = 0,9	Asium	80,02%	74,07%	73,14%	69,82%
	ExpLSA1	***81,28%***	***75,78%***	***75,07%***	***72,36%***
	ExpLSA2	*80,27%*	*74,30%*	*73,29%*	*70,07%*
SA = 0,6	Asium	72,36%	65,75%	66,79%	61,13%
	ExpLSA1	*79,35%*	*73,79%*	*74,19%*	*71,39%*
	ExpLSA2	*78,36%*	*72,07%*	*69,78%*	*65,62%*

TAB. 4.17 – Comparaison des f-scores pour les grand articles

de l'enrichissement comme un "grand corpus". Ces faits ne se vérifient cependant pas avec les petits corpus. Ces derniers ont certes une base d'apprentissage assez faible, mais de qualité car le corpus contient des articles de toutes tailles.

Seuil d'Asium	*Type*	*Dépêches*		*Opinions*	
		MicroMoy	*MacroMoy*	*MicroMoy*	*MacroMoy*
	original	**78,16%**	**75,34%**	65,30%	64,07%
SA = 0,9	Asium	75,94%	72,60%	**65,80%**	64,57%
	ExpLSA1	*77,74%*	*74,72%*	65,68%	64,47%
	ExpLSA2	*76,95%*	*73,74%*	**65,80%**	***64,62%***
SA = 0,6	Asium	72,14%	67,98%	63,01%	61,39%
	ExpLSA1	*76,82%*	*73,43%*	*65,31%*	*64,02%*
	ExpLSA2	*76,76%*	*73,67%*	*65,39%*	*64,04%*

TAB. 4.18 – Comparaison des f-scores pour les petits articles

Finalement, nous proposons d'utiliser quel que soit le thème, la taille ou le type d'article l'approche *ExpLSA_1* avec $SA = 0,9$.

Comparaison avec l'approche TreeTagger

Nous comparons dans cette section les résultats de ExpLSA avec l'approche TreeTagger, déjà expérimentée en section 4.2.2.6. Ces expérimentations ont pour but de comparer ExpLSA avec autre méthode sur des résultats de classifications uniquement. Il ne sera en effet pas discuté de la qualité de nos descripteurs.

Nous montrons dans le tableau 4.19 les résultats obtenus avec cette approche. Notons que les résultats de ce tableau traitent uniquement le corpus d'opinions de petite taille ainsi que celui contenant de petits articles. Les f-scores produits par l'approche TreeTagger

pour les autres corpus ne sont pas significatifs. En effet, ils améliorent ou dégradent les résultats du corpus original de manière non significative.

Corpus	*Type*	*Opinions*	
		MicroMoy	*MacroMoy*
Petits articles	original	65,30%	64,07%
	Asium (SA=0,9)	65,80%	64,57%
	ExpLSA2 (SA=0,9)	65,80%	64,62%
	Tagger	**66,06%**	**64,84%**
Petit corpus	original	65,95%	62,92%
	Asium (SA=0,9)	66,65%	63,72%
	ExpLSA2 (SA=0,9)	65,95%	62,92%
	Tagger	**66,94%**	**64,17%**

Tab. 4.19 – Comparaison de ExpLSA à l'approche "TreeTagger"

Les résultats du tableau 4.19 montrent que l'approche TreeTagger semble adaptée aux corpus d'opinions de faibles tailles, ou contenant de petits articles. En d'autres termes, cette approche améliore les scores de corpus d'opinions à faible base d'apprentissage. Remarquons cependant que cette amélioration reste faible par rapport à l'approche Asium pour $SA = 0,9$. Rappelons par ailleurs que les paramètres de SelDe ont été sélectionnés sur un corpus de même domaine que les dépêches (le corpus de référence). Ainsi, les faibles résultats obtenus pour cette approche avec les opinions peuvent résulter d'un mauvais choix de paramètres.

4.3.4 Synthèse et discussions

La section 4.3 a présenté les expérimentations menées dans le cadre de l'utilisation des descripteurs extraits par le modèle SelDe pour une tâche de classification de textes. Ces expérimentations avaient pour but de classer de manière automatique des documents textuels, pouvant être enrichis par notre approche *ExpLSA*, dans des catégories existantes. Ces expérimentations avaient les mêmes objectifs que celles effectuées avec la classification conceptuelle à savoir :

1. Évaluer la qualité des descripteurs extraits par SelDe
2. Mesurer l'impact d'un enrichissement avec *ExpLSA* par rapport à un enrichissement de type Asium

Dans ce chapitre, **nous avons de nouveau confirmé la qualité des descripteurs** sélectionnés par notre approche. De plus, nous avons identifié suivant le type, le thème,

la taille des corpus et la taille des articles de ces corpus quels étaient les enrichissements les plus adaptés afin de produire un corpus de qualité. Par ailleurs, nous avons montré qu'avec des corpus de tailles importantes ou des corpus constitués de grands articles, **l'approche *ExpLSA* améliorait systématiquement les résultats obtenus avec l'approche Asium**. Cette amélioration est d'autant plus significative avec un $SA = 0,6$, contenant un maximum de bruit avec Asium. Ainsi, *ExpLSA* réduit ce bruit de manière importante.
Néanmoins, dans le but d'améliorer les résultats de classification de textes, l'approche *ExpLSA* ne s'est de nouveau pas montrée pertinente, améliorant ponctuellement et de manière non significative les résultats de f-score. Notons que l'apport de connaissances sémantiques n'améliore pas de manière significative les performances d'algorithmes dans la littérature. Nos expérimentations confirment donc ces résultats, notre expansion pouvant être ramenée à ce type d'approche. Rappelons en effet que nous effectuons une expansion de termes avec d'autres termes sémantiquement proches car partageant le même contexte. Notre valeur ajoutée se situe au niveau même de ces connaissances qui proviennent du corpus même, et non pas de connaissances du domaine parfois difficiles à obtenir.
Ces résultats mitigés peuvent par ailleurs s'expliquer par la non introduction lors de l'enrichissement du contexte de l'article. En effet, la méthode d'enrichissement *ExpLSA* propose d'enrichir tout terme d'un corpus par des descripteurs extraits avec SelDe, dès lors que ce terme apparaît dans un couple de verbes. L'inconvénient est que chaque terme identique sera enrichi d'une façon similaire, quel que soit l'article dans lequel il est situé. Ce dernier point ne doit pas être négligé si l'on cherche à enrichir de manière "intelligente" le corpus en vue d'une classification automatique. Nous évoquerons des pistes permettant d'effectuer un tel enrichissement dans le chapitre 8 en perspective.

Le modèle SelDe utilisé par *ExpLSA* afin d'enrichir un corpus est fondé sur une analyse syntaxique. Qu'en est-il alors des données textuelles syntaxiquement pauvres ? Ce modèle de sélection de descripteurs n'est, par exemple, pas applicable avec des corpus contenant des *Curiculum Vitæ*. En effet, un nombre trop faible de relations syntaxiques sera extrait et le modèle SelDe ne pourra être employé. Le chapitre suivant propose alors une alternative pour ce type de données, des données textuelles dites "complexes".
Notons pour finir que les approches de types sac de mots ne sont pas nécessairement adaptées à la construction de classes conceptuelles (comme l'utilisation de LSA dans ce chapitre). Nous aborderons dans le chapitre 6 une autre méthode de construction de classes conceptuelles utilisant les descripteurs de SelDe et de "SelDeF". Ce dernier est un nouveau modèle proposant un filtrage supplémentaire nous permettant d'utiliser les objets dits "complémentaires". Les expérimentations qui seront menées dans ce chapitre

vont permettre de mettre davantage en avant la qualité de ces modèles de sélection de descripteurs pertinents.

Chapitre 5

Quel modèle appliquer sur les données complexes

Sommaire

Nous présentons dans ce chapitre un récapitulatif de divers travaux.
Les travaux présentés en section 5.3 ont été publiés dans [7 - IIP'08] et [18 - QDC'08].
Ceux présentés en section 5.4 ont été publiés dans [1 - IJDEM] et [2 - RNTI].
Finalement, les travaux décrits en section 5.5 ont été publiés dans [4 - ISMIS'09], [10 - QSI'08] et [13 - TALN'09]

5.1 Introduction

5.1.1 Les limites du modèle SelDe

Le chapitre 4 propose d'utiliser le modèle de sélection de descripteurs SelDe afin d'effectuer des tâches de classification de données textuelles. Nous discutons en fin de ce chapitre la notion de dépendance vis-à-vis de la syntaxe de ce modèle. En effet, suivant le

type de données textuelles traité, SELDE peut être difficile à utiliser. Citons par exemple les données textuelles issues de blogs, ou bien encore les données provenant de *Curriculum Vitæ* (CV). Un exemple d'article issu d'un blog est présenté ci-dessous. Il provient du site de blog de SKYROCK (*skyrock.com*).

```
1 - Le bélouga
Nom scientifique : Delphinapterus leucas
Taille : de 3 à 4,5 m
Poids : de 400 à 1030 kg
Alimentation : Poissons, krill ou autres crustacés, invertébrés
(calmars, vers marins ou poulpes).
Ils ont besoin de manger 12 kg de nourriture par jour
Espérance de vie : 30 ans
Répartition géographique : dans l'océan Arctique surtout près des côtes
des Etats-Unis, du Canada, de l'Alaska, du Groenland, de la Norvège et
de la Russie.
```

La seule relation syntaxique extraite de cet article avec l'analyseur SYGFRAN est "manger nourriture". Ainsi, l'approche d'extraction de descripteurs SELDE ne peut s'appliquer. Pour traiter ces données, que nous qualifions de **complexes** de par leur pauvreté syntaxique, nous proposons dans ce chapitre d'autres modèles d'extraction de descripteurs, moins dépendants des informations syntaxiques d'un corpus. Nous définissons alors dans un premier temps, les données complexes d'une manière générale.

5.1.2 Les données textuelles complexes

De nos jours, avec l'essor grandissant d'internet, de nombreuses données émergeantes pouvant être qualifiées de complexes. Ainsi, des images, des vidéos ou bien encore des textes non structurés sont qualifiés comme tels. Nous pouvons définir une telle donnée si elle possède l'une des caractéristiques suivantes [Zighed & Loudcher, 2004 à 2007].

- **Les données sont très volumineuses**. Elles peuvent représenter une grande quantité de données étant de l'ordre du téraoctet. Citons par exemple la base de données du site Web d'Amazon (*www.amazon.com*) qui contient plus de 42 téraoctets de données[23]. Notons qu'une tâche de *TREC* (Text REtrieval Conference) est dédiée à cette problématique (citons par exemple [Hawking & Thistlewaite, 1998]).
- **Le caractère distribué des données**. Le Dossier Médical Personnalisé (DMP) peut par exemple être stocké dans divers établissements médicaux, en fonction des

[23]Information provenant du site http ://www.businessintelligencelowdown.com

endroits où un patient a été hospitalisé.
- **L'hétérogénéité des données**. Des données peuvent en effet être de différents types. Un blog peut par exemple contenir des données textuelles, des vidéos et des images.
- **L'évolutivité des données**. La meilleure source de données à caractère évolutif est sans conteste le Web, en constante transformation.
- **Le caractère non structuré des données**. Citons par exemple des textes issus de CV, qui bien que respectant un certain schéma, ne sont pas structurés syntaxiquement, contenant peu de phrases.

Finalement nous pouvons définir les données textuelles complexes comme ne respectant pas une structure syntaxique telles une grammaire de langue naturelle ou une syntaxe de langage de programmation. Ainsi, nous nous focalisons dans ce chapitre sur des corpus écrits en français, mais ne respectant peu ou pas une grammaire permettant l'extraction de relations syntaxiques. En effet, bien que le modèle SelDe propose des descripteurs de qualité, le nombre important de données textuelles complexes nous ont poussé à proposer d'autres approches plus adaptées aux données traitées. L'objectif de nos travaux reste cependant lié à la problématique d'extraction de descripteurs pertinents pour différentes tâches décrites dans ce chapitre.

5.1.3 Plan du chapitre

Ce chapitre est organisé de la manière suivante. Nous présentons dans un premier temps l'approche de sélection de descripteurs adaptée aux données complexes (section 5.2). Dès lors nous présenterons trois tâches diverses employant ce modèle. Les deux premières tâches résultent de travaux menés par des stagiaires en entreprise que j'ai co-encadrés, afin de mettre en place une solution de classification automatique de données textuelles complexes. La première approche présentée en section 5.3 propose de classifier des données issues de blogs. Avec ces données plus ou moins bien formulées syntaxiquement, nous évaluerons la qualité de notre modèle. La seconde présentée en section 5.4 s'intéresse aux données bruitées ou incomplètes, résultant de rétro-conversion de numérisation OCR. Nous montrerons dans cette section que notre approche n'est pas nécessairement adaptée, et proposerons alors une alternative de sélection de descripteurs à partir de ces données complexes. Finalement, nous présenterons dans la section 5.5 les résultats obtenus lors de travaux menés en collaboration avec le Laboratoire en Informatique d'Avignon (LIA). Ces derniers travaux proposent de réduire l'implication humaine pour un recruteur. Les données traitées sont mal ou pas structurées. Leur traitement constitue la tâche la plus difficile de ce chapitre.

Notons que les expérimentations menées respectent des protocoles expérimentaux divers,

du fait des différentes collaborations effectuées.

5.2 De la sélection de descripteurs à un modèle de classification de données textuelles complexes

5.2.1 L'extraction des descripteurs

Les informations syntaxiques d'un texte restent selon nous des informations très pertinentes. Cependant, elles ne peuvent pas être utilisées sur des données textuelles complexes, et *a fortiori* avec SelDe. Ainsi, nous présentons dans cette section un modèle d'extraction de descripteurs adapté à ce type de données. Ce modèle se fonde également sur l'information syntaxique contenue dans un corpus. Cependant, les données peuvent être obtenues avec des outils statistiques qui sont moins dépendants de la structure syntaxique d'un corpus. Notre approche se fonde sur l'utilisation de catégories lexicales ou grammaticales telles que "nom", "adjectif" ou encore "verbe". Ce type de données a déjà été utilisé dans différents contextes comme dans l'approche de [Wiemer-Hastings & Zipitria, 2001] précédemment explicitée dans la section 4.1.4 qui propose de représenter un terme par le terme lui même associé à sa catégorie lexicale. D'autres travaux de la littérature comme ceux de [Kohomban & Lee, 2007] montrent par ailleurs que les noms sont très porteurs de sens.
Notre utilisation des catégories lexicales est assez différente et propose non pas d'ajouter de l'information au corpus initial comme c'est le cas dans l'approche de [Wiemer-Hastings & Zipitria, 2001] mais de supprimer de l'information *inutile*. Ainsi, les descripteurs utilisés avec ce type de données seront les mots appartenant à des catégories lexicales données. Ce modèle sera appliqué à la classification de données textuelles complexes. Alors, nous proposons l'approche suivante afin de produire un corpus à classifier.

5.2.2 Le modèle de classification

Le principe de l'approche de classification de données textuelles complexes est de ne conserver que les termes (descripteurs) d'un corpus qui vont appartenir à une catégorie lexicale fixée. Par exemple, nous ne conservons uniquement d'un corpus que les *noms*. Nous introduisons également la notion de pondérations des termes en fonction de la "qualité sémantique" d'une catégorie. Nous pouvons par exemple, en supposant la construction d'un corpus avec uniquement les verbes et les noms de celui-ci, accorder un poids plus important aux noms et un poids moindre aux verbes d'un corpus.
Notons par ailleurs que dans toutes nos expérimentations utilisant le modèle de classification proposé, nous avons extrait les catégories lexicales avec l'outil TreeTagger

[Schmid, 1995] pour sa simplicité, son efficacité et ses performances qualitatives.
Cet étiqueteur est développé au sein du projet TC (Textcorpora and sungswerkzeuge) à l'institut de linguistique computationnelle de l'université de Stuttgart. Ce système d'annotation de catégories morpho-syntaxiques permet d'étiqueter des textes dans différentes langues rendant notre approche multi-langues. Cet étiqueteur est fondé sur la notion d'apprentissage de corpus. Il ressort de cet apprentissage un certain nombre de règles lexicales et syntaxiques permettant l'étiquetage de nouveaux textes. L'étiqueteur, pour une suite de mots données va alors "prédire" la plus probable des catégories lexicales pouvant survenir. Ces probabilités sont construites à partir d'un ensemble de tri-grammes connus de mots, pouvant être définis comme une suite de trois étiquettes grammaticales consécutives constituant l'ensemble d'apprentissage (cf chapitre 2).
Le TreeTagger propose par exemple les résultats suivants pour la phrase : *Les étiquettes lexicales apportent une information supplémentaire.*

Les	DET :ART	le
étiquettes	NOM	étiquette
lexicales	ADJ	lexical
apportent	VER :pres	apporter
une	DET :ART	un
information	NOM	information
supplémentaire	ADJ	supplémentaire
.	SENT	.

La première colonne correspond au terme traité (forme fléchie), la seconde nous renseigne sur la catégorie lexicale de ce terme et la dernière nous donne sa forme lemmatisée.
Une première expérimentation effectuée avec cette approche est présentée dans la section suivante. Elle propose de classer automatiquement des articles issus de blogs en utilisant le modèle présenté dans cette section. Nous évaluons ainsi la qualité de notre approche avec des données textuelles syntaxiquement mal formulées.

5.3 Traitement des données issues de blogs

5.3.1 Contexte

Les travaux présentés dans cette section sont issus d'une collaboration avec la Société PaperBlog (http ://www.paperblog.fr/) qui héberge un site web proposant un référencement de blogs (ou weblog) issus de sites web partenaires. Cette collaboration est intervenue dans le cadre du stage d'ingénieur d'Inès Bayoudh que j'ai co-encadré avec

Mathieu Roche.

Les blogs s'apparentent à des sites Web constitués d'articles souvent ordonnés chronologiquement ou ante-chronologiquement. Chaque article est écrit à la manière d'un journal de bord, pour lequel des commentaires peuvent être apportés. Ce nouveau type de site Web, illustrant les concepts du Web 2.0, s'est popularisé ces dernières années du fait de sa facilité de publication, de son interactivité et pour finir d'une grande liberté d'expression. Ce dernier point pose le problème de la recherche d'information dans de tels articles.

L'idée du site web de PaperBlog est de répondre à la question : comment trouver des articles d'une thématique précise issue de blogs ? Pour cela, les articles des blogs sont évalués suivant leur pertinence puis associés à une catégorie thématique (comme *culture*, *informatique*, *insolite* etc.). Cette approche permet de retrouver des informations d'une thématique précise contenues dans les blogs. L'objectif de nos travaux est d'apporter une méthode qui effectue cette classification thématique de manière automatique (celle-ci étant actuellement réalisée manuellement) en minimisant le taux d'erreur lors de cette classification.

5.3.2 Protocole expérimental

Pour effectuer cette classification, nous avons choisi d'implémenter un algorithme classique de classification de données textuelles, les k plus proches voisins (k-ppv). Se pose alors le choix des descripteurs à utiliser pour réaliser une telle tâche. Nous avons alors proposé l'utilisation des catégories lexicales (notre modèle présenté en section 5.2) afin de réduire la base d'apprentissage mais également le temps de traitement des nouveaux articles à classer.
Les expérimentations ont été menées avec un échantillon d'articles du site de PaperBlog, d'une taille de 3,4 Mo contenant 2520 articles et composé de plus de 400 000 mots. Celui-ci est réparti en cinq classes : "alimentation", "talents", "people", "cuisine" et "bourse". Une première étape de pré-traitement du corpus est alors effectuée (Suppression des balises "*html*" et des *stop words*). Nous sélectionnons alors les différents descripteurs dont nous souhaitons mesurer la qualité :

- La forme fléchie du mot issu du corpus original.
- La forme lemmatisée du mot (obtenu avec l'outil TreeTagger).
- Les approches sélectionnant les catégories lexicales : les Noms, les Verbes et les Adjectifs.

– Les combinaisons de catégories lexicales.
– Les pondérations de ces catégories.

Nous avons alors choisi de comparer deux formes de représentations vectorielles, la matrice de co-occurrences définie par *Salton* et la même matrice mais pondérée par l'approche *tf-idf* (définie en section 2.2.3.1).
La métrique d'évaluation utilisée permettant de mesurer le taux d'articles mal classés est le *taux d'erreur* défini comme suit :

$$taux\ d'erreur = \frac{nombre\ d'articles\ mal\ classés}{nombre\ total\ d'articles} \tag{5.1}$$

Ce taux moyen est mesuré en effectuant une validation croisée en segmentant les données en cinq sous-ensembles et utilisation des k-ppv pour catégoriser les articles.

5.3.3 Résultats expérimentaux

Nous proposons dans un premier temps de mesurer l'impact de la pondération *tf-idf* et de la lemmatisation de notre corpus. Le tableau 5.1 présente les taux d'erreurs obtenus pour la tâche de classification automatique des documents de ce corpus en appliquant ces différents traitements. Nous montrons avec ces résultats que la lemmatisation a tendance

	Taux d'erreur	
Type de corpus	Tf	Tf-Idf
Formes fléchis	0,39	0,25
Lemmes	0,42	**0,21**

Tab. 5.1 – Influence du *tf-idf* et de la lemmatisation sur notre corpus

à augmenter le taux d'erreurs. Cependant, appliqué avec le *tf-idf*, la lemmatisation réduit l'erreur et le taux le plus faible est obtenu avec la combinaison *tf-idf* /Lemmatisation.

Nous présentons alors dans le tableau 5.2 les taux d'erreurs obtenus pour la sélection des catégories lexicales *Nom*, *Verbe*, *Adjectif* et les combinaisons *Nom-Verbe*, *Nom-Adjectif*, *Verbe-Adjectif*, *Nom-Verbe-Adjectif* (avec et sans *tf-idf*). Les résultats du tableau 5.2 montrent que le fait de ne conserver que les noms et les verbes de notre corpus en utilisant l'approche *tf-idf* permet d'obtenir un taux d'erreur similaire à celui obtenu avec la totalité du corpus $(0,21)$. De plus, les adjectifs sont ici, dans le cadre de données complexes issues d'articles de blogs de thématiques générales, peu porteurs d'informations. Par ailleurs, les noms sont les plus riches en termes d'information avec un taux d'erreur à $0,27$ soit assez proche de ceux obtenus avec le corpus entier. Nous

	Taux d'erreur	
Type de corpus	Tf	Tf-Idf
Lemmes	0,42	**0,21**
Nom	0,33	0,27
Verbe	0,58	0,47
Adjectif	0,51	0,44
Nom-Verbe	0,27	**0,21**
Nom-Adj	0,36	0,27
Verbe-Adj	0,34	0,29
Nom-Verbe-Adj	0,36	0,27

TAB. 5.2 – Résultats obtenus avec la sélection de catégories lexicales

montrons finalement avec ces expérimentations sur les catégories lexicales qu'avec une quantité réduite d'informations (uniquement les noms ou les noms et verbes) nous pouvons classer efficacement les données complexes issus de blogs en réduisant ainsi la taille de la base d'apprentissage et *a fortiori* le temps de traitement pour les nouveaux articles à classifier.

Au regard des résultats obtenus dans le tableau 5.2, nous avons montré l'impact de la sélection de catégories lexicales particulières et leur influence sur la tâche de classification. Nous nous focalisons alors sur le *tf-idf* qui donne les meilleurs résultats sur nos données. Nous proposons d'attribuer de poids à différentes catégories lexicales avec le *tf-idf*. Cela se traduit par la multiplication du poids calculé par le *tf-idf* par un scalaire fixé. Par exemple, si nous fixons à 3 le poids des noms, nous multiplions par 3 tous les poids des noms de notre matrice. Les résultats de ces pondérations sont présentés dans le tableau 5.3. La pondération par 3 de catégories lexicales a montré une réduction significative du

Poids			
Nom	Verbe	Adjectif	*Taux d'erreur*
1	2	1	0,31
1	1	2	0,30
2	1	1	0,31
2	2	1	0,29
1	2	2	0,31
2	1	2	0,23

Poids			
Nom	Verbe	Adjectif	*Taux d'erreur*
1	3	1	0,10
1	1	3	0,29
3	1	1	0,11
3	3	1	**0,06**
1	3	3	0,10
3	1	3	0,09

TAB. 5.3 – Pondération des catégories lexicales dans la matrice du *tf-idf*

taux d'erreur avec un taux de $0,06$ pour un poids de 3 attribué aux verbes et aux noms. Ces résultats confirment la qualité des descripteurs extraits avec notre approche.

Après avoir montré la qualité de notre approche avec des données textuelles syntaxiquement mal formulées, nous proposons dans la section suivante d'expérimenter des données

bruitées ou incomplètes.

5.4 La sélection de descripteurs appliquée aux données bruitées

5.4.1 Contexte

Les travaux présentés dans cette section sont issus d'une collaboration avec la Société Itesoft (www.itesoft.fr). Les prestations fournies par cette Société sont des solutions dédiées à la dématérialisation, au traitement automatique et à la gestion de tous les documents entrant dans l'entreprise tels que des courriers, factures, chèques, formulaires ou encore des bons de commande. La collaboration avec Itesoft est intervenue dans le cadre du stage de Master II recherche de Sami Laroum que j'ai co-encadré avec Mathieu Roche et Hatem Hamza (Itesoft).
Outre l'utilisation du modèle présenté en section 5.2, nos travaux ont permis la mise en place d'une méthode permettant de réaliser une classification automatique de documents issus de numérisation OCR (Reconnaissance Optique de Caractères). Un procédé OCR fournit aux systèmes d'imagerie et de numérisation la capacité de transformer les images de caractères imprimés en caractères lisibles numériquement. Il est à dissocier du procédé ICR qui fournit aux systèmes d'imagerie et de numérisation la capacité de transformer les images de caractères écrits à la main en caractères lisibles numériquement.

Dans cette étude conduite pour améliorer les performances de la classification automatique de documents textuels issus d'OCR (Reconnaissance Optique de Caractères), nous proposons d'évaluer la pertinence de différents descripteurs fréquemment employés avec des données textuelles complexes. Nos travaux reposent sur l'évaluation de descripteurs robustes aux données bruitées. Ainsi, une fois la qualité de l'ensemble des descripteurs évaluée, nous présentons une approche nommée **HYBRED** (**HYB**rid **RE**presentation of **D**ocuments) combinant les descripteurs de la littérature, adaptés aux données complexes et ceux de notre modèle.

Afin de tester la pertinence de notre approche, nous nous appuyons sur des corpus de la société Itesoft. Les documents de ces corpus sont répartis dans différentes classes spécialisées telles que *attestation salaire*, *facture optique* ou encore *frais médecin*. Ces corpus ont la particularité de provenir de rétro-conversion d'OCR, engendrant une quantité non négligeable de bruit, pouvant notamment être des fautes d'orthographe, des lettres manquantes, etc. d'où leurs dénomination "complexes". Les données qui seront

traitées lors de ces expérimentations sont assez diversifiées et à faible contenu textuel. Ainsi, ce contexte rend la tâche de classification très difficile. C'est pourquoi, outre le modèle présenté précédemment, nous proposons d'évaluer la qualité d'autres descripteurs et finalement de les combiner. Nous effectuons dans un premier temps une sélection des descripteurs de la littérature les plus pertinents tel que décrit dans la section suivante.

5.4.2 Quelles approches combiner ?

L'objectif de cette section est de motiver le choix des descripteurs qui pourront être combinés.

5.4.2.1 Le choix des descripteurs pertinents de la littérature

Afin de sélectionner les descripteurs les plus pertinents, nous avons effectué un certain nombre d'expérimentations sur les corpus d'Itesoft. Ces expérimentations s'appuient sur les descripteurs précédemment décrits dans l'état de l'art relatif à la sélection de descripteurs (chapitre 2) : le mot, le n-gramme de mots, le n-gramme de caractères et les descripteurs proposés par le modèle décrit en section 5.2 extrayant les catégories lexicales. Nous avons, comme pour les expérimentations menées sur les corpus issus de blogs, utilisé une vectorialisation de type Salon, avec l'emploi de la fréquence des termes *tf* et la pondération *tf-idf*. Ces travaux nous ont permis d'identifier les descripteurs les plus adaptés pour notre tâche de classification. Les expérimentations qui seront présentées dans la section 5.4.5 ainsi que les résultats issus des travaux de la littérature nous ont amené à sélectionner trois méthodes :

- Les descripteurs de notre modèle.
- Les N-grammes de caractères.
- Le filtrage statistique.

Le choix s'est porté sur ces trois descripteurs pour les raisons suivantes.

• L'application de l'étiquetage grammatical (de notre modèle) a pour but de sélectionner les données respectant une catégorie ou un groupe de catégories lexicales données (nom, verbe, adjectif, nom-verbe, nom-adjectif, etc.). L'objectif principal de ce traitement est de ne conserver que les données ayant une information sémantique pertinente pour une tâche de classification. Notons cependant que nous n'utiliserons pas dans nos expérimentations la pondération de ces catégories.

• La représentation des données selon les N-grammes de caractères est motivée par la complexité des données que nous manipulons (données bruitées issues de la rétro-conversion d'OCR). En effet, l'utilisation des N-grammes de caractères est adaptée

dans le cadre des données issues d'OCR [Junker & Hoch, 1997].

• Le dernier processus proposé consiste à appliquer une mesure statistique afin d'attribuer un poids aux descripteurs. En leur attribuant un poids, nous favorisons les plus discriminants pour une classe particulière.

5.4.2.2 Dans quel ordre combiner ces approches

Afin d'obtenir de meilleurs résultats de classification, nous proposons de combiner les approches sélectionnées dans la section précédente. Cependant, une telle combinaison nécessite un ordre de sélection des descripteurs.
L'association d'étiquettes grammaticales peut seulement s'effectuer à partir des mots "au complet". En d'autres termes, elle n'est pas applicable sur les mots tronqués par les N-grammes de caractères. En effet, utiliser des catégories lexicales après l'extraction de N-grammes de mots serait incohérent linguistiquement et impossible avec des n-grammes de caractères, les mots n'étant plus identifiables. L'ordre établi consiste donc à appliquer un filtrage grammatical (sélection des mots selon leur étiquette tel que présenté dans notre modèle section 5.2) suivi par une représentation des N-grammes. Le fait de finir le traitement par une pondération statistique se justifie par la représentation de chaque document par un vecteur de K éléments (les éléments représentent les K N-grammes). Ces éléments ne sont pas tous discriminants, en leur attribuant un poids nous favorisons les plus significatifs pour caractériser une classe.

5.4.3 Description et discussions sur la combinaison des approches de sélection de descripteurs

Dans la section précédente, nous avons établi un ordre dans l'application des différents traitements. Dans cette section, nous allons détailler la manière de combiner nos approches.

Dans un premier temps, la sélection des mots avec des étiquettes grammaticales est effectuée. La sélection selon les étiquettes (Nom-Verbe) sur la phrase "`or le bijoux plaqué or a du charme.`", nous donne le résultat suivant : "`bijoux plaqué or a charme`". Notons que le filtrage grammatical permet de distinguer le mot "or" de type *conjonction de coordination* comparativement au *nom* "or".

Après ce premier traitement, nous représentons les mots extraits par les N-grammes de caractères. L'application de la représentation N-grammes de caractères nous donne trois

possibilités de représentation :

- La première représentation peut être considérée comme un sac de mots sélectionnés grammaticalement. L'application des N-grammes avec N=5 nous donne par exemple le résultat suivant :

  "_bijo, bijou, ijoux, joux_, oux_p, ux_pl, x_pla, _plaq, plaqu, laqué, aqué_, qué_o, ué_or, é_or_, _or_a, or_a_, r_a_c, _a_ch, a_cha, _cham, chamr, harme, arme_"

  Cette application est erronée car elle rajoute du bruit et des N-grammes inutiles, par exemple a_cha est un des N-grammes qui représente du bruit (N-gramme issu du fragment "a du charme" pour lequel le mot "du" a été supprimé). En effet, le fait d'éliminer des mots de la phrase initiale entraîne la construction de suites de mots non pertinents (et donc des N-grammes incorrects).

- Un deuxième type de représentation consiste à appliquer des N-grammes de caractères pour chacun des mots extraits séparément. Nous aurons comme résultat :

  "_bijo, bijou, ijoux, joux_, _plaq, plaqu, laqué, aqué_, _cham, chamr, harme, arme_"

  Cette représentation corrige les défauts causés par la précédente méthode. Elle n'introduit pas de bruit mais elle souffre de perte d'information notamment sur les mots courts. Par exemple, en appliquant les N-grammes de caractères avec $N \geq 5$ le nom "or" ne peut être identifié. Cette suppression occasionne une perte d'information.

- Les deux premières représentations ont donc des défauts majeurs liés à l'introduction de **bruit** (première méthode) et du **silence** (deuxième méthode). Pour cela nous avons introduit un **principe de frontière**. Celui-ci permet de remplacer les mots supprimés par une frontière. L'extraction des n-grammes est alors menée empiriquement entre des frontières et obliées. Cette méthode corrige le défaut de rajout du bruit causé lors de la première représentation. Il permet également de prendre en considération des groupes de mots (par exemple, "plaqué or"). Le résultat obtenu selon le principe de frontière est montré ci-dessous :

  "X bijoux plaqué or a X charme", le "X" représente la frontière.

  L'application de la méthode des 5-grammes donne le résultat ci-dessous :

  "_bijo, bijou, ijoux, joux_, oux_p, ux_pl, x_pla, _plaq, plaqu, laqué, aqué_, qué_o, ué_or, é_or_, _or_a, or_a_, _cham, chamr, harme, arme_ "

Après avoir présenté les descripteurs les plus pertinents pour notre tâche, avoir défini

l'ordre d'application des descripteurs et enfin après avoir discuté la manière de combiner de manière optimale les descripteurs, nous présentons dans la section suivante notre approche appelée HYBRED qui combine les descripteurs de notre modèle à ceux de la littérature.

5.4.4 Approche HYBRED

Dans cette section, nous présentons le principe que nous avons retenu pour notre système de représentation des données. Le principe général est résumé dans l'algorithme ci-dessous qui sera détaillé dans cette section.

Entrées : L'ensemble des textes constituant le corpus.
Sorties : Matrice.
forall *Documents* **do**
  Extraction des mots selon une étiquette grammaticale (a)
  Application du principe de frontière (b)
  Représentation des mots extraits selon les N-grammes de caractères (c)
  Attribution de poids selon la mesure *tf-idf* (d)
end

Algorithme 1 : HYBRED

5.4.4.1 Description d'HYBRED

Étape (a) : sélection selon une étiquette grammaticale
Une sélection des données selon une étiquette grammaticale propose de ne sélectionner que les termes appartenant à une ou plusieurs catégories lexicales données, comme les *noms* et les *verbes*.

Étape (b) : application du principe de frontière
Dans les travaux de [Bourigault, 1994], nous trouvons une application du principe de frontière. LEXTER, développé par D. Bourigault est un outil d'extraction de la terminologie. Il effectue une extraction de groupes nominaux (syntagmes nominaux) par repérage des marqueurs de frontières. Ces frontières sont déterminées linguistiquement (exemples de frontière : "préposition + adjectif possessif", "préposition + article indéfini", etc.). Les candidats termes, à savoir les groupes nominaux maximaux, sont extraits sur la base de leur position relative aux frontières.

Dans notre étude, les mots apportant peu d'informations (avec des étiquettes grammaticales moins pertinentes) sont remplacés par une frontière. L'objectif reste le même que

dans LEXTER. En effet, nous prenons en considération les groupes de mots pertinents situés entre les frontières. Cependant, la différence tient au fait que nos frontières sont les mots ayant des étiquettes grammaticales moins pertinentes pour les tâches de classifications (adverbe, préposition, etc.) et ne s'appuient pas sur des règles linguistiques comme dans LEXTER.

Étape (c) : représentation avec les N-grammes
Après avoir conservé les données appartenant à une étiquette grammaticale et appliqué le principe de frontière, vient l'étape de représentation avec les N-grammes de caractères. Il s'agit d'une fusion des N-grammes des différents fragments séparés par la frontière.

$$\texttt{Nbr-N-grammes} = \sum_{i \in \{ensemble \ \ des \ \ fragments\}} \texttt{N-grammes}(fragment_i)$$

Après avoir effectué la représentation avec les N-grammes de caractères, nous réalisons une étape de filtrage de N-grammes non pertinents. Cette étape consiste à supprimer les N-grammes peu fréquents et qui peuvent constituer du bruit (N-grammes $<$ seuil (fixé manuellement à 30)).

Étape (d) : pondération statistique
Enfin, de manière similaire aux très nombreux travaux de la littérature, nous avons appliqué une pondération statistique fondée sur le *tf-idf* afin de mettre en valeur les descripteurs discriminants. Le principe du *tf-idf* appliqué ici est décrit dans la section 2.2.3.1.

5.4.4.2 Exemple de l'application d'HYBRED

Cette section développe un exemple complet de l'approche HYBRED. Pour ce faire, nous considérons la phrase "`Il faut une infinie patience pour attendre toujours ce qui n'arrive jamais`".

(a) La sélection des données selon la combinaison NVA (Nom Verbe Adjectif) donnera comme résultat : "`faut infinie patience attendre arrive`".
(b) L'application du principe de frontière, nous donne :
"`X faut X infinie patience X attendre X arrive X`".
(c) La représentation sous N-grammes ou N=3 aura comme résultat :

Mot	N-grammes de caractères
[_faut_]	[_fa, fau, aut, ut_]
[_infinie patience_]	[_in, inf, nfi, fin, ini, nie, ie_, e_p, _pa, pat, ati, tie, ien, enc, nce, ce_]
[_attendre_]	[_at, att, tte, ten, end, ndr, dre, re_]
[_arrive_]	[_ar, arr, rri, riv, ive, ve_]

Ainsi, nous pouvons calculer la somme de tous les 3-grammes :
N-grammes("_faut_") + N-grammes("_infinie patience_") + N-grammes("_attendre_")
+ N-grammes("_arrive_").
Nous obtenons :
{_fa, fau, aut, ut_, _in, inf, nfi, fin, ini, nie, ie_, e_p, _pa, pat, ati, tie, ien, enc, nce, ce_, _at, att, tte, ten, end, ndr, dre, re_, _ar, arr, rri, riv, ive, ve_}
Enfin, les termes sont représentés vectoriellement en fonction de leurs fréquences d'apparition dans les documents puis pondérés par l'approche *tf-idf*.

5.4.5 Expérimentations

Dans cette section, nous présentons les différentes expérimentations que nous avons réalisées pour déterminer les descripteurs pertinents et tester la pertinence de notre approche. Tout d'abord, nous présentons le protocole expérimental sur lequel nous nous appuyons puis les résultats obtenus lors de la classification.

5.4.5.1 Protocole expérimental

Pour évaluer la pertinence des différents descripteurs, nous avons utilisé les algorithmes de classification avec apprentissage supervisé suivant : les k plus proches voisins (k-ppv), un classificateur fondé sur les machines à support vectoriel (SVM) et un algorithme probabiliste (NaiveBayes).

Notons que nous n'avons pas implémenté ces algorithmes, mais nous avons utilisé le logiciel "Weka[24]" [Witten *et al.*, 1999]. Les paramètres que nous avons utilisés avec les algorithmes de classification sont ceux définis par défaut dans Weka (k-ppv avec $k = 1$ et les SVM avec un *noyau polynomial du second degrée*). Afin de déterminer quels descripteurs sont les plus pertinents, nous comparons les résultats obtenus avec les trois algorithmes. Les résultats de la classification sont présentés selon la précision (accuracy)

[24] www.cs.waikato.ac.nz/ml/weka

des algorithmes définie comme suit.

$$Précision = \frac{nombre\ de\ documents\ correctement\ classés}{ensemble\ des\ documents\ classés}$$

Par ailleurs, cette mesure de précision a été calculée après l'application d'une 10-validation croisée.
Nous rappelons que les documents sont représentés selon une approche vectorielle. Nous effectuons une suppression des informations rajoutées par OCR comme les coordonnées de chaque mot dans les documents. En outre, nous avons expérimenté le calcul de fréquence de chaque descripteur dans le corpus (*tf*) ainsi que la pondération avec la mesure *tf-idf*.

Pour évaluer les performances de l'utilisation de nos différents descripteurs pour une tâche de classification, nous avons utilisé trois corpus de test :

Corpus A
Le premier jeu d'essai comporte 250 fichiers en provenance d'ITESOFT se répartissant en 18 catégories qui représentent des frais, des factures, des attestations (transport, hospitalière, médecin, etc). Ce corpus en français liste des fichiers images qui peuvent être :

- Des images sur des documents manuscrits.
- Des formulaires.
- Des documents imprimés ou dactylographiés en fichiers de texte.

Les textes contiennent peu de phrases bien formulées en langage naturel. Notons que certaines classes contiennent plus de documents que d'autres (distribution non équilibrée entre les classes). La grande complexité des données est due au faible contenu des documents, mais surtout à un ensemble d'apprentissage de faible taille.

Corpus B
Le deuxième corpus utilisé pour les expérimentations provient aussi d'ITESOFT. Il comporte 2000 fichiers répartis dans 24 catégories. Les documents en français sont issus d'une rétro-conversion d'OCR.
Les catégories représentent des bulletins, des certificats, des avis d'impôt, etc. La principale caractéristique du corpus est la diversité des documents et le faible contenu de ces derniers. Mais la grande difficulté est la distribution des documents. Nous avons des catégories qui ne contiennent que 1 à 5 documents en comparaison avec d'autres qui contiennent plus de 250 documents (répartition très hétérogène).

Corpus C

Le dernier corpus est une collection de dépêches (en français) obtenue manuellement à partir de sites d'actualité sur internet. Il comporte 65 documents répartis en 5 catégories. L'idée de l'utilisation de ce dernier est de pouvoir comparer les performances du système avec un corpus plus riche et moins bruité.

Les catégories sont des articles de presse sur le Président Sarkozy traitant du thème de vie privée, la diffusion de vidéo de TF1 sur les sites Daylimotion et Youtube, les élections à la ville de Paris, du sport (ski) et la dernière classe représente les actions de l'association "des enfants Don Quichotte" sur le problème des personnes sans domicile fixe. Les documents contiennent peu de bruit et sont d'une taille conséquente pour l'apprentissage.

Le tableau 5.4 résume les principales caractéristiques des trois corpus exploités.

	Corpus A	Corpus B	Corpus C
Nombre de documents	250	2000	65
Nombre de catégories	18	24	5
Taille (en Mo)	0.434	2.12	0.211
Nombre de mots	73855	260752	31711
Type de textes	rétro-conversion OCR	rétro-conversion OCR	Articles journaux

TAB. 5.4 – Résumé des principales caractéristiques des trois corpus.

5.4.5.2 Résultats expérimentaux

Dans cette section, nous allons présenter les expérimentations selon les différents descripteurs et l'approche HYBRED. Notons que nous avons appliqué un filtrage des N-grammes peu fréquents, en fixant un seuil à 30. En dessous de ce seuil, les N-grammes ne sont pas pris en compte. Notons que nous présenterons uniquement les résultats obtenus pour le corpus B. Les résultats obtenus avec les autres corpus sont présentés en annexe. En effet, le corpus B est le plus représentatif en termes de taille et de difficulté de traitement (données hétérogènes et bruitées). En outre, une des difficultés du traitement du corpus B tient au fait qu'en moyenne, chaque document est moins riche en termes de nombre de mots (130 mots par document pour le corpus B) comparativement aux corpus A (295 mots par document) et C (488 mots par document).

Le tableau 5.5 présente les résultats obtenus avec les différents descripteurs en appliquant une validation croisée (10-CV). Nous observons que les meilleurs résultats sont

Algorithmes	k-ppv		SVM		NB	
mesure	Fréquentiel	*tf-idf*	Fréquentiel	*tf-idf*	Fréquentiel	*tf-idf*
Mot	91.1	**91.1**	**95.8**	**95.8**	**94.1**	**93.8**
2-mots	**92.2**	90.9	93.7	93.7	91.9	92.2
3-mots	90.5	90.5	90.1	89.9	82.8	86.1
2-caractères	73.7	72.6	89.6	88.2	74.3	58.5
3-caractères	85.7	86.0	**96.5**	**96.8**	**93.4**	**91.9**
4-caractères	**95.0**	**96.1**	96.0	96.3	93.1	90.7
5-caractères	91.4	92.5	96.2	95.6	92.0	90.8
Lemme	92.3	93.8	95.4	**95.5**	93.7	94.4
N	91.1	**93.0**	95.6	95.1	93.6	94.6
V	88.2	87.5	88.4	87.8	85.2	84.9
NV	92.4	92.7	95.5	95.5	**94.1**	94.3
NVA	**93.3**	92.6	**95.6**	**95.8**	**94.1**	94.5
NA	92.8	92.4	**95.6**	95.4	93.9	**94.8**
VA	92.0	91.4	93.7	93.7	91.7	91.4

TAB. 5.5 – Résultats du corpus B avec les différents descripteurs (précision).

obtenus avec l'algorithme SVM par rapport à k-ppv et à Naïve Bayes. La représentation avec les N-grammes de caractères se comporte bien avec les trois corpus. En général, nous remarquons que les résultats se détériorent significativement quand N=2. L'application de l'analyse morpho-syntaxique ou la sélection selon une étiquette grammaticale peut, dans certains cas, se révéler efficace dans son utilisation et sa capacité à ne sélectionner que les données discriminantes (ce qui réduit significativement l'espace de représentation comme nous allons le montrer dans la section 5.3). Nos résultats montrent que les combinaisons selon les NV (**N**om**V**erbe), NVA (**N**om**V**erbe**A**djectif) et NA (**N**om**A**djectif) sont les plus pertinentes parmi les combinaisons testées.

Dans le corpus A, les résultats (présentés en Annexe) sont en général meilleurs que les résultats obtenus avec le corpus B. Ceci est dû à la complexité du corpus B (corpus bruité possédant en moyenne moins de mots par document). Enfin, avec le corpus C, nous obtenons les meilleurs résultats. Cela s'explique par le corpus en lui-même (textes journalistiques), qui est peu bruité et chaque document est assez riche en terme de nombre de mots.

Le tableau 5.6 présente les résultats obtenus en appliquant l'approche HYBRED avec le corpus B. Les résultats des corpus A et C sont donnés en Annexe. Dans tous les cas, nous obtenons les meilleures performances avec l'algorithme **SVM** par rapport aux k-ppv et Naïve Bayes. Nous remarquons que la sélection NV (Nom Verbe) associée

Algorithme k-ppv (*tf-idf*)						
Descripteurs	N	V	NV	NVA	NA	VA
2-caractères	74.8	85.5	74.6	72.7	74.6	77.3
3-caractères	85.0	85.5	95.8	87.0	86.3	86.6
4-caractères	85.0	86.5	**96.7**	92.6	92.1	**90.2**
5-caractères	**91.8**	**88.4**	93.0	**93.4**	**92.4**	90.0
Algorithme SVM (*tf-idf*)						
Descripteurs	N	V	NV	NVA	NA	VA
2-caractères	89.5	89.9	87.4	86.43	89.0	88.1
3-caractères	96.4	**94.2**	96.6	**96.9**	**96.8**	**96.3**
4-caractères	**96.5**	93.8	**98.0**	96.8	96.7	95.8
5-caractères	96.4	93.2	96.8	96.8	96.7	95.2
Algorithme NB (*tf-idf*)						
Descripteurs	N	V	NV	NVA	NA	VA
2-caractères	61.4	**92.6**	60.7	59.4	73.5	63.6
3-caractères	61.4	88.3	60.7	**92.2**	73.5	91.4
4-caractères	**92.6**	88.3	**96.9**	**92.2**	**92.7**	**91.9**
5-caractères	**92.6**	86.9	93.1	92.1	92.4	91.5

Tab. 5.6 – Précision obtenue avec l'approche HYBRED pour le corpus B

aux 4-grammes donne des résultats très satisfaisants sur le corpus B. Nous remarquons globalement ces mêmes comportements sur les corpus A et C.

5.4.5.3 Synthèse

Comme nous avons pu le constater lors des expérimentations, la méthode HYBRED fondée sur la combinaison des descripteurs a tendance à améliorer les performances des classificateurs par rapport à l'utilisation de chaque descripteur séparément. Nous pouvons observer cette amélioration sur les trois corpus, notamment sur le corpus B qui est le plus complexe à classifier. Le tableau 5.7 présente une comparaison entre l'approche proposée avec une combinaison NV (**N**om**V**erbe) associée à une représentation 4-grammes de caractères et les différents descripteurs.

En général, des améliorations de la précision ont été obtenues en appliquant l'approche HYBRED. Ainsi, le tableau 4 montre qu'HYBRED améliore toujours les résultats (de manière plus ou moins significative selon les corpus) avec l'algorithme SVM. Ceci est particulièrement intéressant, car cet algorithme est celui qui a le meilleur comportement, ce que nos expérimentations ont confirmé à partir des trois corpus. Par ailleurs, cette amélioration est particulièrement importante avec le corpus le plus représentatif et le plus complexe (corpus B).

	Corpus A			Corpus B			Corpus C		
algorithmes	k-ppv	SVM	NB	k-ppv	SVM	NB	k-ppv	SVM	NB
mot	96.5	97.5	**96.7**	91.1	95.8	93.8	95.3	98.4	**98.4**
3-caractères	94.7	97.9	93.5	86.0	96.8	91.9	98.4	**100**	82.8
4-caractères	**97.5**	98.3	94.3	96.1	96.3	90.7	98.4	**100**	82.8
5-caractères	96.7	98.3	95.1	92.5	95.6	90.8	**100**	**100**	85.9
NV	95.9	98.0	**96.7**	92.7	95.5	94.3	96.4	98.0	87.5
NVA	95.5	98.0	**96.7**	92.6	95.8	94.5	96.8	98.2	79.6
NA	95.1	98.0	**96.7**	92.4	95.4	94.8	98.0	98.0	85.9
HYBRED	96.8	**98.4**	93.6	**96.7**	**98.0**	**96.9**	**100**	**100**	79.6

TAB. 5.7 – Tableau de comparaison entre HYBRED et des méthodes "classiques"

Dans le tableau 5.8, nous présentons une comparaison de l'espace de représentation (taille) avec et sans appliquer l'approche HYBRED. Nous remarquons que l'application d'HYBRED réduit de manière significative l'espace de représentation. Les résultats obtenus selon l'approche HYBRED sont donnés avec la combinaison NV (Nom Verbe) suivie d'une représentation 4-grammes de caractères.

Espace de représentation	Sans application d'HYBRED		Après application d'HYBRED
	Mot	N-grammes (N=4)	NV + N-grammes (N=4)
Corpus A	12307	2087	1603
Corpus B	37837	4485	3294
Corpus C	5417	1274	876

TAB. 5.8 – Tableau de comparaison de l'espace de représentation avec et sans HYBRED

Nous avons montré au cours de ces expérimentations que l'approche proposée en section 5.2 était adaptée aux données syntaxiquement pauvres. Cependant, les résultats de cette section ont mis en évidence les limites d'une telle approche avec des données bruitées. Nous avons ainsi proposé une approche hybride combinant des descripteurs adaptés aux données bruités et ceux de notre modèle. Nous avons obtenu des résultats encourageants avec cette méthode. Nous cherchons à présent à confronter notre approche aux données dépourvues de syntaxes tel que présenté dans la section suivante.

5.5 Traitement des données liées aux Ressources Humaines

5.5.1 Contexte

Nous présentons dans cette section des travaux ayant été menés avec en collaboration avec le Laboratoire en Informatique d'Avignon (LIA). Nous avons ainsi collaboré avec Marc El Beze, Juan Manuel Torres Moreno et Rémy Kessler. Les travaux présentés dans cette section sont une étape de l'application du système E-Gen, objet de la thèse de Rémy Kessler, effectués dans le cadre d'un financement CIFRE avec la société Aktor Intéractive (www.aktor.fr). Le système E-Gen est proposé pour répondre à l'émergence d'un grand nombre de sites d'emplois [Bizer & Rainer, 2005] et d'un marché du recrutement en ligne en expansion significative (août 2003 : 177 000 offres, mai 2008 : 500 000 offres)[25]. Ce nombre conséquent d'informations pose le problème de la rapidité et l'efficacité de leur traitement, ce dernier devant répondre aux demandes des entreprises [Bourse *et al.*, 2004],[Rafter *et al.*, 2000]. Ainsi il est nécessaire de traiter cette masse de documents de manière automatique ou assistée. Il est alors proposé le système E-Gen, pouvant être décomposé en trois modules dont le rôle de chacun étant :

1. Extraire de l'information à partir de corpus de courriels provenant d'offres d'emplois extraites de la base de données d'Aktor [Kessler *et al.*, 2007].
2. Analyser les courriels de réponses des candidats pour distinguer lettre de motivation (LM) et Curriculum Vitæ(CV) [Kessler *et al.*, 2008b].
3. Analyser et calculer un classement de pertinence des candidatures (LM et CV).

Notre collaboration a porté sur le dernier module en proposant d'utiliser notre modèle fondé sur les catégories lexicales afin d'améliorer les résultats du classement automatique des candidatures. Ainsi, nous disposons d'un certain nombre d'offres d'emplois. À chacune d'elles sont associées des candidatures sous forme de CV et de lettres de motivations. Notre tâche est alors de proposer une sélection des meilleurs candidats à chaque offre d'emploi, afin de limiter la tâche des recruteurs.

5.5.2 Méthode de classement automatique des candidats

Le protocole expérimental suivi diffère de ceux déjà présentés dans ce chapitre car nous n'avons pas effectué de classification automatique de textes. Une telle tâche a en effet été menée précédemment avec ces données mais n'a pas révélé de résultats probant.

[25] Site d'emploi www.keljob.com

Ainsi, nous proposons d'effectuer une comparaison des candidatures avec les offres d'emploi. Nous obtenons alors un classement des candidatures en fonction de leurs pertinences vis-à-vis des offres d'emploi.
Nous disposons d'un corpus fourni par la société *Aktor* contenant 14 offres d'emploi avec des thématiques différentes (emplois en comptabilité, commercial, informatique, etc) associées aux réponses des candidats (1917 candidatures). Chaque candidature est identifiée comme pertinente ou non pertinente par un expert.
Nous construisons à partir de ce corpus une matrice de co-occurrences termes/documents afin d'obtenir une représentation vectorielle des offres d'emploi et des candidatures. Nous pouvons finalement mesurer la proximité des offres avec les candidatures en utilisant trois mesures de proximité : le *cosinus* [Salton, 1971], la mesure de *Minkowski* [Sokal, 1977] et la mesure *Okabis* [Bellot & El-Bèze, 2001]. Notons que ces mesures, plus complexes, ont le même comportement que la mesure cosinus [Kessler *et al.*, 2008a]. Après avoir calculé la proximité des candidatures aux offres d'emploi sur la base des différentes mesures, nous pouvons établir un classement des candidatures.
Afin de mesurer la qualité de ce classement, nous employons les *Courbes ROC* (Receiver Operating characteristic) [Ferri *et al.*, 2002], fréquemment utilisée afin de mesurer la qualité des fonctions de rang. La méthode des courbes ROC met en relation dans un graphique le taux de faux positifs (c'est-à-dire les candidatures non pertinentes) en abscisse et le taux de vrais positifs (c'est-à-dire les candidatures pertinentes) en ordonnée. La surface sous la courbe ROC ainsi créée est appelée AUC (Area Under the Curve). Nous nous sommes appuyés sur cette métrique dans nos expérimentations. Notons que cette approche sera appliquée également dans le chapitre 7, dans lequel nous détaillerons davantage le principe des courbes ROC.

5.5.3 Expérimentations

Nous résumons ici les résultats expérimentaux obtenus en utilisant différents descripteurs. Notons que nous ne détaillerons pas dans cette section l'ensemble des résultats expérimentaux obtenus, ces derniers étant décrits dans [Kessler *et al.*, 2009] et [Kessler *et al.*, 2008a]. Nous avons évalué le terme, sous sa forme fléchie ou lemmatisée. Ce dernier a été pondéré suivant sa fréquence ou bien son *tf-idf*. Nous avons également utilisé les descripteurs fondés sur les catégories lexicales avec diverses pondérations sur les noms verbes et adjectifs. Les résultats des AUC en utilisant ces différents descripteurs sont obtenus en calculant une moyenne des AUC obtenus avec chaque mesure de similarité. Chaque descripteur obtient des résultats similaires autour de 0,64. Ainsi, le modèle présenté dans ce chapitre ne semble pas adapté aux données syntaxiquement pauvres comme des CV. Notons par ailleurs que les n-grammes de caractères utilisés dans le

modèle HYBRED ont obtenu des valeurs d'AUC plus faibles au cours de nos expérimentations. Nous n'avons ainsi pas évalué l'approche HYBRED avec ces données, la jugeant non adaptée.
Nous nous sommes alors intéressés à la structure des CV et des lettres de motivation (LM). Nous cherchons alors à déterminer où se situe l'information contenue dans les CV et les LM. Nos expérimentations ont montré, en effectuant un découpage par tiers des CV et LM, que l'information était contenue dans le second tiers du CV et dans le premier tiers des LM. En effet, avec uniquement le second tiers du CV, les AUC obtenues sont relativement proches de celles obtenues avec la totalité des candidatures. Notons que ces travaux sont en cours et que d'autres expérimentations seront prochainement effectuées.
Une dernière méthode a été utilisée afin d'améliorer les AUC précédemment obtenues, le modèle présenté dans ce chapitre ne donnant pas de résultats satisfaisants. La méthode testée, nommée *Relevance Feedback* ou retour de pertinence proposée par [Spärck-Jones, 1970] est une approche classique de reformulation de requête afin d'améliorer les résultats obtenus au préalable. Par exemple, un ensemble de résultats faisant suite à une requête est analysé par un utilisateur, qui va reformuler sa requête en prenant compte des résultats. Notons que cette approche a déjà été employée dans le domaine des Ressources Humaines. En effet, [Rafter *et al.*, 2000] proposent un système de Relevance Feedback afin de guider l'internaute dans sa recherche d'emploi à partir d'informations extraites du site d'emploi *JobFinder*[26]. Dans notre cas, cette approche va permettre de prendre en compte les choix du recruteur lors de l'évaluation de quelques candidatures. En d'autres termes, cette approche permet d'introduire les connaissances de l'expert dans le modèle de sélection de candidatures. Cette méthode permet en quelque sorte d'effectuer un apprentissage sur les données positives. Ainsi, nous effectuons un tirage aléatoire de quelques candidatures (de une à six dans nos expérimentations) parmi l'ensemble des candidatures étiquetées comme pertinentes. Celles-ci sont finalement ajoutées à la Mission (description de l'offre), produisant un espace vectoriel enrichi par les termes (descripteurs) jugés pertinents par le recruteur. L'approche de "Relevance Feedback" a permis d'améliorer nos résultats d'AUC [Kessler *et al.*, 2009].

5.6 Synthèse

Nous avons présenté dans ce chapitre une méthode d'extraction de descripteurs pertinents adaptée aux *données complexes*. Ce type de données peut être défini comme des données textuelles écrites dans une langue naturelle, mais ne respectant pas ou peu une grammaire décrivant cette langue. La méthode d'extraction de descripteurs proposée dans

[26] jobfinder.com

ce chapitre a été mise en place afin de répondre aux contraintes du modèle SelDe. Rappelons en effet que SelDe ne peut être appliqué sur des données complexes car notre modèle est fondé sur les relations syntaxiques contenues dans un corpus. Le modèle de sélection de descripteurs de ce chapitre propose l'extraction de termes appartenant à des catégories lexicales préalablement fixées. Ces termes sont extraits en employant un étiqueteur grammatical qui s'appuie sur une approche statistique. Ainsi, notre modèle est assez robuste aux données complexes.
Nous avons évalué la qualité de ce modèle dans diverses expérimentations sur différents types de données complexes. Nous avons expérimenté les données suivantes.

1. Données syntaxiquement mal formulées et mal orthographiées (corpus formé d'articles de blogs).
2. Données bruitées ou incomplètes (corpus formé de documents numérisés par reconnaissance OCR).
3. Données dépourvues de syntaxe (corpus de Ressources Humaines contenant des CV).

Notons que pour les deux premières expérimentations, notre tâche fut de classer automatiquement des articles dans des catégories définies en employant une approche avec apprentissage supervisé. La dernière tâche quant à elle, consiste à sélectionner des candidats (sur la base de CV et de lettres de motivations) pertinents par le biais de mesures de proximité, sans apprentissage.

• Le modèle proposé dans ce chapitre a donné des résultats encourageants lors de nos premières expérimentations. Ainsi, il nous semble adapté aux données syntaxiquement mal formulées.

• Les résultats de la deuxième expérimentation avec le modèle se sont révélés moins pertinents. Alors, une approche nommée Hybred a été définie consistant à combiner l'approche des catégories lexicales avec des descripteurs de type n-grammes de caractères. Cette approche a donné des résultats encourageants. Ainsi, nous proposons avec des données bruitées et/ou incomplètes d'utiliser l'approche Hybred.

• La dernière expérimentation fut l'occasion de proposer une nouvelle approche, le *Relevance Feedback*, les autres approches ne s'étant pas révélées probantes. Bien qu'améliorant les résultats des autres méthodes, le *Relevance Feedback* produit des résultats assez mitigés, nous invitant à travailler sur de nouvelles pistes, consistant à intégrer des connaissances sur les CV.

• Notons pour finir que nous n'avons pas appliqué l'approche LSA au cours des expérimentations menées sur les données textuelles complexes. En effet, les contraintes imposées par les applications industrielles empêchent l'utilisation d'approches trop coûteuses en termes de temps d'exécution.

Après avoir expérimenté la sélection de descripteurs adaptés aux données textuelles complexes, nous revenons dans la section suivante sur la sélection de descripteurs fondée sur la syntaxe. Rappelons en effet que le modèle SelDe ne traite pas les objets complémentaires entre verbes, fournissant une information supplémentaire. Le chapitre suivant propose de traiter ces objets en présentant un nouveau modèle.

Chapitre 6

SelDeF : la sélection de descripteurs avec filtrage

Sommaire

Nous présentons dans ce chapitre les modèles théoriques de diverses approches ayant été expérimentées dans le chapitre 7. Ces travaux ont été publiés dans [3 - MAW'09], [5 - ECIR'09], [12 - TOTh'09], [14 - EGC'09] et [17 - EvalECD'09].

6.1 Vers un nouveau modèle

Après avoir présenté dans la section 3.4 un modèle de sélection de descripteurs pertinents en se fondant sur l'information syntaxique donnée par un corpus, nous avons appliqué ce modèle à la classification automatique de données textuelles. Notre approche d'enrichissement de corpus à partir de ces descripteurs est nommée ExpLSA. L'utilisation des différents paramètres proposés dans le modèle SelDe ont permis de mettre en valeur celui-ci face au modèle initial proposé par *D. Faure*[27]. Cependant, l'approche proposée dans la section 3.4 utilisant SelDe n'a pas permis d'améliorer dans tous les cas les résultats de classification, qu'il s'agisse de classification de textes ou de classification conceptuelle.

Par ailleurs, les résultats obtenus avec la classification conceptuelle sont d'une manière

[27] Ce modèle est noté Asium dans les expérimentations du chapitre 4

générale assez décevants en utilisant l'approche LSA. L'approche ExpLSA permet cependant d'améliorer les résultats dans certaines configurations. Cela signifie qu'une approche de type "statistique" comme LSA peut être pertinente mais pas nécessairement adaptée à la tâche de classification conceptuelle.
Nous employons en effet dans le chapitre 4 un paradigme visant à regrouper des termes via l'approche statistique LSA (que nous pouvons enrichir avec ExpLSA mais le paradigme reste le même). Ces termes sont alors regroupés afin de former des concepts. Cette approche est néanmoins discutable au vue des résultats obtenus lors de nos expérimentations. Une hypothèse que nous formulons est le manque d'utilisation d'informations syntaxiques lors de la sélection et du regroupement de ces termes. Bien que l'approche ExpLSA ait pour objectif de combler ces lacunes, la sélection des termes servant à construire une classification conceptuelle est toujours menée par le biais d'une approche statistique. Celle-ci comprend les étapes de la vectorialisation des phrases du corpus (enrichis ou non) via LSA puis la classification conceptuelle proprement dite est effectuée par des algorithmes statistiques usuels tels que les k plus proches voisins.
Nous présentons dans ce chapitre un nouveau modèle de sélection de descripteurs : SelDeF, pour **Sél**ection de **De**scripeurs avec **F**iltrage. Comme nous le montrerons dans les chapitres 6 et 7, le paradigme de ce modèle est plus adapté à la tâche de classification conceptuelle.

6.2 SelDeF

6.2.1 Description générale du modèle

Le second modèle SelDeF s'appuie sur SelDe pour la sélection de descripteurs. Les étapes communes entre les deux modèles sont rappelées ci-dessous :

1) Extraction des relations syntaxiques d'un corpus d'origine (en utilisant l'analyseur syntaxique Sygfran).
2) Mesure de la proximité sémantique des verbes des relations syntaxiques extraites (en utilisant la mesure de proximité d'Asium).
3) Sélection des couples de verbes jugés proches (couples de verbes partageant un certain nombre d'objets en commun).
4) Regroupement des objets des verbes jugés proches.
5) Distinction des objets communs des complémentaires.

L'ensemble de ces points sont représentés par l'étiquette "SelDe" dans la figure 6.1, synthétisant SelDe. La valeur ajoutée de SelDeF par rapport au précédent modèle

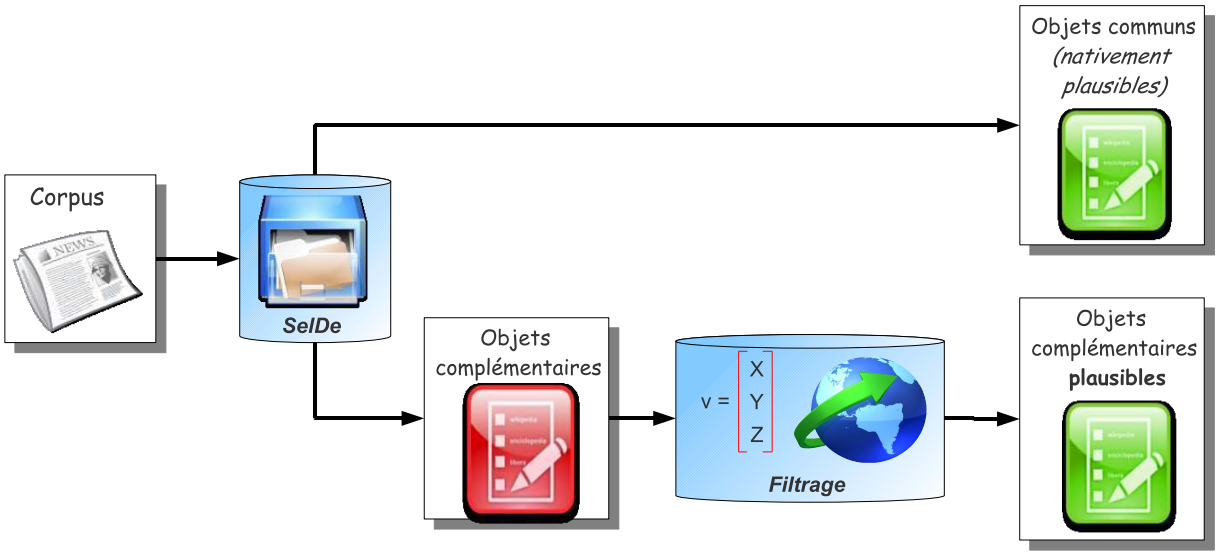

FIG. 6.1 – Le modèle SelDeF

est le filtrage des objets complémentaires, sur lequel nous reviendrons dans la section suivante. Ainsi, en sortie du modèle, nous obtenons les objets communs (nativement de qualité car originalement présents dans le corpus) et les objets complémentaires filtrés par différentes approches qui vont définir les descripteurs pertinents.
La section suivante montre l'intérêt du modèle SELDEF en motivant la nécessité de filtrer les objets complémentaires. Nous rappelons dans un premier temps comment distinguer ces objets par rapport aux objets communs de verbes.

6.2.2 Pourquoi un second modèle ?

En se référant à la figure 6.2, rappelons qu'une relation induite est formée d'un verbe et d'un objet complémentaire. Par exemple sur cette figure, une relation induite peut être la relation "*consommer fruit*". Les relations syntaxiques ainsi formées ne sont pas nativement présentes dans le texte et bien qu'apportant une information nouvelle, elles peuvent introduire une quantité non négligeable de bruit. Par exemple la relation induite formée par les termes "*manger essence*", extraite de la figure 6.2, n'est pas d'un point de vue pragmatique "acceptable". Dans le système d'ASIUM original [Faure & Nedellec, 1999], une sélection manuelle des objets complémentaires est effectuée. Une telle tâche peut se révéler trop coûteuse, trouvant ses limites avec un nombre de relations induites trop important, rendant l'expertise très fastidieuse.
Le modèle SELDEF permet quant à lui d'utiliser outre les objets communs des relations syntaxiques d'un corpus, les objets complémentaires (contrairement à SELDE).
Pour cela, il est proposé un filtrage automatique de ces objets en termes de pertinence.

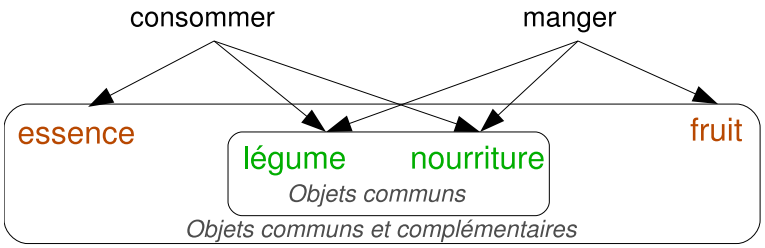

FIG. 6.2 – Objets communs et complémentaires des verbes “consommer” et “manger”.

Un objet complémentaire pertinent est ainsi défini comme *un objet, qui, associé au verbe avec lequel il forme une relation induite, produit une relation syntaxique plausible et acceptable d'un point de vue pragmatique.*
Nous proposons deux approches afin de filtrer les objets complémentaires. Le principe est de produire une liste ordonnée en termes de cohérence de relations syntaxiques induites. Une première approche considère une relation syntaxique comme une combinaison de différents concepts en se fondant sur un thésaurus [Larousse, 1992]. Ce thésaurus propose une indexation d'un nombre conséquent de termes de la langue française en définissant chacun d'eux comme une combinaison de concepts d'ordres généraux. Une telle représentation se fondant sur les concepts du thésaurus est nommée “approche des *vecteurs sémantiques*”, qui sont un type de vecteurs d'idées présentés en section 2.2.2.3. La seconde approche est fondée sur les ressources du Web. Elle utilise le nombre de résultats retournés par un moteur de recherche et différentes mesures statistiques afin de former une mesure de “popularité” des relations syntaxiques. Nous combinerons finalement ces deux approches. Nous présentons dans la section suivante le détail des méthodes proposées dans cette thèse.

6.3 Le filtrage des objets complémentaires

6.3.1 Les vecteurs sémantiques

Les vecteurs sémantiques ont été introduits par [Chauché, 1990]. Ils définissent une représentation de documents textuels pouvant être de granularité différente comme un terme, une phrase, un texte, etc. dans un même espace vectoriel. L'originalité d'une telle approche se traduit par l'utilisation d'un thésaurus comme base d'espace vectoriel, et permet la représentation dans une même base de corpus de différents domaines.

Cette section est organisée de la manière suivante. Nous présentons dans un premier temps certains travaux relatifs à l'emploi de vecteurs se fondant sur des thésaurus.

Puis, nous décrirons de quelle manière un vecteur sémantique est obtenu. Alors, nous définirons deux approches utilisant les vecteurs sémantiques afin de mesurer la plausibilité d'une relation sémantique induite. Finalement, nous montrerons plus précisément comment calculer la proximité de deux vecteurs sémantiques.

6.3.1.1 Travaux relatifs aux vecteurs fondés sur les thésaurus

[Wilks, 1998] discute du fait que les différents descripteurs d'un thésaurus comme le Roget peuvent être bénéfiques pour des tâches relatives au TAL.
Des approches dites *à la Roget* sont employées dans divers domaines du TAL, comme dans la désambiguïsation sémantique [Yarowsky, 1992], la recherche d'information [Boyd *et al.*, 1993], la cohésion textuelle [Morris & Hirst, 1991], ou comme mesure de similarité entres termes [McHale, 1998], [Jarmasz & Szpakowicz, 2003].
[Jarmasz & Szpakowicz, 2003] utilisent la taxonomie de la structure du thésaurus Roget pour déterminer la proximité sémantique entre deux termes. Ils obtiennent des résultats de bonne qualité pour les tests du TOEFL, ESL et Reader's Digest. Notre approche utilise quant à elle une approche *à la Roget* dans un contexte différent. Nos deux approches présentées dans la section 6.3.1.3 proposent d'utiliser une méthode *à la Roget* afin de mesurer la qualité des constituants d'un verbe donné. Nous présentons au préalable dans la section suivante comment sont obtenus les vecteurs sémantiques.

6.3.1.2 La représentation vectorielle

Nous avons choisi de représenter les mots et relations syntaxiques avec des vecteurs sémantiques afin d'ordonner en termes de qualité les relations syntaxiques induites d'un corpus. Nous présentons dans cette section comment de tels vecteurs sont calculés.

Les vecteurs sémantiques de mots
La base vectorielle représentant l'espace dans lequel chaque mot provenant des phrases d'un corpus est définie par 873 concepts décrits dans le thésaurus de la langue française Larousse, une version française du Roget (qui définissait 1 043 concepts pour l'anglais, réduit plus récemment à 1 000). Cette base vectorielle peut être vue comme une "*ontologie conceptuelle*" qui référence tous les mots du dictionnaire (dictionnaire Larousse). Actuellement, plus de 60 000 mots sont indexés par cette ontologie et leurs vecteurs sémantiques associés sont ainsi définis. Pour chaque terme sont indexés un ou plusieurs champs de l'ontologie (autrement appelés concepts). Par exemple le verbe "*consommer*" est relatif aux concepts de "*fin*[28]*, nutrition, accomplissement, usage, dépense et repas*". Ainsi, le vecteur

[28]Notons ici que bien qu'il prête à confusion dans un contexte de nutrition, ce concept est bien la "fin", et non pas la "faim". Ce concept "fin" doit être vu dans le sens de "achever", qui signifie par exemple

sémantique résultant est composé de zéros, excepté pour les concepts actifs tel qu'illustré dans la figure 6.3. Les concepts de l'ontologie (la base vectorielle) sont numérotés (de 1 à

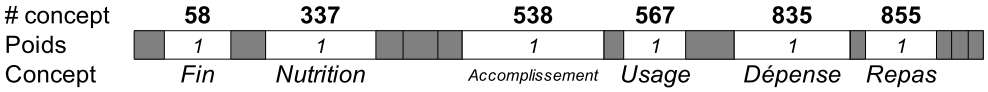

FIG. 6.3 – Vecteur sémantique du verbe "consommer".

873). La composante du vecteur se voit attribuer la valeur "1" si elle active un concept de l'ontologie, 0 sinon. De tels vecteurs ainsi formés peuvent être qualifiés d'inertes. En effet, toutes les significations possibles d'un mot sont évoquées mais les concepts actifs ne sont pas différenciés en terme d'intensité.

Les vecteurs sémantiques de phrases

Un vecteur sémantique de phrase est calculé en effectuant une combinaison linéaire de vecteurs de *groupes de mots*. À leurs tours, les vecteurs sémantiques de groupes de mots sont des combinaisons linéaires de vecteurs de *mots* ou de *sous-groupes de mots*.

Deux questions se posent alors :

– Comment sont calculés les vecteurs de mots ?

– Comment sont attribués les poids aux groupes de mots ou bien aux mots lors de la combinaison linéaire ?

Les vecteurs de mots sont représentés tel que décrit dans la section précédente en fonction des concepts définis dans le thésaurus Larousse.

Les différents poids sont quant à eux associés aux vecteurs de mots ou bien de groupes de mots. Ces poids sont calculés en fonction du rôle syntaxique du mot (ou du groupe de mots) dans son contexte (un groupe de mots ou une phrase). C'est pourquoi le groupe de mots est au préalable analysé syntaxiquement avec SYGFRAN puis transformé en arbre de constituants. Les poids des constituants gouvernants sont plus importants que ceux des constituants gouvernés. Ces pondérations permettent de donner plus d'importance d'un point de vue sémantique aux gouverneurs d'un groupe de mots. Par exemple dans le groupe de mots "*consommer de la nourriture*", le verbe "*consommer*" aura un poids plus important que "*nourriture*" (le *verbe* gouvernant l'*objet*). Ces poids sont définis comme des puissances de 2, en commençant par 2^0 pour les feuilles de l'arbre syntaxique, qui sont les constituants les plus dépendants, jusqu'à 2^p. p représente le rang du composant "le plus gouvernant" de l'arbre résultant de l'analyse syntaxique. Ainsi, dans l'exemple précédent, *consommer* aura un poids de 2, et *nourriture* un poids de 1.

Finalement, la représentation formelle du calcul d'un vecteur sémantique d'un groupe

pour le groupe verbal "consommer une pomme" que la pomme est totalement consommée.

de mots est la suivante. Soit γ, un groupe de mots analysé par SYGFRAN afin d'en connaître la structure syntaxique. Ils peuvent être définis comme un ensemble de mots ordonnés v_1, v_2,..., v_n. Ces mots sont représentés par les vecteurs suivants $\overrightarrow{v_1}$, $\overrightarrow{v_2}$, ..., $\overrightarrow{v_n}$. Soit λ une puissance de 2 représentant le poids de chaque mot dans l'arbre du groupe de mots. Par exemple, λ_2 est le poids du second mot du groupe de mots. Le vecteur du groupe de mots λ est obtenu par la somme récursive normée des :
1) mots appartenant au dit groupe de mots
2) sous-groupes de ce groupe de mots
Soit j appartenant à $[1, n]$. Alors, pour chaque groupe de mots d'un niveau i dans l'arbre issu de l'analyse syntaxique, en sachant que la racine de l'arbre possède le niveau 0 (le plus haut niveau), et les feuilles terminales le plus bas niveau (n), nous avons alors la formule récursive suivante pour calculer le vecteur d'un groupe de mots $\overrightarrow{\gamma_i}$:

$$\overrightarrow{\gamma_i} = \frac{\sum_j \overrightarrow{(\lambda_j v_{j,i+1})}}{\| \sum_j (\lambda_j v_{j,i+1}) \|} \tag{6.1}$$

Ainsi, le vecteur est normalisé à chaque appel récursif.
Définissons dès lors la représentation mathématique du calcul d'un vecteur sémantique de phrase. Soit σ une phrase analysée syntaxiquement. Si σ est un groupe de mots de niveau $i = 0$, ϕ_j sont alors des groupes de mots de niveau $i = 1$. En d'autres termes, ϕ_j sont les groupes de mots placés directement sous la racine de l'arbre syntaxique lors de l'analyse de la phrase σ. Alors, la formule permettant de calculer un vecteur sémantique de phrase $\overrightarrow{\sigma}$ est la suivante :

$$\overrightarrow{\sigma} = \frac{\sum_j \overrightarrow{(\lambda_j \phi_{j,i})_{nor}}}{\| \sum_j (\lambda_j \phi_{j,i})_{nor} \|} \tag{6.2}$$

Notons dans cette équation la présence de la notation “nor”. Celle-ci montre que les vecteurs des groupes de mots ϕ_j sont normalisés tel que montré dans la formule les définissant. Le calcul d'un vecteur sémantique de phrase est nécessaire lors du calcul d'un vecteur sémantique de mot “contextualisé”. Nous présentons ci-dessous ce type de vecteurs en motivant le calcul.

Les vecteurs sémantiques contextualisés de mots

Le paragraphe précédent présente le calcul d'un vecteur sémantique de phrase. Ce calcul prend en compte les constituants de cette phrase en fonction de leurs rôles syntaxiques comme “sujet” ou “verbe”. Une phrase et le vecteur sémantique la représentant véhiculent donc un contexte sémantique. En effet, des mots peuvent avoir plusieurs significations suivant la phrase d'où ils proviennent. Citons par exemple le verbe “consommer” qui

possède des sens différents dans les phrases suivantes :

1. Tu veux donc jusqu'au bout consommer ta fureur.[29]
2. Le pape exigea que ces deux enfants consommassent le mariage, le jour même de sa célébration, tant il craignit les subterfuges de la politique et les ruses en usage à cette époque.[30]
3. La Prusse rhénane et Lyon fabriquent tout le velours d'Utrecht qui se consomme dans le monde.[31]
4. La France trouvera de l'avantage dans la vente de ses grains, si, ne se bornant pas à vendre à ceux qui consomment chez elle, elle vend encore à ceux qui consomment dans les États où il lui est permis d'importer.[32]
5. Stevens parlait ce soir (...) de l'effrayant avalement de bière et d'alcool de Courbet, qui consommait trente bocks dans une soirée et prenait des absinthes où il remplaçait l'eau par du vin blanc.[33]

Le principe des vecteurs sémantiques contextualisés de mots est de tenir compte du contexte de la phrase afin de désambiguïser un mot. En d'autres termes, la signification contextuelle d'un mot dans une phrase reflète l'impact des autres mots composant cette phrase. Afin de produire un tel vecteur, le produit du mot "vecteur dictionnaire" (vecteur obtenu avec les concepts du thésaurus Larousse tel que décrit dans le premier paragraphe de cette section) avec le vecteur de la phrase d'où il provient (vecteur décrivant le contexte du mot) est calculé. Ce vecteur contextualisé peut être formulé de la manière suivante. Soit $\overrightarrow{v_p}$ le vecteur dictionnaire du mot v_p appartenant à la phrase σ_k, dont le vecteur sémantique s'écrit $\overrightarrow{\sigma_k}$. Ainsi, un vecteur contextualisé $\overrightarrow{v_p/\sigma_k}$ s'écrit :

$$\overrightarrow{v_p/\sigma_k} = \overrightarrow{v_p} \times \overrightarrow{\sigma_k} \tag{6.3}$$

Les vecteurs sémantiques normalisés du terme "consommer" pour chaque phrase précédente sont alors présentés dans la figure 6.4 (de manière respective). L'impact de la sémantique des phrases est dans cet exemple bien visible. Par exemple, la première phrase est plutôt relative à la *finalité* ou l'*achèvement* (concept numéro 58), la quatrième phrase quant à elle est partagée entre ce même concept d'*achèvement* et celui de la *dépense* (concept numéro 835), notamment influencé par le verbe "vendre".

[29] Citation tirée d'"Alzire" (François Marie Arouet dit Voltaire)
[30] Citation tirée de "Sur Catherine de Médicis" (Honoré de Balzac)
[31] Citation tirée de "En Hollande" (Maxime Du Camp)
[32] Citation tirée de "Le Commerce et le gouvernement considérés relativement l'un à l'autre" (Étienne Bonnot de Condillac)
[33] Citation tirée de "Journal" (Edmond et Jules de Goncourt)

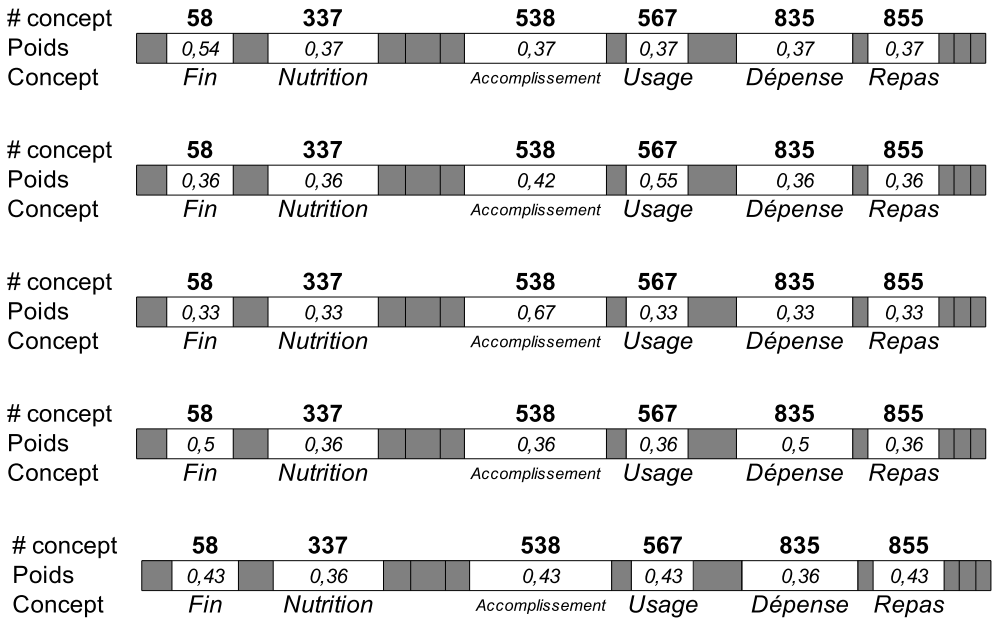

FIG. 6.4 – Vecteur sémantique contextualisé du verbe "consommer" pour cinq phrases sémantiquement distinctes

Cette contextualisation se révèle ici pertinente mais ne peut être appliquée en tant que telle pour notre approche. En effet, nous cherchons à mesurer la plausibilité d'une relation syntaxique induite qui n'est, rappelons-le, pas présente dans le corpus. Ainsi, nous ne pouvons identifier la phrase d'où elle serait hypothétiquement issue. Nous proposons alors une alternative présentée dans la section suivante, consistant à "globaliser" ces vecteurs sémantiques contextuels.

Les vecteurs sémantiques contextualisés globaux

Le fait de ne pouvoir identifier la phrase de laquelle provient une relation induite est problématique et nous a conduit à proposer ce type de vecteurs : **les vecteurs sémantiques contextualisés globaux**. Le principe est de produire des vecteurs prenant en compte le contexte de la totalité d'un corpus. Ainsi, en représentant le verbe d'une relation induite et son objet complémentaire par des vecteurs sémantiques contextualisés globaux, nous cherchons à savoir si **dans ce corpus**, ces deux termes peuvent "cohabiter". En d'autres termes, formeraient-ils une relation syntaxique correctement formulée sémantiquement, **dans la thématique de ce corpus**.

Ce vecteur global est obtenu de la manière suivante. Pour un terme donné et pour chaque phrase dans laquelle ce terme apparaît dans le corpus, nous calculons un vecteur sémantique contextualisé relatif. Le vecteur sémantique global contextualisé de ce terme est alors défini par le barycentre de tous les vecteurs sémantiques contextualisés de ce mot,

provenant des phrases du corpus. La figure 6.5 donne un exemple de vecteur sémantique

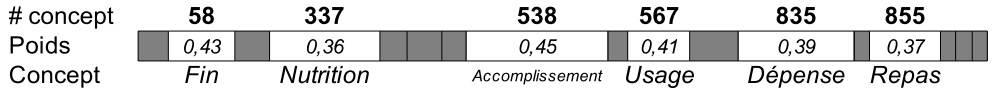

FIG. 6.5 – Vecteur sémantique contextualisé global du verbe “consommer”.

global contextualisé pour le verbe “consommer” en utilisant les cinq phrases de notre exemple précédent. Notons que ce vecteur ne peut refléter un contexte global. Les phrases qui ont été utilisées pour le construire ne sont en effet pas issues d'un même corpus.
Après avoir montré comment calculer différents types de vecteurs sémantiques, nous présentons dans la section suivante comment utiliser ces vecteurs afin de sélectionner les relations induites pertinentes.

6.3.1.3 Deux approches pour mesurer la qualité d'une relation syntaxique induite

L'objectif de l'utilisation de vecteurs sémantiques est de mesurer la pertinence de l'association d'un verbe avec son objet complémentaire afin de classer en termes de plausibilité les relations syntaxiques induites. La finalité est d'obtenir des descripteurs de bonne qualité avec notre modèle SELDEF en utilisant les objets complémentaires. Nous présentons ici deux approches permettant de sélectionner ces objets.

L'approche non contextuelle

Une première méthode **consiste à mesurer la qualité d'une relation syntaxique induite par rapport à la relation syntaxique dont elle est issue**. En d'autres termes, l'objectif est d'évaluer si deux relations syntaxiques, l'une classique, l'autre induite partagent les mêmes concepts. Ainsi, nous allons valider la proximité sémantique entre le verbe et l'objet d'une relation induite avec le verbe et l'objet de la relation originale. Concrètement avec l'exemple de la figure 6.2, il s'agit de mesurer la proximité sémantique des relations *manger fruit* (relation originale) et *consommer fruit* (relation induite).
Nous représentons alors chaque relation syntaxique sans prendre en compte le contexte des phrases d'où elles proviennent[34] (c.-à.-d. **sans les contextualiser**). Ceci permet de mesurer la proximité sémantique des deux relations. Ainsi, une relation syntaxique est pondérée uniquement avec les concepts du verbe et de l'objet, en tenant compte de la position du gouverneur dans l'arbre morphosyntaxique. Le contexte pris en compte est alors uniquement fondé sur la position hiérarchique des composantes de la relation

[34]Notons que une relation syntaxique induite ne peut être contextualisée du fait qu'elle ne provient pas d'une phrase du corpus.

syntaxique. Nous présentons, afin d'illustrer cette approche, la représentation vectorielle de la relation syntaxique "consommer fruit" dans la figure 6.6. Nous calculons dans un premier temps les vecteurs sémantiques du verbe "consommer" puis du nom "fruit". Nous obtenons finalement le vecteur résultant de la relation syntaxique[35]. Dans cet exemple, le

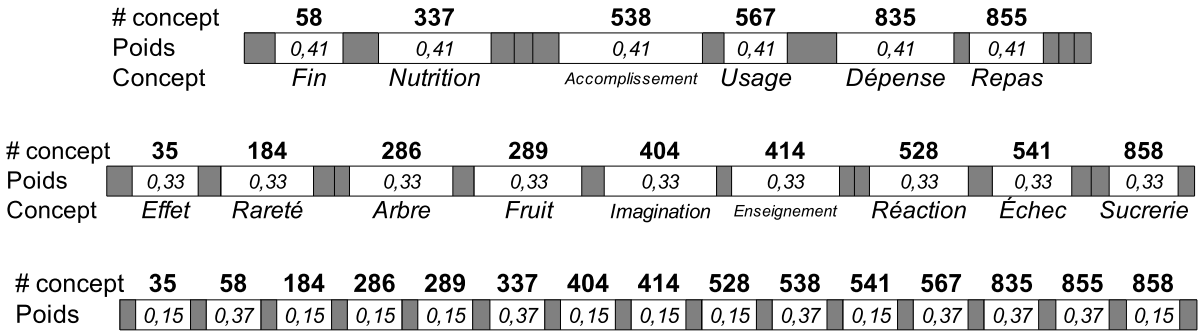

FIG. 6.6 – Vecteurs sémantiques respectifs non contextualisés de "consommer", "fruit", et "consommer fruit".

poids du verbe influe sur ses concepts, cela explique les différentes valeurs des composantes. Notons que cette méthode n'est pas très discriminante, car elle ne contextualise pas les relations syntaxiques. En effet, dans cet exemple, la thématique de la nourriture est mise en avant, mais n'est pas répercutée dans le vecteur sémantique de la relation syntaxique résultante. Les concepts "*nutrition, fruits, repas et sucrerie*" devraient en effet être mis en avant. Cette méthode, qui fut la première implémentée afin d'extraire des descripteurs de qualité à partir de relations induites, montre ses limites dans nos expérimentations qui seront présentées dans le chapitre 7. Nous proposons ci-dessous une approche tenant compte du contexte en employant les vecteurs sémantiques contextualisés globaux.

L'approche contextuelle

L'utilisation de vecteurs sémantiques contextualisés permet une richesse sémantique plus fine. Ainsi, dans cette seconde approche visant à mesurer la qualité des relations induites, nous proposons non plus de mesurer la proximité de relations syntaxiques, l'une induite et l'autre classique, mais de **mesurer directement la proximité d'un verbe et de son objet complémentaire**. Pour cela, nous utilisons des vecteurs sémantiques globaux contextualisés. Plusieurs représentations globales peuvent alors être employées pour le verbe et l'objet. En d'autres termes, quels peuvent-être les meilleurs descripteurs pour le verbe et l'objet ? Nous avons retenu les représentations suivantes :

– pour le verbe : le verbe "lui même" (le gouverneur du groupe verbal) et le groupe verbal complet

[35]Nous n'avons pas reporté les noms des concepts dans la figure représentant le vecteur sémantique de la relation syntaxique pour des soucis de lisibilité.

– pour l'objet : le nom (le gouverneur du groupe nominal) et le groupe nominal.
L'utilisation de groupes nominaux et verbaux afin de décrire respectivement un verbe et un objet s'apparente à l'utilisation de syntagmes verbaux ou nominaux. En effet, ne pouvant produire assez de relations syntaxiques en utilisant les syntagmes, nous pouvons ainsi utiliser les ressources sémantiques des contextes générés par les verbes et/ou les objets. Rappelons en effet que nos modèles SelDe et SelDeF dépendent du nombre d'objets communs et distincts de couple de verbes. Nous sommes ainsi contraint d'utiliser comme objets le gouverneur du "groupe objet". Dans le cas contraire, nous extrairions un nombre trop limité d'objets communs afin d'appliquer nos approches (cf section 3.4.2). C'est pourquoi nous proposons avec cette approche fondée sur des vecteurs sémantiques, d'utiliser des **représentations plus riches d'un verbe et de son objet**. En effet, un groupe nominal est plus porteur de sens qu'un simple nom. De plus, en les contextualisant, les groupes nominaux seront d'autant plus informatifs.

Par exemple, prenons la phrase "J'ai consommé des fruits rouges". Afin de représenter le vecteur du mot "fruit", nous prendrions alors, si nous choisissions d'utiliser les groupes nominaux, l'expression "des fruits rouges" afin de décrire "fruit". Le vecteur résultant donné dans la figure 6.7 en sort plus riche avec de nouveaux concepts activés par rapport au terme "fruit" seul. Ces nouveaux concepts ont les numéros *15, 21, 24, 102, 335, 356, 357* et *471* correspondant respectivement à : "*identité, ressemblance, uniformité, un, pilosité, brun,* et *rouge*". Notons que la sélection d'un groupe nominal ou verbal se fait avec l'appui de l'arbre syntaxique généré par l'analyse de Sygfran. Après avoir défini

# concept	15	21	24	35	102	184	286	289	235	356	357	404	414	471	528	541	611	858
Poids	0,2	0,2	0,2	0,27	0,2	0,27	0,27	0,27	0,18	0,18	0,18	0,27	0,27	0,18	0,27	0,27	0,18	0,27

Fig. 6.7 – Vecteurs sémantiques de "fruit" avec pour descripteur "des fruits rouges".

nos différentes approches pour représenter les relations syntaxiques induites, nous devons définir une mesure de qualité pour ordonner les vecteurs sémantiques obtenus. Nous présentons dans la section suivante les choix que nous avons effectués dans ce but.

6.3.1.4 Comment mesurer la proximité sémantique des vecteurs sémantiques ?

La proximité de vecteurs sémantiques peut être obtenue par des mesures "classiques" de similarité comme le *cosinus*. Cependant, bien que les vecteurs sémantiques s'apparentent à des vecteurs "classiques", certaines propriétés les caractérisant permettent de produire une mesure de similarité plus adaptée. Nous présentons ci-dessous deux mesures

de similarité, une classique : le *cosinus* et une plus spécifique aux vecteurs sémantiques : la *distance de concordance*.

Le cosinus

Rappelons que le cosinus entre deux vecteurs $\overrightarrow{u}$ et $\overrightarrow{v}$ de telle sorte que θ soit l'angle formé par ces deux vecteurs est défini par l'équation suivante (cf. 2) :

$$\theta = arccos \frac{\overrightarrow{u}.\overrightarrow{v}}{\|\overrightarrow{u}\|.\|\overrightarrow{v}\|} \tag{6.4}$$

Ainsi, nous calculons la proximité de chaque couple de relations syntaxiques (induites et classiques) pour la première approche et chaque couple de verbes et objets complémentaires provenant de relations induites pour la seconde, en utilisant le cosinus. Finalement, les résultats sont ordonnés de manière décroissante. Notons que dans le cas de la seconde approche, la similarité entre ces couples ne doit pas être trop importante. En effet, une similarité importante peut refléter une tautologie, comme par exemple “manger nourriture” ou bien “marcher à pied”. Certes, la plupart des relations ainsi formées sont pertinentes, mais n'apportent pas réellement de nouvelles informations. Ainsi, les verbes et objets utilisés lors de la seconde approche doivent être assez éloignés. Autrement, il n'y aurait aucun élément sémantique qui permettrait de mesurer leur proximité.
Une alternative à ce problème, qui est également une mesure plus adaptée afin d'évaluer la proximité de vecteurs sémantiques est présentée dans le paragraphe suivant.

La distance de concordance

La mesure de similarité “cosinus” n'est pas en mesure d'évaluer correctement la pertinence d'un objet pour un verbe donné. Il est possible d'obtenir le même score de proximité sémantique en employant le cosinus avec plusieurs composantes de vecteurs actifs ayant une faible valeur, et peu de composantes de vecteurs actifs ayant de fortes valeurs. Cependant, la pertinence de ces proximités n'est pas la même sémantiquement. En outre, si le vecteur gouverneur (le verbe) est réduit en ne conservant uniquement que les composantes des concepts actifs, et qu'il en est de même pour le second vecteur (l'objet), il est tout à fait possible de modifier leurs similarités (en utilisant le cosinus). Ainsi nous modifions la plausibilité d'une relation syntaxique induite formée par les deux vecteurs. La réduction vectorielle est envisagée pour un vecteur de 873 composantes dont la plupart sont inactives. Le vecteur résultant possède donc un nombre important de composantes nulles. L'objectif de la distance de concordance est d'être plus discriminante que le cosinus en ne considérant pas uniquement les valeurs des composantes des vecteurs mais également leurs rangs dans la hiérarchie donnée par le thésaurus. Ainsi, nous pouvons

utiliser une mesure de proximité sémantique sur des vecteurs de dimensions réduites.
La première utilisation de cette mesure fut expérimentée par [Chauché *et al.*, 2003] dans le but d'effectuer une classification automatique de textes, les textes étant représentés par des vecteurs sémantiques. Les auteurs ont montré dans cet article mais également dans [Chauché & Prince, 2007] que la distance de concordance améliorait les résultats obtenus. Cette mesure a également été utilisée dans le cadre de la segmentation thématique de textes [Labadié & Prince, 2008]. Les auteurs proposent de comparer dans cet article les résultats de deux méthodes de segmentation thématique de textes, *C99* [Choi, 2000] et TRANSEG [Labadié & Prince, 2008]. Les résultats expérimentaux ont montré que TRANSEG, fondé sur l'utilisation de vecteurs sémantiques avec une distance de concordance, obtient les meilleurs résultats en moyenne. Le paragraphe suivant défini de manière formelle la distance de concordance.

Considérons deux vecteurs $\vec{A}$ et $\vec{B}$. Nous les classons en fonction des valeurs de leurs composantes, de la plus active à la moins active. Nous appliquons alors une réduction des vecteurs triés en ne conservant que les $1/s$ premières composantes. Reste ainsi uniquement les composantes les plus "fortes". Les vecteurs résultants sont notés $\vec{A_{tr}}$ et $\vec{B_{tr}}$. Si les deux vecteurs ainsi formatés n'ont aucune composante commune, la distance de concordance vaut alors 1 (ils sont le plus éloignés possible). Dans les autres cas, nous devons calculer deux différences : la différence de rang et la différence d'intensité.

La différence de rang $E_{i,\rho(i)}$ est définie comme suit :

$$E_{i,\rho(i)} = \frac{(i-\rho(i))^2}{Nb^2 + (1+\frac{i}{2})} \tag{6.5}$$

Avec i qui est le rang de la composante C_t du vecteur $\vec{A_{tr}}$, et $\rho(i)$ le rang de la même composante mais pour le vecteur $\vec{B_{tr}}$, où Nb est le nombre de composantes conservées.

La différence d'intensité $I_{i,\rho(i)}$ qui compare la différence d'intensité des différentes composantes communes des deux vecteurs est définie par la formule suivante :

$$I_{i,\rho(i)} = \frac{\left\| a_i - b_{\rho(i)} \right\|}{Nb^2 + (\frac{1+i}{2})} \tag{6.6}$$

Avec a_i qui est l'intensité de la composante de rang i du vecteur $\vec{A_{tr}}$ et $b_{\rho(i)}$ l'intensité de la même composante de l'autre vecteur $\vec{B_{tr}}$ (et dont le rang est $\rho(i)$).

Après avoir défini ces deux différences, nous pouvons mesurer la concordance P :

$$P(\vec{A_{tr}}, \vec{B_{tr}}) = (\frac{\sum_{i=0}^{Nb-1} \frac{1}{1+E_{i,\rho(i)} * I_{i,\rho(i)}}}{Nb})^2 \tag{6.7}$$

Cependant, la concordance P se concentre sur l'intensité et le rang des composantes et ne développe pas la notion de direction que possède la distance angulaire. Ainsi, cette notion est introduite en combinant la concordance avec la distance angulaire notée $\delta(\vec{A}, \vec{B})$ pour les vecteurs $\vec{A}$ et $\vec{B}$. Nous notons alors $\Delta(\vec{A_{tr}}, \vec{B_{tr}})$ le résultat de cette combinaison dont la définition est donnée ci dessous :

$$\Delta(\vec{A_{tr}}, \vec{B_{tr}}) = \frac{P(\vec{A_{tr}}, \vec{B_{tr}}) * \delta(\vec{A}, \vec{B})}{w * P(\vec{A_{tr}}, \vec{B_{tr}}) + (1-w) * \delta(\vec{A}, \vec{B})} \tag{6.8}$$

Dans cette formule, w est un coefficient pondérant l'importance qui doit être donnée à la distance angulaire P. Néanmoins, la mesure résultante n'est pas symétrique, cette mesure ayant été au départ conçue pour une tâche de classification automatique de textes. Nous la rendons donc symétrique en proposant la distance de concordance notée D telle que définie ci-dessous :

$$D(\vec{A}, \vec{B}) = \frac{\Delta(\vec{A_{tr}}, \vec{B_{tr}}) + \Delta(\vec{B_{tr}}, \vec{A_{tr}})}{2} \tag{6.9}$$

Notons que cette distance de concordance est une distance au sens mathématique respectant les propriétés de symétrie et d'inégalité triangulaire [Labadié, 2008]. Finalement, pour permettre une combinaison avec un scalaire nous proposons la mesure suivante notée D_{Final}. Le score résultant sera ainsi compris entre 0 et 1 avec un score de 1 pour les vecteurs similaires.

$$D(\vec{A}, \vec{B})_{Final} = 1 - D(\vec{A}, \vec{B}) \tag{6.10}$$

Après avoir calculé la distance de concordance entre chaque vecteur représentant les relations syntaxiques (induites et classiques) dans le cas de la première approche, et représentant le verbe et l'objet complémentaire de relations induites dans le cas de la seconde approche, nous obtenons un classement en termes de plausibilité des relations induites.
La section suivante présente notre seconde méthode permettant d'obtenir ce classement : la validation par le Web.

6.3.2 La validation par le Web

L'approche présentée dans la section précédente se fonde sur d'importantes ressources et processus du TAL afin de déterminer la qualité d'une relation syntaxique induite : un thésaurus, une représentation vectorielle de mots (s'appuyant sur ce thésaurus), un

analyseur syntaxique calculant la dépendance des relations syntaxiques, une procédure calculant les vecteurs sémantiques de phrases. Les vecteurs contextualisés ainsi produits sont alimentés par les phrases extraites du corpus. La question que nous pouvons nous poser est alors la suivante. Est-il possible d'évaluer la qualité d'une relation syntaxique induite par le biais d'autres connaissances, n'utilisant pas de ressources plus ou moins coûteuse de TAL (qui d'un autre côté doivent garantir une certaine cohérence linguistique) ?
Par conséquent, nous proposons une approche fondée sur les ressources du Web pour mesurer la proximité sémantique de l'objet complémentaire d'un verbe donné dans une relation syntaxique induite. La proximité sémantique est ici supposée comme reflétée par la popularité de la dite relation sur le Web, par le biais d'un moteur de recherche Web. Ce type d'approche a déjà été utilisé dans divers travaux de la littérature que nous présentons de manière non exhaustive dans le paragraphe suivant.

6.3.2.1 Travaux relatifs à la validation par le Web

La "validation Web" utilisée pour mesurer la dépendance entre verbe et objet d'une relation induite, est proche de la méthode de [Turney, 2001], utilisant le Web pour définir une fonction de rang. L'algorithme PMI-IR (Pointwise Mutual Information and Information Retrieval) décrit par [Turney, 2001] utilise le moteur de recherche AltaVista pour déterminer les synonymes appropriés à une requête donnée.
Pour un mot donné, noté *word*, PMI-IR permet de choisir un synonyme parmi une liste donnée. Les termes sélectionnés notés $choice_i$ avec $i \in [1, n]$ correspondent à des questions du TOEFL. L'objectif est alors de trouver le synonyme parmi les termes $choice_i$ qui donnera le meilleur score. Le calcul du score de l'algorithme PMI-IR utilise diverses mesures fondées sur la proportion de documents où les deux termes sont présents (c.-à-d. le terme dont on cherche un synonyme – *word* – et le synonyme candidat – $choice_i$). La formule de Turney est donnée ci-dessous. C'est l'une des mesures de base utilisée dans [Turney, 2001]. Cette mesure s'appuie sur l'information mutuelle précédemment évoquée (cf chapitre 2).

$$score(\ choice_i\ ) = \frac{nb(\ word\ NEAR\ choice_i\ )}{nb(\ choice_i\ )} \tag{6.11}$$

Avec :
– $nb(x)$ qui donne le nombre de documents contenant le mot x,
– $NEAR$ (utilisé avec la section "*recherche avancée*" du moteur de recherche Altavista) est un opérateur indiquant si deux mots sont présents dans une fenêtre de 10 mots.

Avec cette formule (6.11), la proportion de documents contenant les deux mots *word* et $choice_i$ est calculée et comparée avec le nombre de documents contenant le mot $choice_i$. Plus cette proportion est importante et plus les deux mots sont considérés comme synonymes. D'autres formules plus sophistiquées sont appliquées. Elles prennent en compte la négation dans la fenêtre de 10 mots. Par exemple, les mots "petit" et "gros" sont potentiellement synonymes, si dans une fenêtre donnée, a été détectée une négation associée à l'un de ces deux mots.

D'autres approches utilisent le Web dans la littérature comme les travaux récents de [Cilibrasi & Vitanyi, 2007] qui proposent de calculer une distance de similarité entre des termes en utilisant le moteur de recherche Google. Le principe est de considérer le nombre de pages retournées par le moteur de recherche de Google comme la fréquence d'apparition d'un terme dans une base de données qui n'est autre que le Web. Notons que cette distance est nommée "Normalized Google Distance" (NGD) par ses auteurs, et qu'il est démontré dans l'article que NGD est bien une distance mathématique. La formule permettant un tel calcul de distance entre deux mots x et y est donnée ci-dessous :

$$NGD(x,y) = \frac{max\{log[f(x)], log[f(y)]\} - log[f(x,y)]}{logN - min\{log[f(x)], log[f(y)]\}} \tag{6.12}$$

Avec :
– $f(x)$ le nombre de pages[36] contenant le mot x
– $f(x,y)$ le nombre de pages contenant les mot x **et** y
– N est un paramètre couramment fixé à $f(x)$ ou $f(y)$
Les auteurs indiquent cependant que la variation du paramètre N n'a que peut d'effet sur la distance, en ce sens que le classement obtenu est peu modifié.
De nombreux travaux de la littérature utilisent des approches similaire comme [Roche & Prince, 2008] proposant une mesure nommé AcroDef, permettant de lever les ambiguïtés d'acronymes. Cette mesure utilise le Web afin d'attribuer un score aux définitions d'acronymes en s'appuyant sur des mesures statistiques et un contexte. D'autres travaux se fondent également sur les ressources du Web afin d'acquérir un corpus. Citons [Baroni & Bernardini, 2004] qui utilisent un moteur de recherche. Ils proposent un outil nommé BootCaT (Bootstrapping Corpora and Terms from the Web) qui génère de manière automatique un corpus en effectuant un certain nombre de requêtes via un moteur de recherche. La seule information à fournir est une liste de mots clés représentant le domaine pour lequel on souhaite acquérir un corpus. Nous pouvons également citer les travaux de [Larkey *et al.*, 2000]. Ces derniers présentent dans cet article l'outil *Acrophile*,

[36] Le nombre de page sur le Web selon le référencement de Google

permettant la construction automatique d'une base de données d'acronymes et d'abréviations en se fondant notamment sur les ressources du Web.
Notre approche est assez similaire à celle de [Cilibrasi & Vitanyi, 2007]. Nous proposons également de mesurer la proximité de termes en utilisant un moteur de recherche. Notons cependant des divergences avec notre approche. Nous ne proposons pas de calculer une distance entre deux termes mais simplement une "proximité sémantique". De plus notre approche est spécifique aux relations syntaxiques comme nous allons le montrer dans la section suivante. Enfin, nous associons dans notre approche les ressources du Web à des mesures classiques de la littérature.

6.3.2.2 Notre approche de validation Web

L'approche de validation par le Web que nous proposons a pour objectif de **mesurer la dépendance entre un verbe et un objet d'une relation induite** afin d'établir un classement par pertinence des relations. Pour cela, nous interrogeons le Web en fournissant à un moteur de recherche une requête (relation syntaxique) sous forme de chaîne de caractères (par exemple, "consommer un fruit"). Cette approche présente la particularité de refléter la popularité d'une relation syntaxique sur le Web, s'adaptant ainsi à une époque ou une certaine mode d'écriture. Plusieurs questions se posent alors.

1. L'information donnée par le Web est-elle un bon indice de plausibilité ?
2. Comment formater une chaîne de caractères afin de décrire une relation syntaxique ?
3. Comment introduire des informations contextuelles du Web afin de prendre en compte la popularité d'une relation syntaxique et de chaque terme la composant. En d'autres termes, est-il suffisant d'interroger le Web avec uniquement une relation syntaxique ?

• Le premier point pose le problème de la qualité des informations fournies par le Web. Le moteur de recherche de *Yahoo !* indexait plus de 10 milliards de pages Web en Août 2005[37]. Ceci donne alors un ordre d'idée du nombre d'utilisateurs contribuant à la création de ces pages Web. Ainsi, est il audacieux d'affirmer que la fréquence des pages retournées par *Yahoo !* lors de la soumission de termes reflète, dans une certaine mesure, la popularité de ces termes dans une société (francophone dans notre cas, sachant que nous ne traitons que la langue française) ? Nous émettons cette hypothèse afin de proposer cette approche visant à mesurer la dépendance entre un verbe et un objet d'une relation induite. En effet une simple expérience consistant à interroger le moteur de

[37] Cette information a été donnée sur le blog de Yahoo ! (*http : //www.ysearchblog.com*) à cette date.

Yahoo ! avec un terme mal orthographié et un second bien orthographié montre que les ressources du Web reflètent une certaine qualité. Par exemple :
– antropofagie = 122 pages indexées contenant ce terme
– anthropophagie = 197 000 pages indexées contenant ce terme

• Bien que notre approche soit similaire à celle de [Cilibrasi & Vitanyi, 2007], rappelons que nous devons soumettre au moteur de recherche une relation syntaxique. Supposons respectivement que *consommer* et $fruit$ représentent un verbe et un objet d'une relation syntaxique[38]. Le fait de mesurer la qualité d'une telle relation en s'intéressant au nombre de pages retournées par un moteur de recherche contenant *consommer* et $fruit$ n'est pas pertinent car l'information portée par la relation syntaxique n'est pas prise en compte. Ainsi, nous proposons d'effectuer une requête sous forme de chaîne de caractères entre doubles quottes, afin de conserver le caractère séquentiel d'une relation syntaxique.

Nous devons alors dans un premier temps lemmatiser la relation syntaxique induite. Puis nous soumettons par exemple la requête "*consommer* $fruit$" au moteur de recherche. Cependant, il est peu probable que cette séquence retourne des résultats. En effet, en français, une relation syntaxique est souvent séparée par des articles comme "*le*" ou "*un*" produisant les relations "*consommer le* $fruit$" ou bien "*consommer un* $fruit$". Ainsi, il nous est nécessaire d'introduire dans notre requête cinq articles fréquemment utilisés en français : *un*, *une*, *le*, *la*, *l'*. Il en résulte alors cinq requêtes contenant chacune un article différent. Nous devons par ailleurs déterminer quel est l'article le plus adapté à une relation syntaxique donnée afin de limiter la quantité de bruit pouvant apparaître. Nous proposons deux variantes : une première dont le principe est de sélectionner la requête retournant un nombre maximal de pages Web et une seconde faisant la somme des résultats de chaque requête. Nous présentons ci-dessous un exemple de calcul de plausibilité pour la relation syntaxique induite "*consommer*$fruit$" issue de la figure 6.2. Nous soumettons dans un premier temps au moteur de recherche les cinq requêtes prenant en compte les articles.
nb("*consommer* un $fruit$") = 571 pages contenant cette séquence
nb("*consommer* une $fruit$") = 0 page contenant cette séquence
nb("*consommer* le $fruit$") = 875 pages contenant cette séquence
nb("*consommer* la $fruit$") = 2 pages contenant cette séquence
nb("*consommer* l'$fruit$") = 0 page contenant cette séquence
Notons que la fonction "$nb(x)$" calcule le nombre de pages retournées lors de l'interroga-

[38]Notons que l'approche présentée dans cette section, bien qu'originalement créée pour mesurer la qualité de relations induites peut tout à fait être appliquée à toutes relations syntaxiques.

tion d'un moteur de recherche en lui soumettant une requête x.

Dans le cas d'une relation syntaxique de type verbe-objet nous noterons cette fonction $nb(v, o)$ avec v le verbe et o l'objet. Alors, nos deux méthodes utilisant la somme ou le maximum seront notées respectivement $nb_{sum}(v, o)$ et $nb_{max}(v, o)$. Nous obtenons avec notre exemple les résultats suivants :
$nb_{sum}(v, o) = 571 + 0 + 875 + 2 + 0 = 1448$
$nb_{max}(v, o) = max\{571, 0, 875, 2, 0\} = 875$
Remarquons à titre comparatif que nous avons obtenu les scores suivants avec l'autre relation induite issue de la figure 6.2 "*manger essence*" :
$nb_{sum}(v, o) = 1 + 1 + 0 + 1 + 39 = 42$
$nb_{max}(v, o) = max\{1, 1, 0, 1, 39\} = 39$
Nous montrons par cet exemple qu'avec la validation Web, la relation syntaxique "manger essence" est moins plausible que "consommer fruit".

• Le dernier point de cette section s'interroge sur le fait d'introduire le nombre de requêtes individuelles pour chaque composante de la relation syntaxique à savoir le verbe et l'objet. En effet, si nous sommes en présence d'un verbe ou d'un objet assez rare, la simple interrogation du Web avec $nb(v, o)$ peut se voir pénalisée. Par exemple avec le verbe "quérir" qui est assez rare, la relation "*quérir un sac*" renvoie seulement 4 résultats soit moins que "*manger l'essence*". L'association sémantique de "*quérir*" et "*sac*" semble pourtant plus cohérente que celle de "*manger*" et "*essence*".
Afin de contourner ce problème, nous utilisons différentes mesures statistiques permettant d'introduire l'information portée par le *verbe* et l'*objet* de manière individuelle. Certaines de ces mesures ont déjà été utilisées dans la section 3.3.4 dans le but de les comparer à la mesure d'Asium. L'objectif dans ce chapitre est différent, visant à faire ressortir la popularité de relations syntaxiques sur le Web et de mesurer celle-ci sur la base de différentes mesures statistiques. Rappelons que ces dernières mesurent une certaine forme de dépendance entre les mots (verbes et objets dans notre cas). Ces mesures sont les suivantes :
– La fréquence[39]
– L'information mutuelle
Adaptée à notre approche, l'information mutuelle entre le verbe noté "v" et l'objet noté

[39]La fréquence représente ici uniquement l'interrogation de la relation syntaxique seule avec $nb(v, o)$ et n'est pas une mesure statistique proprement dite, mais peut être utilisée comme telle.

"o" d'une relation syntaxique s'écrit :

$$IM(v,o) = \frac{nb(v,o)}{nb(v)nb(o)} \tag{6.13}$$

– L'information mutuelle au cube :

$$IM^3(v,o) = \frac{nb(v,o)^3}{nb(v)nb(o)} \tag{6.14}$$

– Le coefficient de Dice :

$$Dice(v,o) = \frac{2 \times nb(v,o)}{nb(v) + nb(o)} \tag{6.15}$$

Ainsi, nous pouvons utiliser huit variantes pour la validation de relations syntaxiques induites avec l'approche "validation Web". En effet, les quatre mesures présentées ci-dessus utilisent $nb(v,o)$ qui peut être la *somme* ($nb_{sum}(v,o)$) ou bien le *maximum* ($nb_{max}(v,o)$).
Remarquons finalement qu'avec cette approche, les relations populaires sont valorisées. En d'autres termes, une relation qui n'est plus "à la mode" obtiendra un score de plausibilité inférieur à une autre qui pourrait être de moins bonne qualité d'un point de vue sémantique.

Après avoir défini nos deux approches permettant une validation automatique des relations syntaxiques induites, nous proposons dans la section suivante deux manières de les combiner en motivant par la même notre choix.

6.3.3 Les approches hybrides

Chacune des deux approches présentées dans les sections précédentes, les *vecteurs sémantiques* (VS) et la *validation Web* (VW), ont des avantages et des inconvénients. L'approche VS est précise, linguistiquement fondée, mais elle repose sur des ressources importantes, et pourrait être biaisée par la nature du corpus étudié. D'un autre côté, l'approche VW est facile à mettre en œuvre, parcourt une grande quantité de données (provenant du Web) mais peut être critiquée dans ses principes fondamentaux. La popularité n'est pas nécessairement une preuve de qualité, et les pages Web peuvent être plus ou moins bien écrites avec des phrases d'un style assez pauvre, conduisant à des relations syntaxiques bruitées. De plus, cette approche peut se révéler très coûteuse en

termes de temps. Ne pouvant réduire le temps nécessaire au calcul de cette approche, et les ressources linguistiques de VS étant le fondement même de l'approche, nous avons cherché des combinaisons entre ces deux méthodes afin d'obtenir de meilleurs résultats. Nous proposons alors deux types de combinaisons présentées dans les sections suivantes.

6.3.3.1 Combinaison 1 : Une combinaison pondérée par un scalaire (HyPon)

La première combinaison entre ces approches consiste à introduire un paramètre $k \in [0, 1]$ pour donner un poids supplémentaire à l'une ou l'autre des approches. Nous normalisons au préalable les résultats donnés par les deux approches à combiner. Alors, pour une relation syntaxique r, nous combinons les approches avec le calcul suivant :

$$combine\_score_r = k \times VS + (1 - k) \times VW \tag{6.16}$$

6.3.3.2 Combinaison 2 : Un système hybride adaptatif (HybAd)

Nous présentons une seconde approche combinant VS et VW, l'**hybridation adaptative**. Le principe de cette combinaison est de classer dans un premier temps la totalité des relations syntaxiques par l'approche VS. Nous retenons et plaçons en tête les n premières relations syntaxiques. Ensuite, l'approche VW effectue le classement des n relations retenues par la méthode VS. Ainsi, avec notre approche adaptative, VS effectue une sélection globale sur la base des connaissances sémantiques et VW affine la sélection préalablement effectuée. Notons que si n correspond au nombre total de relations syntaxiques, ceci revient à appliquer une validation Web "classique". L'ensemble des approches présentées dans cette section va fournir une liste de candidats aux différents concepts ordonnés par valeurs décroissantes des différentes mesures (VS, VW, HyPon, HybAd).

6.3.4 Exemple de classement avec cinq relations induites

Cette section présente un exemple de classements de relations syntaxiques induites avec les différentes approches présentées précédemment : les vecteurs sémantiques, la validation Web et les deux approches de combinaisons HyPon et HybAd.
Afin de simplifier l'exemple ci-dessous, nous avons limité les résultats à une liste triée par méthode. Ainsi, pour l'approche VS, nous avons utilisé le cosinus afin de mesurer la proximité des verbes et objets complémentaires représentés avec des vecteurs sémantiques contextualisés globaux. La méthode VW quant à elle employée avec l'information mutuelle au cube comme mesure statistique. Concernant les combinaisons, le paramètre k a été fixé à 0.5 donnant un même poids à VS et VW dans le cadre de la première

Relations Verbe-Objet	
Induites	Originales
lancer recherche	mener recherche
poursuivre réforme	demander réforme
réussir évaluation	faire évaluation
dépasser recherche	faire recherche
dire croisade	poursuivre croisade

TAB. 6.1 – Résultats avec les vecteurs sémantiques.

combinaison. Le seuil pour la seconde combinaison est fixé à 3 qui représente un assez bon compromis entre les deux approches VS et VW.
Les cinq relations syntaxiques induites que nous avons choisies d'évaluer en termes de plausibilité pour cet exemple sont présentées dans le tableau 6.1. Nous calculons tout d'abord les scores résultant de l'approche des vecteurs sémantiques. Nous représentons

Relations Verb-Object	Cosinus
poursuivre réforme	0,60
dépasser recherche	0,52
réussir évaluation	0,41
dire croisade	0,37
lancer recherche	0,27

TAB. 6.2 – Résultats avec les vecteurs sémantiques.

sous forme de vecteurs sémantiques contextualisés globaux les verbes et objets des relations induites avec SYGFRAN. Nous pouvons alors calculer le cosinus entre ces verbes et objets complémentaires. Les résultats normalisés obtenus pour les cinq relations testées sont présentés dans le tableau 6.2.
Nous calculons ensuite les scores résultant de la validation Web. Nous effectuons pour cela des requêtes sur le Web avec les objets, verbes et relations syntaxiques induites afin de déterminer le nombre de résultats retournés par le moteur de recherche (fonction nb). Nous pouvons donc calculer l'IM^3 pour les cinq relations induites. Les scores obtenus sont présentés dans le tableau 6.3, qui sont également normalisés afin d'appliquer la première combinaison (avec k=0,5). Pour la seconde combinaison (HYBAD), nous classons dans un premier temps les relations syntaxiques induites avec l'approche VS, pour ensuite classer les 3 relations jugées par VS les plus plausibles avec VW (paramètre de HYBAD fixé à 3 dans cet exemple).
Les résultats obtenus pour les deux combinaisons[40] sont présentés dans le tableau 6.4.

[40] Afin de représenter sous forme de scores l'application de l'approche adaptative pour les 3 meilleures relations classées par VS, nous reportons leurs scores respectifs obtenus avec VW, auxquelles nous ajou-

Verbe-Objet	nb(Verbe)	nb(Objet)	nb(Verbe, Objet)	IM³
lancer recherche	82 700 000	863 000 000	2 299 288	0,71
poursuivre réforme	46 200 000	39 000 000	45 914	0,49
dire croisade	370 000 000	4 120 000	72	0,13
réussir évaluation	27 600 000	57 900 000	1 366	0,35
dépasser recherche	15 900 000	863 000	363	0,33

TAB. 6.3 – Résultats avec la validation Web.

Les résultats des combinaisons sont présentés dans le même tableau afin de comparer les

Verb-Object relations	VS	VW	Combinaison 1	Combinaison 2
poursuivre réforme	0,60	0,49	0,55	1,49
réussir évaluation	0,41	0,35	0,38	1,35
dépasser recherche	0,52	0,33	0,43	1,33
dire croisade	0,37	0,13	0,25	0,37
lancer recherche	0,27	0,71	0,49	0,27

TAB. 6.4 – Relations syntaxiques triées avec l'ensemble des approches.

approches VS et VW. Une fois l'ensemble des scores obtenus pour chacune des approches, nous pouvons ordonner les relations syntaxiques. Le classement finalement obtenu pour cet exemple est donné dans le tableau 6.5.

VS	VW	Combinaison 1	Combinaison 2
poursuivre réforme	lancer recherche	poursuivre réforme	poursuivre réforme
dépasser recherche	poursuivre réforme	lancer recherche	réussir évaluation
réussir évaluation	réussir évaluation	dépasser recherche	dépasser recherche
dire croisade	dépasser recherche	réussir évaluation	dire croisade
lancer recherche	dire croisade	dire croisade	lancer recherche

TAB. 6.5 – Classement obtenu des relations syntaxiques.

Plusieurs points relatifs à cet exemple sont discutés dans le paragraphe suivant.

- Les résultats donnés avec le cosinus pour l'approche VS sont assez faibles. La seule relation à posséder un score acceptable est obtenue avec la relation induite "poursuivre réforme". En effet, un score acceptable pour le cosinus est généralement significativement bien supérieur à $0,5$.
- Une relation syntaxique comme "lancer recherche" est considérée comme improbable avec l'approche VS. Ce résultat non pertinent peut être expliqué par deux points.

tons 1, ce qui place ces 3 relations automatiquement en tête de liste (car les scores sont normalisés entre 0 et 1).

(1) *La nature du corpus*. Le terme "recherche" peut en effet être présent dans le corpus dans des contextes très divers, et il en est de même pour le verbe "lancer".
(2) *Le pouvoir discriminant de la mesure*. Le cosinus n'est pas une mesure très discriminante. C'est la raison pour laquelle nous avons proposé d'utiliser dans nos expérimentations la distance de concordance.

- La mesure VW fournit presque des résultats opposés à ceux de VS. "Lancer recherche" obtient en effet le meilleur score avec cette approche et "poursuivre réforme" le deuxième score, qui sont les deux derniers dans le classement de l'approche VS. Finalement, "réussir évaluation", relation plausible, se voit mal classée par les deux approches.
- **Les résultats fournis par les combinaisons sont pertinents**. La première classe en tête les relations "poursuivre réforme" et "lancer recherche" quant à la seconde elle classe en tête les relations "poursuivre réforme" et "réussir évaluation". Rappelons que cette dernière relation était considérée comme "non significative" par les deux approches VS et VW seules (classée troisième sur cinq). Ainsi, nous montrons ici l'intérêt des combinaisons, renforçant le score des relations de qualité.

6.4 Synthèse

Nous avons présenté dans ce chapitre un modèle de sélection de descripteurs SelDeF. Après avoir analysé syntaxiquement un texte, et identifié parmi les relations syntaxiques de type verbe-objet quels étaient les verbes proches, nous avons distingué deux types d'objets provenant de ces couples. Les objets communs aux verbes, servent à décrire un texte au sens du modèle SelDe et les objets complémentaires non utilisés dans SelDe car contenant trop de bruit. L'objet de ce chapitre, en présentant le modèle SelDeF fut de proposer différentes approches afin de valider ces objets. Ainsi, ils peuvent être utilisés comme descripteurs. Nous présentons dans le chapitre suivant les expérimentations que nous avons menées avec les descripteurs fournis par ce modèle. Ces expérimentations proposent de construire des classes conceptuelles et de les enrichir en suivant un nouveau paradigme (par rapport à celui utilisé dans le chapitre 4). En effet, les descripteurs fournis par SelDeF ne servent pas à enrichir un corpus mais à définir les termes de concepts.

Chapitre 7

La construction et l'enrichissement de classes conceptuelles via SelDeF

Sommaire

7.1 Des descripteurs de SelDe aux classes conceptuelles

7.1.1 Préambule

Le modèle de sélection de descripteurs SelDeF présenté dans le chapitre précédent va être expérimenté dans ce chapitre. Rappelons que ce modèle, fondé sur le modèle SelDe présenté dans le chapitre 3, propose de sélectionner des descripteurs en filtrant des objets de verbes dits "complémentaires". Rappelons que SelDe s'inspire de l'approche d'Asium. Ce dernier propose à l'origine de former des groupes de mots sémantiquement proches. Ainsi, nous proposons dans ce chapitre d'expérimenter les descripteurs fournis par SelDe et SelDeF dans le cadre de la construction et l'enrichissement de classes conceptuelles. Notons que la construction de telles classes se distingue de la classification conceptuelle présentée dans le chapitre 4. En effet, dans ce précédent chapitre, nous cherchions à construire des classes par le biais d'un apprentissage supervisé, en se fondant sur des algorithmes classiques de classification. Un certain nombre de termes furent sélectionnés par un expert et notre tâche a consisté à les classer automatiquement dans des catégories,

elles aussi définies par l'expert. Le modèle SelDe a alors été employé afin d'enrichir le corpus via la méthode *ExpLSA*. Une représentation de type "sac de mots" du corpus enrichi a alors été utilisée pour la classification conceptuelle. Ce principe est très différent de la méthode développée dans ce chapitre. Dans ce dernier, nous nous intéressons à une nouvelle méthode de construction de classes conceptuelles, n'utilisant pas d'apprentissage. Ainsi, les descripteurs fournis par SelDe ne servent pas à enrichir un corpus mais à définir les instances des futurs concepts. Les descripteurs de SelDeF vont quant à eux permettre d'enrichir les classes construites.

7.1.2 La terminologie issue d'un corpus

La terminologie est un domaine ayant de nombreuses applications en TAL (Traitement Automatique des Langues). Elle peut-être vue comme l'étude des mots techniques propres à un domaine et de leurs significations. Nous distinguons deux types d'études terminologiques : l'approche sémasiologique et l'approche onomasiologique. La première s'intéresse à l'étude des significations partant du mot pour en étudier le sens. La seconde, propose de partir du concept. Un concept peut être défini comme la représentation mentale d'une chose ou d'un objet [Desrosiers-Sabbath, 1984]. Nous proposons de définir un concept comme un "*ensemble de connaissances partageant des caractéristiques sémantiques communes*". Nous utilisons dans ce chapitre l'approche sémasiologique en apportant un début de réponse aux problèmes générés par ce type d'approches. La terminologie ainsi extraite est très dépendante du corpus. Cela implique alors qu'une terminologie répondant à des besoins spécifiques est vouée à une faible réutilisabilité [Roche, 2005]. Nous proposons dans nos travaux de construire dans un premier temps des classes conceptuelles spécifiques en nous appuyant sur les données issues de corpus. Pour cette tâche, nous construisons des classes conceptuelles en étudiant la dépendance syntaxique des termes d'un corpus en se fondant sur le modèle SelDe.

7.1.3 La construction de classes conceptuelle fondée sur le modèle SelDe

Les classes conceptuelles proposées dans ce chapitre sont construites avec les descripteurs proposés par le modèle SelDe. Le principe de notre méthode consiste à rassembler les objets des verbes jugés proches (cf 3.4). Nous illustrons dans la figure 7.1 un exemple de verbes sémantiquement proches : *agiter* et *brandir*. Les **objets communs** à ces deux verbes peuvent alors définir le concept *Objets symboliques*, où en d'autres termes, les descripteurs fournis par SelDe. Nous remarquerons que le concept formé ici est d'une thématique peu fréquente montrant l'intérêt d'une telle construction de

classes conceptuelles spécifiques. Cette méthode de construction de classes conceptuelles

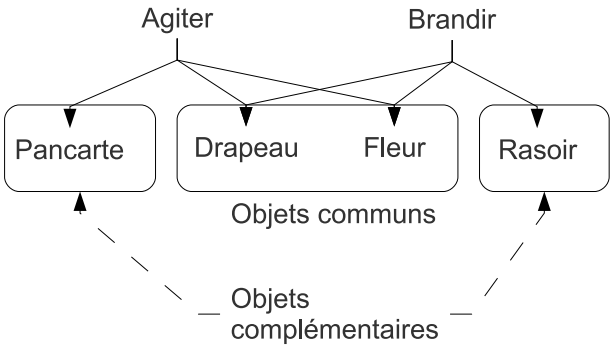

FIG. 7.1 – Objets communs et complémentaires des verbes “Agiter” et “Brandir”

s'inspire du modèle de David Faure, le système ASIUM. Cependant plusieurs points diffèrent en comparaison avec notre approche.

- **Les type de relations syntaxiques**. Rappelons que le système d'Asium considère les relations de type verbe-objet et sujet-verbe. Nous ne considérons que les relations syntaxiques de type verbe-objet, jugeant les relations sujet-verbe difficilement exploitables, notamment sur des corpus contenant beaucoup d'anaphores comme les pronoms. Ces dernières peuvent être définie comme *la répétition d'un même mot ou groupe de mots au début de plusieurs phrases ou propositions successives* (TLFi *Trésor de la Langue Française Informatisé*).
- **Les paramètres de filtrage**. Le modèle SELDE introduit des paramètres de filtrage afin de ne pas considérer l'ensemble des objets provenant des relations syntaxiques.
- **Les objets complémentaires**. Le système ASIUM utilise les objets complémentaires en les faisant valider par un expert. Nous ne considérons pas ces objets dans la phase de construction des classes. Nous proposons ainsi de les traiter comme pouvant être des instances potentielles des concepts en les validant via notre modèle SELDEF.

Les expérimentations menées dans ce chapitre sont organisées comme suit. Nous présentons tout d'abord dans la section suivante (section 7.2) une expérimentation menée sur un ensemble de concepts afin d'évaluer la qualité de l'enrichissement de SELDEF. Notons que les concepts qui vont être enrichis sont obtenus avec les descripteurs du modèle SELDE. Nous proposons également une autre approche d'enrichissement de classes conceptuelles fondées sur le Web. Celle-ci sera également expérimentée dans

la section suivante. Nous présentons ensuite dans la section 7.3 des expérimentations menées sur la validation d'objets complémentaires afin d'évaluer la qualité du modèle SelDeF. Ces dernières expérimentations se focalisent uniquement sur la validation d'un nombre important d'objets complémentaires. Ainsi, nous évaluerons la qualité de chaque approche de validation du modèle SelDeF présentée dans le chapitre 6.

7.2 Évaluation de la construction et de l'enrichissement de classes conceptuelles

Nous présentons dans cette section une première expérimentation mesurant la qualité des descripteurs fournis par le modèle SelDeF. Ces expérimentations sont organisées comme suit. Nous montrons dans un premier temps la construction des classes conceptuelles "d'origine" (section 7.2.1). Nous proposons alors deux méthodes pour effectuer cet enrichissement. Nous enrichissons dans un premier temps les classes en employant les descripteurs du modèle SelDeF tel que décrit dans la section 7.2.2 . La seconde approche utilise quant à elle les ressources du Web afin de proposer de nouveaux termes. Elle sera présentée en section 7.2.3. La figure 7.2 synthétise la construction et l'enrichissement de classes conceptuelles présentés dans ce chapitre. L'objectif de l'expérimentation présentée dans cette section est de mesurer la qualité de l'enrichissement des méthodes proposées.

7.2.1 La construction des classes conceptuelles

Afin de construire nos classes conceptuelles en utilisant le modèle SelDe, nous devons disposer d'un corpus. Nous utilisons ainsi un corpus écrit en français. Il est extrait du site Web d'informations de *Yahoo !* (http ://fr.news.yahoo.com) appartenant au domaine "actualités avec un style journalistique". Il contient 8 948 articles. Nous appliquons alors à ce corpus le modèle SelDe avec des paramètres fixés par un expert. Nous avons ainsi sélectionné les couples de verbes ayant obtenu un minimum de 0,8 comme score de proximité avec Asium ($SA = 0,8$). Nous n'avons conservé que les objets communs ayant entre 3 et 9 occurrences dans un couple (*NbOccMin=3* et *NbOccMax=9*). Ce couple peut avoir un maximum de 9 objets communs distincts et un minimum de 3. La sélection suivant les paramètres a fourni une liste de 131 couples de verbes. L'expert a finalement sélectionné parmi cette liste **cinq couples de verbes** qui, avec leurs objets communs respectifs définissent nos concepts initiaux dont les objets communs sont les instances. Ces cinq concepts sont présentés dans le tableau 7.3.

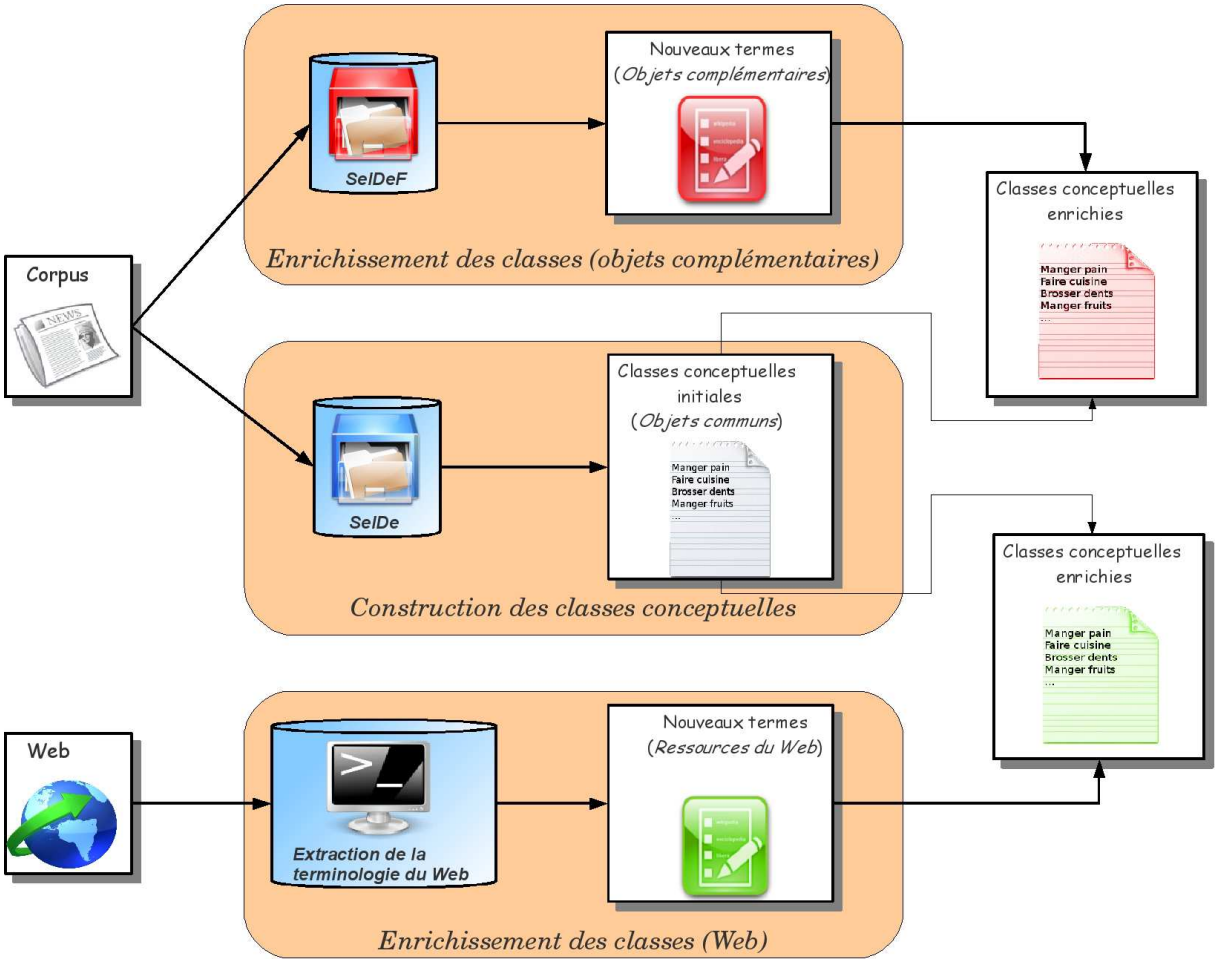

FIG. 7.2 – La construction et l'enrichissement de classes conceptuelles

Concepts	*Organisme /Administration*	*Fonction*	*Objets symboliques*	*Sentiment*	*Manifestation de protestation*
Instances	parquet	négociateur	drapeau	mécontentement	protestation
	mairie	cinéaste	fleur	souhait	grincement
	gendarme	écrivain	spectre	déception	indignation
	préfecture	orateur		désaccord	émotion
	pompier			désir	remous
	onu				tollé
					émoi
					panique

FIG. 7.3 – Les cinq concepts sélectionnés et leurs instances

7.2.2 Enrichissement avec SelDeF

L'enrichissement de classes conceptuelles avec le modèle SelDeF consiste à utiliser les descripteurs de ce modèle comme nouvelles instances des classes conceptuelles formées avec SelDeF. En d'autres termes, nous validons chaque objet complémentaire des couples de verbes constituant les classes conceptuelles, avec les différentes approches de filtrage du modèle SelDeF à savoir l'approche fondée sur les vecteurs sémantique, la validation Web et les combinaisons de ces approches. Nous proposons, afin de mesurer la qualité des objets complémentaires, de suivre le protocole d'évaluation suivant.

7.2.2.1 Protocole d'évaluation

Évaluation manuelle

Nous effectuons dans un premier temps une évaluation manuelle des termes candidats pouvant être retenus comme nouvelles instances de concepts. Nous disposons de huit évaluateurs auxquels nous avons soumis un formulaire. Celui-ci a pour objectif de faire valider manuellement des termes pouvant appartenir à un concept. Pour chacun des cinq concepts extraits, nous soumettons aux évaluateurs les objets candidats, qui ne sont autres que les objets complémentaires de l'un ou l'autre des verbes. L'évaluateur doit alors mesurer la pertinence d'un terme pour un concept donné en respectant le barème suivant :
- 2 : Parfaitement pertinent
- 1 : Susceptible d'être pertinent
- 0 : Non pertinent
- N : Ne se prononce pas

La figure 7.4 présente une capture d'écran du formulaire soumis aux experts.
Nous présentons alors deux variantes permettant d'utiliser les scores attribués par les juges : une *moyenne* des scores obtenus et un système de *votes*.

La moyenne. Après l'évaluation des objets candidats (553 termes) par les experts, nous effectuons une moyenne des résultats en faisant varier la tolérance. Nous distinguons alors différents intervalles afin de considérer un résultat comme pertinent ou non. Par exemple, un terme peut-être pertinent si les experts lui ont attribué en moyenne un score supérieur à 1.
Le vote. Nous proposons de soumettre les scores donnés par les juges à un système de vote. Nous qualifions alors de pertinent un candidat qui a été jugé de bonne qualité par un pourcentage p de juges. Un juge définit un terme comme étant de bonne qualité en lui attribuant par exemple un score supérieur ou égal à 1.

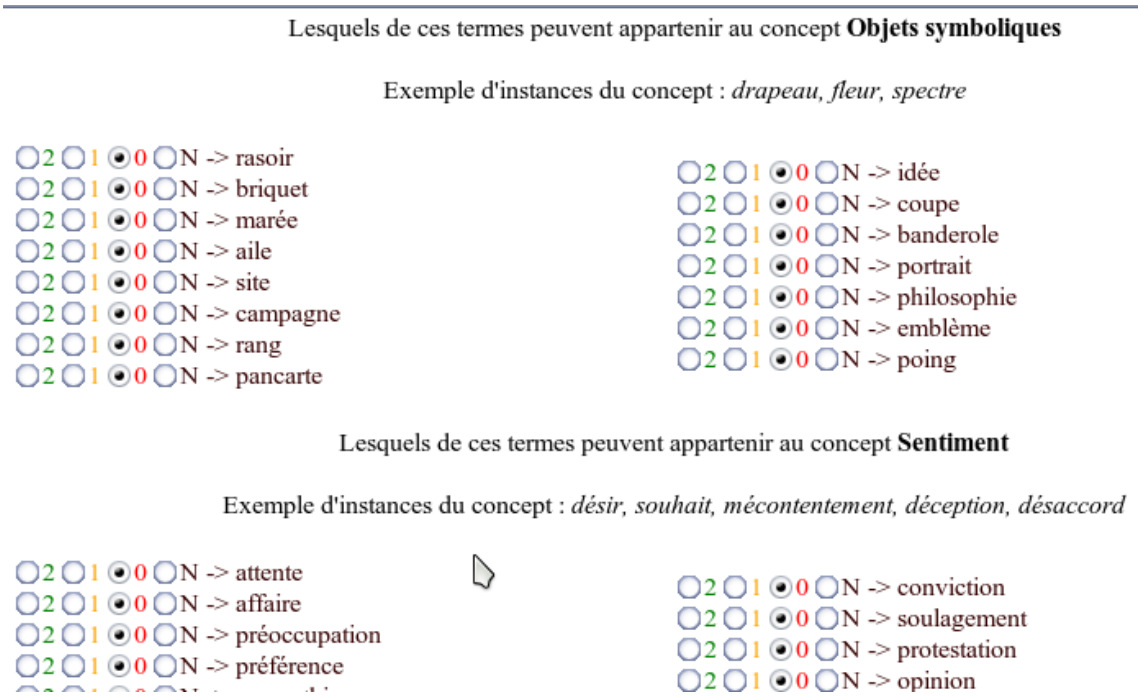

FIG. 7.4 – Capture d'écran du formulaire d'évaluation manuelle

Ce protocole d'évaluation a l'avantage de faire intervenir un ou plusieurs experts garantissant ainsi la qualité de l'évaluation. Cependant, il parait peu raisonnable de faire évaluer par un expert un ensemble de 50 000 termes comme nous le proposons dans la section 7.3. Ainsi, nous présentons ci-dessous une alternative à ce protocole d'évaluation manuel.

Évaluation automatique

Le principe de l'évaluation automatique est d'utiliser un second corpus écrit également en français, de taille plus conséquente que celui d'où proviennent les objets complémentaires à valider. Les deux corpus sont du même domaine. Le corpus utilisé pour nos expérimentations est constitué de plus de 60 000 articles provenant du quotidien *Le Monde* pour une taille de 125 Mo. Nous jugeons alors comme *bien formées* des relations induites qui vont être présentes *nativement* dans le second corpus. Une telle relation sera alors qualifiée de **positive**. À l'inverse, une relation induite non retrouvée dans le second corpus sera qualifiée de **négative**. Notons que les relations jugées négatives peuvent être de faux négatifs. En effet, une relation qui n'a pas été retrouvée dans le second corpus n'est pas pour autant non pertinente. De plus, un objet complémentaire dans une relation syntaxique jugée pertinente peut également ne pas être un "bon" candidat pour un concept. Ainsi, avec un tel protocole, nous pouvons mesurer de manière

automatique la qualité des approches proposées, et ceci pour un très grand nombre de relations syntaxiques.
Une fois la notion de candidats pertinents définie avec les protocoles présentés, nous proposons d'évaluer le classement issu de nos différentes approches en utilisant les courbes ROC. Cette méthode, déjà abordée dans le chapitre 5, est décrite ci-dessous de manière plus précise.

Les courbes ROC

Terme	***Validation Manuelle***
Conviction	+
Opinion	+
Préférence	-
Attente	-
Colère	+

TAB. 7.1 – Exemple de classement de termes du concept "Sentiment"

La méthode des courbes ROC (Receiver Operating Characteristic), détaillée par [Ferri *et al.*, 2002], fut utilisée à l'origine dans le domaine du traitement du signal. Cette méthode est fréquemment employée en médecine afin d'évaluer automatiquement la validité d'un diagnostique de tests. Nous trouvons en abscisse des axes représentant une courbe ROC le *taux de faux positifs* et en ordonnée le *taux de vrais positifs*. Rappelons que dans notre cas le taux de vrais positifs représente pour le protocole automatique le taux de relations syntaxiques retrouvées dans le corpus du *Monde*, et pour le protocole manuel le taux de termes jugés positifs par les experts. La surface sous la courbe ROC ainsi créée est appelée AUC (Area Under the Curve). En outre, cette surface est équivalente au test statistique de Wilcoxon-Mann-Whitney tel que montré dans [Yan *et al.*, 2003]. Notons qu'une courbe ROC représentée par une diagonale correspond à un système où les relations syntaxiques ont une distribution aléatoire, la progression du taux de vrais positifs est accompagnée par la dégradation du taux de faux positifs. Considérons le cas d'une validation de relations syntaxiques induites. Si toutes les relations étaient positives (ou pertinentes), l'AUC vaudrait 1, ce qui signifie avoir toutes les relations pertinentes en début de liste, donc ordonnées de manière optimale. Par ailleurs, un des avantages de l'utilisation des courbes ROC réside dans leur résistance au déséquilibre entre le nombre d'exemples positifs et négatifs comme le montrent par exemple [Fisher *et al.*, 2004].
Le tableau 7.1 présente un exemple de termes ordonnés avec l'approche HYBAD (décrite en section 6.3.3.2) évalués par une validation manuelle pour le concept "Sentiment". La

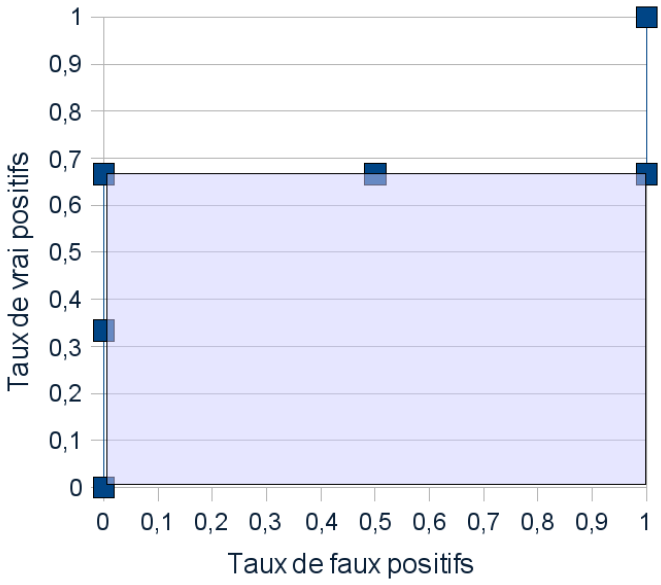

FIG. 7.5 – Courbe ROC résultante de l'exemple présenté dans le tableau 7.1

courbe ROC alors obtenue est présentée dans la figure 7.5. Nous obtenons finalement une AUC (en bleu sur la figure 7.5) de 2/3 avec cet exemple.

7.2.2.2 Résultats expérimentaux

Les critères de la validation manuelle

Les différents critères proposés pour la validation manuelle doivent être fixés afin de pouvoir évaluer la qualité de nos approches de validation. Rappelons tout d'abord que les juges doivent attribuer à chaque terme candidat (objets complémentaires provenant de relations syntaxiques induites) un score de qualité pouvant être 0, 1 ou 2[41]. Ainsi, pour les deux variantes proposées afin de prendre en compte les scores de juges (la moyenne et le système de vote), nous devons fixer les critères qui définiront un terme comme pertinent ou non.
(1) Pour la moyenne, nous devons déterminer le score moyen au delà duquel un candidat sera pertinent.
(2) Le système de vote nécessite quant à lui de fixer deux critères. Déterminer quelle note minimale donnée par un juge qualifiera un terme comme pertinent. Puis nous devons fixer le pourcentage de juges nécessaire afin de considérer un candidat comme pertinent. Finalement, un exemple de critères pour le système de vote peut-être : un terme est pertinent si 75% des juges lui ont attribué une note supérieure ou égale à 1.

[41] Les termes ayant obtenu la note "N" ne seront pas pris en compte.

Pour nos expérimentations, nous avons évalué les critères suivants, en fonction des variantes *moyenne* et *vote* :
– Les scores moyens supérieurs ou égales à 1, 1,5 et 2 ont été testé soit trois variantes.
– Pour le vote, 50 et 75% de juges pour un score égal à 1 ou 2 soit quatre variantes.
Nous présentons dans cette section les résultats obtenus avec les critères de pertinence pour le vote et la moyenne fixés comme suit. Un terme sera pertinent s'il a obtenu une moyenne supérieure à 1,5. Pour le vote, un objet complémentaire sera pertinent si 75% des experts lui ont attribué la note de 2. Les expérimentations menées avec d'autres critères sont présentées en annexe.
Notons que nous nous sommes appuyés sur le nombre de termes pertinents obtenus pour déterminer les critères les plus adaptés à l'obtention de résultats expérimentaux cohérents. Par exemple, en considérant les termes qui ont été jugés pertinents avec un score moyen supérieur ou égal à deux, seul un terme répond à ce critère.

Résultats expérimentaux

L'objectif que nous nous fixons avec l'enrichissement de classes conceptuelles est de réduire la tâche de l'expert en filtrant le nombre de relations syntaxiques induites candidates à un concept. Ceci valide implicitement des objets complémentaires. Les expérimentations ci-dessous ont pour but de montrer dans quelle mesure nos approches de validation automatique sont intéressantes. Ainsi, nous introduisons un **seuil** qui n'est autre que le nombre de relations syntaxiques considérées. Rappelons que les approches de validation du modèle SelDeF fournissent en sortie une liste triée de relations syntaxiques induites. Par exemple, un seuil fixé à 100 indique que l'on ne mesure l'AUC que pour les 100 premières relations syntaxiques triées avec une approche de validation. Nous avons utilisé pour ces expérimentations la première approche des vecteurs sémantiques, comparant la proximité de deux vecteurs non contextualisés. Notons que la seconde approche sera expérimentée dans la section 7.3 avec un nombre plus conséquent de relations syntaxiques induites à valider. La validation par le Web[42] a été quant à elle expérimentée avec les mesures statistiques *Fréquence*, IM, IM^3 et *Dice*. Nous ne montrerons cependant que les résultats obtenus avec la mesure IM, mesure ayant obtenu les meilleurs résultats. Nous employons par ailleurs pour la fonction *nb* l'approche "*somme*" telle que décrite en section 6.3.2.2.

Nous présentons dans un premier temps les scores obtenus pour l'approche des

[42]Précisons qu'il s'agit de la mesure de validation des objets complémentaire utilisant le Web. L'approche utilisant le Web afin de fournir de nouveaux candidats pour les concepts sera présentée en section 7.2.3.

Vecteurs Sémantiques (VS). Cette approche utilisée sans contextualisation donne ici des

AUC	*Protocole de validation utilisé*		
	Manuelle		*Automatique*
Seuil	Moyenne	Vote	
50	0,54	0,50	0,54
100	0,57	0,66	0,54
150	0,52	0,55	0,55
200	0,50	0,57	0,48
250	0,42	0,53	0,52
300	0,34	0,35	0,47
350	0,42	0,42	0,50
400	0,47	0,46	0,51
450	0,46	0,46	0,50
500	0,41	0,41	0,53
550	0,39	0,39	0,55

TAB. 7.2 – AUC obtenues avec les vecteurs sémantiques

résultats non pertinents comme le montre le tableau 7.2. En effet, les AUC obtenues sont proches d'une distribution aléatoire (AUC=0,5). Les évaluations effectuées avec le protocole manuel donnent des résultats assez similaires pour le vote ou la moyenne. Notons que plus nous considérons de relations syntaxiques (seuil qui augmente) et plus les AUC diminuent. Le protocole d'évaluation automatique **se comporte quant à lui de manière similaire au protocole manuel en fournissant des AUC du même ordre**.

Nous présentons ci-après les résultats obtenus avec l'approche *Validation Web* (VW). Le tableau 7.3 montre les AUC obtenues avec la mesure statistique IM. Ces résultats, bien

AUC	*Protocole de validation utilisé*		
	Manuelle		*Automatique*
Seuil	Moyenne	Vote	
50	0,62	0,64	0,59
100	0,53	0,5	0,6
150	0,58	0,62	0,66
200	0,61	0,61	0,65
250	0,57	0,56	0,66
300	0,52	0,51	0,65
350	0,56	0,57	0,67
400	0,58	0,59	0,67
450	0,6	0,61	0,67
500	0,57	0,56	0,68
550	0,53	0,52	0,69

TAB. 7.3 – AUC obtenues avec la Validation Web

que meilleurs que ceux obtenus avec l'approche VS restent assez faibles. En effet, un résultat d'AUC peut être considéré comme pertinent au delà de 0,8 [Lehr & Pong, 2003].

Le protocole d'évaluation automatique se comporte de manière identique par rapport aux expérimentations effectuées avec VW. En effet, au delà d'un seuil de 400 relations, les scores qu'il fournit sont trop éloignés de la validation manuelle. Nous expliquons ces résultats par le fait qu'une telle validation génère un certain nombre de faux négatifs. La présence de ces derniers est moins visible avec les premiers seuils, indiquant que les faux négatifs sont classés en fin de liste par l'approche Validation Web. Ces relations ne sont donc pas populaires sur le Web. Dès lors, il est cohérent de ne pas retrouver ces relations syntaxiques dans le corpus du *Monde* (validation automatique). En effet, il est peut probable qu'une relation non présente sur le Web soit présente dans un corpus, si grand soit il. En outre, un évaluateur humain peut juger ces relations comme bien formées, là où l'approche proposée par le protocole de validation automatique est plus faillible.

Après avoir présenté les résultats de nos deux approches de validation VS et VW, nous proposons de combiner ces dernières. Nous évaluons dans un premier temps la combinaison HyPon (définie en section 6.3.3.1 page 174). Nous présentons dans le tableau 7.4 uniquement les résultats de la combinaison des approche VS et VW (IM) pour un paramètre k valant $0,3$. Nous avons effet obtenu les meilleurs résultats en moyenne avec cette valeur. Les résultats pour d'autres valeurs de k sont présentés en annexes. L'approche HyPon donnent de meilleurs résultats que VS ou VW pour un

AUC	*Protocole de validation utilisé*		
	Manuelle		*Automatique*
Seuil	Moyenne	Vote	
50	0,79	0,78	0,65
100	0,54	0,54	0,65
150	0,69	0,69	0,73
200	0,75	0,74	0,80
250	0,56	0,58	0,66
300	0,49	0,55	0,64
350	0,51	0,53	0,62
400	0,53	0,53	0,64
450	0,52	0,53	0,62
500	0,47	0,46	0,62
550	0,43	0,43	0,62

Tab. 7.4 – AUC obtenues pour l'approche HyPon combinant VS et VW

seuil inférieur ou égal à 200. Le protocole automatique montre également ce même résultat. Ainsi, nous pouvons réduire la tâche d'un expert en proposant 200 candidats afin d'enrichir nos classes conceptuelles. Parmi ces 200 candidats, 75% sont des candidats pertinents.

Nous évaluons finalement la qualité de l'approche HybAd. Les résultats obtenus

sont présentés dans le tableau 7.5. Rappelons que cette approche consiste à classer la totalité des candidats avec l'approche VS puis à ordonner les n premiers candidats avec l'approche VW. Nous fixons pour ces expérimentations le paramètre n à la valeur du seuil considéré. Ainsi, pour un seuil à 200, nous classons toutes les relations avec l'approche VS puis les 200 premières avec VW. Nous obtenons avec cette approche les

AUC	*Protocole de validation utilisé*		
	Manuelle		*Automatique*
Seuil	Moyenne	Vote	
50	0,74	0,81	0,90
100	0,82	0,83	0,87
150	0,79	0,80	0,84
200	0,75	0,76	0,79
250	0,65	0,71	0,75
300	0,68	0,70	0,74
350	0,67	0,69	0,75
400	0,66	0,67	0,74
450	0,63	0,65	0,71
500	0,57	0,57	0,70
550	0,53	0,52	0,69

Tab. 7.5 – AUC obtenues avec HybAd, comparées aux autres approches

meilleurs résultats. En effet, l'évaluation manuelle donne d'excellents résultats pour les premières relations (AUC jusqu'à 0,83). Les résultats sont de qualité moyenne (AUC de 0,70) jusqu'au seuil de 350, pour se dégrader avec la totalité des candidats (AUC proche de l'aléatoire 0,5). Ainsi, cette approche semble être la plus adaptée afin de réduire le nombre de candidats à expertiser. Nous ne pouvons cependant pas fournir à un expert une liste triée de l'ensemble des candidats mais une liste contenant un sous ensemble. Ainsi, nous privilégions la précision et la qualité de la liste fournie à l'expert en réduisant en contre partie le nombre de candidats disponibles initialement (plus faible rappel). Nous constatons par ailleurs que les résultats sont du même ordre pour les deux types de protocoles (trois protocoles au total, les deux manuels et l'automatique). En effet, les résultats de l'approche HybAd sont de très bonne qualité pour les faibles seuils et se dégradent avec la totalité des résultats. Nous montrons alors, sur cet échantillon de candidats, que notre protocole de validation automatique est de bonne qualité. Il permet en effet de montrer que les premières relations sont les mieux classées avec l'approche HybAd. Il reflète également le fait que cette approche fournit les meilleurs classements. Néanmoins, les scores obtenus ont tendance à être surévalués avec le protocole automatique. Ces scores s'expliquent notamment par la diversité des tâches effectuées par les deux protocoles. Le protocole manuel cherche à connaître la pertinence d'un terme dans un concept. Le protocole automatique propose de mesurer la cohérence d'une relation syntaxique formée d'un verbe et d'un objet complémentaire. Ces tâches, bien qu'assez

proches, ne visent pas les mêmes objectifs. Il est en effet plus difficile de mesurer de manière automatique la qualité d'un candidat potentiel à un concept que la qualité d'une relation syntaxique.

Synthèse de la première approche d'enrichissement de classes conceptuelles

Cette section a présenté une approche d'enrichissement de classes conceptuelles. Nous utilisons les descripteurs du modèle SelDeF afin d'obtenir de nouveaux termes pour les classes conceptuelles. Rappelons que ce modèle utilise des informations syntaxiques (les relations syntaxiques de type verbe-objet) et sémantiques (mesure de proximité des verbes). L'approche présentée propose des nouveaux termes de qualité comme nous avons pu le montrer lors de nos expérimentations. Notons cependant que ces termes restent assez dépendants de leur corpus d'origine. En effet, bien que les approches de validation utilisent des connaissances extérieures comme le Web ou un thésaurus, au final, les termes proposés restent des mots provenant du corpus d'origine.
Parmi les approches de validation des relations syntaxiques induites du modèle SelDeF, la validation Web et l'approche HybAd se sont révélées particulièrement efficaces. Nous reviendrons cependant sur la qualité de ces approches en effectuant des expérimentations plus poussées avec un nombre plus conséquent de relations induites dans la section 7.3.
Nous avons finalement proposé un protocole d'évaluation automatique afin de mesurer la qualité des relations syntaxiques induites. Ce dernier semble de bonne qualité pour la moitié des relations syntaxiques, et se dégrade pour les suivantes. Notons cependant que la qualité de ce protocole a été évaluée comparativement aux résultats fournis par le protocole manuel. Plusieurs points expliquent ces résultats.
– La présence de faux négatifs.
– La diversité des tâches des deux protocoles.
– La subjectivité humaine.
Nous présentons dans la section suivante une autre approche spécifique à l'enrichissement de classes conceptuelles. Cette dernière n'utilise pas le modèle SelDeF. Nous cherchons dès lors à produire des termes étant moins dépendant d'un corpus.

7.2.3 Le modèle d'enrichissement fondé sur le Web

Avec notre précédente méthode d'enrichissement, nous utilisons les informations d'un corpus afin de proposer de nouveaux termes pour enrichir des concepts. Une telle approche utilise des connaissances spécifiques aux corpus pour enrichir les concepts. Nous présentons dans cette section une autre approche d'enrichissement fondée sur le Web utilisant des ressources de domaines plus généraux que celles d'un corpus.

7.2.3.1 L'acquisition de nouveaux termes

L'objectif de cette méthode est de fournir de nouveaux candidats aux concepts formés tel que décrit en section 7.1.3. Elle se fonde sur l'énumération de termes sémantiquement proches présents sur le Web. Par exemple, en saisissant dans un moteur de recherche la requête (chaîne de caractères) "lundi, mardi et", nous obtenons d'autres jours de la semaine en résultats.

Afin d'appliquer cette méthode, nous considérons dans un premier temps les objets communs des verbes jugés sémantiquement proches. Ils constituent les instances de références des classes ainsi formées. Nous proposons alors d'utiliser le Web afin d'acquérir de nouveaux candidats. Cette méthode présente l'avantage de ne plus se limiter aux termes du corpus dont les classes conceptuelles sont issues.
Considérons alors les N concepts $C_{i\in\{1,...,N\}}$ et leurs instances respectives $I_j(Ci)$. Pour chaque concept Ci nous soumettons alors à un moteur de recherche les requêtes suivantes : "$I_{jA}(Ci)$, $I_{jB}(Ci)$ et" et "$I_{jA}(Ci)$, $I_{jB}(Ci)$ ou" avec j_A et $j_B \in \{1, ..., NbInstanceCi\}$ et $j_A \neq j_B$. Plus concrètement avec l'exemple de la figure 7.1, nous fournissons au moteur de recherche les requêtes : "drapeau, fleur et", "drapeau, fleur ou", "fleur, drapeau et", "fleur, drapeau ou".
Le moteur de recherche nous retourne alors un ensemble de résultats desquels nous extrayons de nouveaux candidats à un concept. Après avoir identifié la requête dans nos résultats, le terme qui suit notre requête constitue une nouvelle instance du concept, tel qu'illustré dans l'exemple suivant.

Soit la requête : "drapeau, fleur et", le moteur nous retourne alors :
"*Tu joues version normale (Carreau, pique, cœur et trèfle) ou version bourbi... heu... suisse-allemande (Gland,* ***Drapeau, Fleur et*** *Grelot)*".
Après avoir identifié notre requête dans le résultat retourné (en gras sur notre exemple), nous ajoutons au concept le terme suivant directement la requête (ici, le terme Grelot). Nous obtenons ainsi de nouveaux candidats pour nos concepts. Afin de pouvoir réitérer l'opération permettant d'acquérir de nouveaux termes, nous devons filtrer les candidats précédemment obtenus par cette approche tel que nous le montrons dans la section suivante.

7.2.3.2 Le filtrage des candidats

L'objectif de cette section est de conserver uniquement les instances cohérentes de chaque concept, suite à une acquisition de nouveaux termes. Notre objectif est alors de favoriser la précision de l'approche au détriment du rappel. Rappelons en effet que

notre but est l'acquisition de nouveaux termes via ces termes filtrés. Dans ces conditions, nous avons tout intérêt à ce que les termes servant à l'interrogation du Web soient les "meilleurs" possibles. Ainsi, nous pouvons juger la qualité de ces termes en les faisant valider par un expert, ou bien en utilisant une approche automatique. Cette dernière utilise à nouveau le Web afin de sélectionner les termes pertinents. L'approche est assez similaire à celle permettant l'acquisition de termes. Le principe est de considérer un terme pertinent s'il existe sur le Web une énumération contenant ce terme parmi d'autres déjà considéré comme pertinents. Ainsi, nous cherchons à valider un terme "en contexte". De manière formelle, cette approche peut être définie de la manière suivante. Soit N concepts $C_{i\in\{1,N\}}$, leurs instances respectives $I_j(C_i)$ et les nouveaux candidats pour un concept C_i, $N_{i_k\in\{1,NbNI(C_i)\}}$.
Pour chaque concept C_i et pour chaque nouveau candidat N_{i_k} sont soumises à un moteur de recherche Web les requêtes :

– "$I_{jA}(C_i), I_{jB}(C_i)$ et N_{i_k}"
– "$I_{jA}(C_i), I_{jB}(C_i)$ ou N_{i_k}"
– "$I_{jA}(C_i), I_{jB}(C_i), N_{i_k}$"

– "$I_{jA}(C_i)$ et $N_{i_k}, I_{jB}(C_i)$"
– "$I_{jA}(C_i)$ ou $N_{i_k}, I_{jB}(C_i)$"
– "$I_{jA}(C_i), N_{i_k}, I_{jB}(C_i)$"

– "$N_{i_k}, I_{jA}(C_i)$ et $I_{jB}(C_i)$"
– "$N_{i_k}, I_{jA}(C_i)$ ou $I_{jB}(C_i)$"
– "$N_{i_k}, I_{jA}(C_i), I_{jB}(C_i)$"

Ainsi, nous soumettons à un moteur de recherche par le biais de chaînes de caractères, des requêtes composées du nouveau candidat et deux à deux les termes de ce concept. Les trois termes sont alternativement séparés par une virgule[43], le terme "ou" et le terme "et". Le moteur de recherche retourne pour chaque requête le nombre de résultats obtenus. Alors, la somme de ces résultats constitue le score de pertinence de ce terme. Nous présentons ci-dessous un exemple de filtrage de termes.

Nous utilisons pour cet exemple le concept "Organisme et Administration" avec les instances définies dans le tableau 7.3. Un des candidats obtenu avec l'approche d'acquisition de nouveaux termes est le terme "*police*". Les instances originales de ce

[43]Notons que les virgules sont automatiquement éliminées par les moteurs de recherche. Ces dernières sont utilisées uniquement pour rendre les requêtes plus lisibles.

concept sont les termes "*mairie, préfecture, pompier, ONU, gendarme*". Ainsi, nous interrogeons le Web avec les requêtes suivantes :

– "parquet mairie et police" : 0 résultat
– "parquet mairie ou police" : 0 résultat
– "parquet mairie police" : 0 résultat
– "parquet police et mairie" : 0 résultat
– "parquet police ou mairie" : 0 résultat
– "parquet police mairie" : 0 résultat
– "police mairie et parquet" : 0 résultat
– "police mairie ou parquet" : 0 résultat
– "police mairie parquet" : 0 résultat
– "parquet ONU et police" : 0 résultat
...
– "mairie préfecture police" : 29 résultats
etc.
Nous effectuons alors la somme de tous les résultats obtenus. Il en résulte un total de *8510* résultats.

Après avoir effectué ce filtrage pour chaque nouveau terme candidat, nous classons ceux-ci par pertinence (par classe). Alors, le filtrage de candidats consiste à ne sélectionner que les ***n* premiers candidats** par classe afin d'effectuer une nouvelle acquisition de termes. Ainsi, nous réitérons cette acquisition en effectuant les nouvelles requêtes incluant les nouveaux termes. Cette opération d'*acquisition/filtrage* peut alors être répétée un certain nombre de fois suivant le nombre d'instances de concepts finalement désiré. Nous présentons dans la section suivante les expérimentations que nous avons menées afin de mesurer la qualité de cette approche.

7.2.3.3 Expérimentations

Protocole expérimental

Nous avons expérimenté cette seconde méthode d'acquisition de termes afin d'enrichir nos cinq concepts déjà expérimentés dans la section 7.2.2.2. Nous utilisons pour nos expérimentations l'API du moteur de recherche *Yahoo !* afin d'obtenir nos nouveaux termes.
Nous appliquons les **post-traitements** suivants à chaque nouveau terme candidat. Ils sont dans un premier temps lemmatisés. Nous supprimons alors les termes candidats

déjà présents dans le concept. En effet, il peut arriver qu'un terme déjà présent dans une classe soit proposé par l'approche d'acquisition de termes. Dès lors, nous ne conservons que les *noms*, après avoir étiqueté les termes avec l'outil TreeTagger déjà présenté en section 5.2.2. Finalement, nous supprimons des "stop words" qui ne sont autres que des termes génériques comme des pronoms, articles ou des noms et verbes fréquents. (par exemple *être, avoir, faire*, etc.).
Après avoir appliqué ces différents post-traitements, nous faisons valider manuellement les nouveaux termes par trois experts. Nous leurs soumettons la question suivante. *Ce terme est-il une bonne instance pour ce concept* ? Nous calculons alors la précision de notre approche pour chaque juge et faisons la moyenne de ces dernières, définissant la qualité des termes. La précision est définie comme suit.

$$\text{Précision} = \frac{\text{Nombre de termes positifs parmi les candidats}}{\text{Nombre de candidats}} \tag{7.1}$$

Ainsi, pour chaque "passe" *acquisition/filtrage*, nous faisons valider aux experts les termes qu'ils jugent pertinents.
Nous appliquons finalement notre filtrage présenté en section 7.2.3.2 en sélectionnant les **quatre termes les mieux classés**. Ainsi, nous pouvons évaluer la qualité de l'approche d'enrichissement de classe conceptuelle ainsi que celle du filtrage automatique.
Notons que nous distinguons lors de nos expérimentations deux types de filtrage.
– Un filtrage automatique, effectué par notre approche présentée en section 7.2.3.2. Nous ne conservons ainsi que quatre termes par classe comme nous l'avons précédement évoqué.
– Un filtrage manuel. Ce dernier consiste à ne retenir que les termes jugés pertinents par les experts.
Une fois l'un des filtrages appliqué, nous réitérons l'étape d'acquisition de nouveaux termes.

Résultats expérimentaux

Le tableau 7.6 présente les résultats obtenus avec une validation automatique des termes. En d'autres termes, à chaque étape d'acquisition de nouveaux termes (une passe), nous appliquons notre approche de filtrage des candidats. Le tableau présente pour chaque passe la précision obtenue après expertise pour :
– **Tous les candidats**. Nous calculons alors la précision avant que l'étape de filtrage automatique soit appliquée.
– **Des candidats filtrés**. Après avoir appliqué le filtrage automatique qui rappelons le revient à sélectionner quatre termes par classe, nous recalculons la précision obtenue.

Notons que le filtrage automatique va réduire le nombre de termes proposés, et donc réduire le rappel[44].
Nous indiquons finalement dans le tableau 7.6 le nombre de termes générés par l'acquisition via le Web pour chaque passe. Nous remarquons avec ces résultats qu'un

	Précision		*Nombre de termes*
Passe #	*Tous*	*Termes filtrés*	*(hors filtre)*
1	0,69	0,83	29
2	0,69	0,77	47
3	0,56	0,65	103

TAB. 7.6 – Résultats obtenus avec le filtrage automatique

nombre important de termes est généré avec l'étape d'acquisition (103 termes pour la troisième passe). Cette approche est donc efficace afin de produire de nouveaux termes enrichissant des classes conceptuelles. La précision obtenue lors des deux premières passes est de bonne qualité, une fois les termes filtrés. La troisième passe donne quant à elle un score plus faible. En effet, globalement, ces termes obtiennent une précision de 0,56. Ainsi, même en ne sélectionnant que les 4 termes les plus pertinents pour chaque classe, du bruit se voit introduit. Nous n'avons alors pas régénéré de nouvelles passes, la précision des termes filtrés étant trop faible pour générer de nouveaux termes de qualité.

Après avoir effectué une validation automatique des termes avec notre approche de filtrage et en avoir montré les limites, nous proposons de faire valider pour chaque passe nos termes par nos experts. Ce filtrage consiste à ne retenir que les termes jugés pertinents par les experts (cf section précédente). Notons que le calcul de la précision après ce filtrage sera toujours égal à 1 car il s'agit précisément de la sélection de juges.
Les résultats obtenus sont présentés dans le tableau 7.7. Ce tableau montre uniquement la précision obtenue après génération de nouveaux termes. Cette validation manuelle

	Précision	*Nombre de termes*
Passe #	*Sélection manuelle*	
1	0,69	29
2	0,70	40
3	0,76	93

TAB. 7.7 – Résultats obtenus avec le filtrage des experts

permet la génération de nouveaux termes de qualité. En effet, la précision obtenue reste

[44]Le rappel n'a pas été calculé au cours de ces expérimentations. Il n'a en effet pas de sens car il vaut toujours 1 sans filtrage et est automatiquement plus faible avec.

assez constante et semble même augmenter. Nous passons en effet de 0,69 à 0,76 ce qui nous indique la bonne qualité de l'approche d'acquisition de termes, une fois les nouveaux termes correctement filtrés. Notons enfin que les précisions obtenues lors des deux premières passes sont assez équivalentes à celles obtenues avec l'approche automatique.

Synthèse de la seconde approche d'enrichissement de classes conceptuelles

Cette section a présenté une approche d'enrichissement de classes conceptuelles fondée sur le Web. Nous appliquons en effet le principe de l'énumération afin d'obtenir via un moteur de recherche Web de nouveaux "candidats" aux concepts déjà existants. Cette approche a l'avantage d'être moins dépendante du corpus d'origine que la première approche d'expansion. Notons que l'utilisation du Web implique une validation des candidats. Ainsi, nous avons proposé une méthode de filtrage de termes qui reste cependant à perfectionner. Ainsi, la sélection des termes par un ou plusieurs experts semble le choix le plus adapté en l'état actuel de l'approche de validation automatique.

7.2.4 Synthèse

7.2.4.1 Les approches d'enrichissement

Cette première partie expérimentant les descripteurs fournis par le modèle SelDeF a consisté à construire et à enrichir des classes conceptuelles. Nous avons pour cela utilisé les descripteurs fournis par SelDe afin de construire les classes. Un couple de verbes proches et ces objets communs définissent une classe et ses instances. Deux approches d'enrichissement de classes furent présentées.
Une première utilisant les descripteurs du modèle SelDeF. Ainsi, les objets complémentaires des couples de verbes formant les classes sont considérés. Dès lors, nous appliquons différentes approches de validation afin d'ordonner en termes de qualité les relations induites. Ces dernières seront alors sélectionnées afin d'enrichir les classes conceptuelles. La seconde approche d'enrichissement propose quant à elle d'utiliser les ressources du Web afin d'enrichir les classes. Cette approche a également donné de bons résultats en faisant valider chaque génération de nouveaux candidats par un expert.
Alors, les applications de ces approches d'enrichissement peuvent être les suivantes.
– *Approche* SelDeF. Enrichissement de classes conceptuelles très spécialisées.
– *Approche Web*. Enrichissement de classes conceptuelles spécialisées ou non. En effet, avec cette dernière approche, l'utilisation d'un corpus n'est en rien obligatoire. Les termes initiaux constituant les classes conceptuelles originales peuvent en effet être proposés par un expert.

7.2.4.2 Analyse des résultats

Les résultats obtenus avec nos deux approches d'enrichissement sont de bonne qualité. Ils peuvent cependant être améliorés. En effet, la thématique même du concept présenté aux experts est discutable car elle est établie manuellement. Rappelons en effet que les concepts formés par les objets communs des couples de verbes jugés proches sont définis par les experts. Alors la subjectivité de l'évaluation humaine joue un rôle non négligeable dans cette évaluation. Citons par exemple le concept "*manifestation de protestation*". La question posée est alors : le terme "*adhésion*" est-il une instance correcte de ce concept ? Une définition du terme adhésion (provenant du *TLFI*) est : "*Reconnaissance implicite ou explicite de l'autorité d'une loi, d'un gouvernement, etc.*". D'une manière triviale en se fondant sur cette définition, nous dirions que ce terme appartient à un concept de sens opposé. Mais si nous considérons une adhésion comme un engagement politique ou associatif s'opposant aux règles ou aux lois établies, il peut être perçu comme un moyen de protestation. Cette subjectivité humaine pose là une question importante au niveau de la qualité de l'évaluation humaine pour des systèmes de fouilles de textes.

7.2.4.3 Exemple de classe enrichie

Nous proposons en conclusion à cette section de présenter une synthèse des deux approches d'enrichissement des concepts avec un exemple concret.
Nous allons nous appuyer sur le concept "*Sentiment*". Ce concept a été formé par les objets communs de deux verbes jugés proches : *exprimer* et *manifester*. Les instances de ce concept sont : *désir, souhait, mécontentement, déception, désaccord*. Nous proposons alors d'utiliser les deux méthodes présentées dans ce chapitre afin d'enrichir ce concept.

Première méthode : Induction d'informations provenant du corpus.
Nous générons une liste d'objets dits complémentaires à partir du corpus. Ces objets vont alors être ordonnés avec la seconde combinaison des approches *Validation Web* et *Vecteurs Sémantiques*. Nous les soumettons alors à un expert qui va sélectionner les plus pertinentes. Les objets retenus par l'expert sont :
"*sympathie, regrets, doute, crainte, exaspération, satisfaction, sensibilité, espoir, indignation, dédain, joie, amertume, désarroi, solidarité, confiance, colère*".

Seconde méthode : Enrichissement via le Web.
La seconde méthode propose d'utiliser des ressources extérieures généralisant ainsi notre concept original. Elle se fonde sur l'envoi de requêtes à un moteur de recherche. Les candidats obtenus par le Web après validation de l'expert sont (pour deux passes validées manuellement) :

"*horreur, satisfaction, déprime, faiblesse, tristesse, désenchantement, folie, amusement, mal, souffrance, enthousiasme, chagrin, passion, fatalisme, angoisse, inconscience*".

7.3 Expérimentations avec un grand nombre de relations induites

Les expérimentations précédentes ont évalué la qualité du modèle SelDeF et *a fortiori* des approches de validation de relations syntaxiques induites. Ces expérimentations ont porté sur 550 relations induites. L'objectif était de proposer une application concrète de création et d'enrichissement de classes conceptuelles. Cependant, lors de la construction de telles classes, nous pouvons être amenés à travailler sur un nombre bien plus conséquent de classes et d'instances de celles-ci. Un expert pouvant en effet souhaiter construire une classification conceptuelle plus fournie. Ainsi, nous présentons dans cette section une seconde expérimentation fondée sur le modèle SelDeF. Nous nous concentrons uniquement dans cette section sur les approches de validation du modèle SelDeF. L'enrichissement par le Web est en effet une approche récente méritant d'être approfondie. De plus, les expérimentations à grande échelle ne sont pour le moment pas réalisables. Nous ne pouvons en effet pas faire expertiser la qualité de nos approches autrement que manuellement ce qui engendre un coût non négligeable.
Ainsi, nous cherchons à savoir si, au delà de la construction et l'enrichissement de classes conceptuelles, les approches de validation du modèle SelDeF sont pertinentes. Nous pouvons en effet supposer diverses applications utilisant des ressources terminologiques ainsi extraites comme par exemple améliorer l'approche ExpLSA présentée dans le chapitre 4. Cette section est organisée comme suit. Nous présentons dans un premier temps les données et le protocole suivi afin de mener ces expérimentations. Nous présentons alors dans la section 7.3.2 les résultats expérimentaux obtenus. Nous discutons pour finir les résultats obtenus dans la section 7.3.3.

7.3.1 Démarche expérimentale

7.3.1.1 Description des données

Afin de mener ces expérimentations, nous utiliserons le même corpus que précédemment expérimenté. Ce corpus extrait du site Web d'informations de *Yahoo !* compte un total de 8948 articles. Nous allons alors valider toutes les relations induites obtenues à partir de ce corpus, pour un seuil d'Asium supérieur à 0,8 (cf chapitre 3)). Notons que les couples de verbes sémantiquement proches extraits n'ont pas été expertisés. Ainsi, aucun filtrage ou traitement des objets communs ne sera effectué dans cette section, cela n'étant

pas l'objet de ce chapitre. Avec $SA = 0,8$, nous avons alors "créé" 50 000 relations induites. Il est donc nécessaire de valider 50 000 objets complémentaires avec nos approches de validation.

7.3.1.2 Les différentes variantes des approches de validation de SelDeF

Nous proposons d'évaluer lors de nos expérimentations les approches du modèle SelDeF à l'instar des précédentes expérimentations effectuées. Néanmoins, nous proposons d'évaluer les deux approches de vecteurs sémantiques, les quatre mesures statistiques de la validation Web et finalement les deux combinaisons HyPon et HybAd. Un certain nombre de points concernant les approches de validation vont être abordés lors des expérimentations proposées dans cette section.
Nous expérimenterons les deux types de vecteurs sémantiques, contextualisés ou non, ainsi que les deux mesures de similarité cosinus et distance de concordance. Nous utiliserons avec la validation Web les mesures statistiques "*fréquence*", IM, IM^3 et $Dice$. De plus, nous expérimenterons la somme et le maximum avec la fonction nb.

7.3.1.3 Le protocole expérimental

Se pose alors le problème de l'évaluation. Une évaluation humaine n'est en effet pas concevable à grande échelle, cette dernière étant trop coûteuse en termes de temps. Rappelons également que les approches manuelles et automatiques d'évaluation donnaient en général des résultats du même ordre lors de nos précédentes expérimentations (section 7.2.2.2). Ceci renforce la validité des résultats obtenus par le protocole automatique. Nous proposons alors d'évaluer la qualité des approches de validation avec **le protocole d'évaluation automatique** décrit en section 7.2.2.1. Nous rappelons brièvement ce dernier. Le principe est d'utiliser un second corpus de taille plus conséquente que celui d'où proviennent les relations induites. Les deux corpus doivent être du même domaine. Nous jugeons alors comme bien formées des relations induites qui vont être présentes nativement dans le second corpus. Une telle relation sera alors qualifiée de positive. Alors, nous utilisons les courbes ROC afin de mesurer la qualité des listes triées avec les approches de validation. Nous calculons l'aire sous cette dernière (AUC) afin d'obtenir un indice de qualité pour une approche de validation. Rappelons que notre objectif est d'obtenir en tête de liste tous les objets complémentaires pertinents. Ainsi, afin de mesurer de manière plus précise la qualité des approches de validation de SelDeF, nous proposons de calculer les AUC pour différents *seuils* comme effectué lors des précédentes expérimentations (section 7.2.2.1).

7.3.2 Résultats expérimentaux

Les expérimentations sont organisées de la manière suivante. Nous évaluons dans un premier temps la qualité de l'approche fondée sur les vecteurs sémantiques. Puis nous nous intéressons aux approches de validation, afin de présenter pour finir les résultats des combinaisons de ces approches.

7.3.2.1 Les vecteurs sémantiques

Nous montrons dans le tableau 7.8 les résultats obtenus avec la première approche des vecteurs sémantiques. Nous utilisons afin de mesurer la proximité sémantique des vecteurs les mesures *cosinus* et *distance de concordance*. Deux paramètres doivent être fixés afin d'utiliser cette dernière. Ainsi, nous avons sélectionné la moitié des composantes du vecteurs (paramètre s=2) et présentons des variantes pour le paramètre w, qui rappelons le, pondère l'importance de la concordance vis-à-vis de la distance angulaire (cf section 6.3.1 page 156). Les résultats obtenus sont cependant assez faibles, quel que

	Cosinus	***Distance concordance*** *(s=2)*			
Seuil		*w=0*	*w=0.25*	*w=0.5*	*w=0.75*
5000	0,510	0,505	**0,530**	0,507	0,502
10000	0,520	0,546	**0,558**	0,549	0,535
15000	0,544	0,551	**0,577**	0,555	0,548
20000	0,543	0,539	0,544	**0,574**	0,562
25000	0,553	0,440	0,507	0,539	**0,556**
30000	**0,547**	0,440	0,505	0,517	0,535
35000	**0,546**	0,449	0,490	0,514	0,531
40000	**0,558**	0,454	0,490	0,525	0,540
45000	**0,551**	0,462	0,504	0,531	0,543
50000	**0,544**	0,468	0,511	0,528	0,535

TAB. 7.8 – AUC obtenues pour les vecteurs sémantiques "simples"

soit le seuil ou bien le type de mesure utilisé. Ces résultats sont similaires à ceux obtenus lors des expérimentations effectuées avec un nombre réduit de relations induites. Nous utilisons au cours de ces expérimentations des vecteurs sémantiquement "pauvres". Ce type de vecteurs peut expliquer ces faibles scores.

Nous proposons alors d'expérimenter la seconde approche utilisant les vecteurs sémantiques contextualisés globaux. Notons que nous ne présentons ici que les meilleurs résultats obtenus à savoir ceux décrivant le verbe par lui même et l'objet complémentaire par le groupe nominal d'où il est issu. Les *AUC* obtenues avec notre protocole d'évaluation automatique sont présentées dans le tableau 7.9. Notons que nous avons conservé les mêmes valeurs pour les paramètres de la distance de concordance. Les résultats montrent

V-GN	***Cosinus***	***Distance concordance*** *(s=2)*			
Seuil		*w=0*	*w=0.25*	*w=0.5*	*w=0.75*
5000	0.451	0.517	0.518	0.520	**0.527**
10000	0.502	0.518	**0.519**	0.516	0.517
15000	0.510	**0.533**	0.532	0.532	0.531
20000	0.501	0.548	0.547	**0.551**	0.551
25000	0.506	0.572	**0.574**	0.573	0.572
30000	0.512	0.589	0.590	**0.591**	0.589
35000	0.550	**0.606**	0.605	0.606	0.604
40000	0.578	0.620	**0.620**	0.620	0.617
45000	0.596	**0.635**	0.634	0.634	0.631
50000	0.603	0.651	**0.651**	0.651	0.649

TAB. 7.9 – AUC obtenues pour les vecteurs sémantiques contextualisés

des scores plus faibles avec cette deuxième approche pour les premiers seuils (jusqu'à 30000). Au delà, les scores sont supérieurs à ceux obtenus avec la première approche, allant jusqu'à atteindre 0,65. Ces AUC restent cependant assez faibles, ne permettant pas de fournir des relations syntaxiques induites correctement filtrées. **La contextualisation des vecteurs sémantiques rend ces derniers plus riches** mais ils restent limités aux concepts du thésaurus. En effet, les 873 concepts définissant ces vecteurs ne sont pas assez discriminants, notre corpus utilisant un vocabulaire assez hétérogène (corpus d'actualités). Cette faible dimension des vecteurs sémantiques ne permet donc pas de classer assez finement nos relations syntaxiques induites. Nous remarquons pour finir que la distance de concordance obtient de meilleurs résultats avec cette approche, ce qui confirme que les vecteurs contextualisés sont plus riches mais néanmoins pas assez discriminants.

7.3.2.2 La validation Web

Les résultats obtenus avec la validation Web sont présentés dans les tableaux 7.10 et 7.11. Ils résultent des expérimentations effectuées respectivement avec la fonction nb "maximum" et "somme". Ces deux approches donnent sensiblement les mêmes résultats. L'approche "somme" donnant cependant des résultats légèrement supérieurs en moyenne. Les scores obtenus avec cette approche sont de bonne qualité pour les seuils supérieurs à 25 000. En d'autres termes, cette approche ne permet pas de fournir un classement correct des premières relations syntaxiques, mais seulement de la totalité, privilégiant ainsi le rappel. Cela reste problématique suivant la tâche désirée. L'enrichissement de classes conceptuelles, comme expérimenté précédemment, nécessite par un exemple un nombre limité de termes afin de les faire expertiser. Par ailleurs, les mesures statistiques ayant le meilleur comportement sont la *fréquence* et l'*information mutuelle au cube*, quel que soit le type de fonction nb utilisé. Cette dernière mesure (IM^3) donne globalement des

maximum	***mesure statistique***			
Seuil	*Freq.*	*IM*	*IM³*	*Dice*
5000	0,615	0,617	0,615	0,623
10000	0,651	0,636	0,656	0,659
15000	0,678	0,668	0,680	0,684
20000	0,709	0,678	0,709	0,706
25000	0,726	0,693	0,731	0,718
30000	0,743	0,703	0,746	0,734
35000	0,756	0,710	0,758	0,748
40000	0,769	0,720	0,770	0,758
45000	0,784	0,724	0,782	0,767
50000	0,800	0,744	0,797	0,784

TAB. 7.10 – AUC obtenues pour la validation Web avec *nb_max*

somme	***mesure statistique***			
Seuil	*Freq.*	*IM*	*IM³*	*Dice*
5000	0.614	0.621	0.608	**0.628**
10000	0.650	0.641	**0.661**	0.658
15000	0.683	0.667	0.685	**0.687**
20000	0.708	0.681	**0.711**	0.708
25000	0.728	0.697	**0.732**	0.719
30000	0.744	0.708	**0.748**	0.737
35000	0.758	0.714	**0.760**	0.748
40000	0.772	0.722	**0.773**	0.759
45000	**0.786**	0.728	0.785	0.770
50000	**0.802**	0.747	0.800	0.786

TAB. 7.11 – AUC obtenues pour la validation Web avec *nb_somme*

résultats légèrement meilleurs que la fréquence. Ainsi, l'information la plus utilisée lors de la validation Web semble être celle fournie par la requête contentant la relation syntaxique induite. Le contexte dans lequel cette dernière apparait dégrade en effet les résultats (ce dernier est accentué avec les mesures Dice et IM). À partir de nos tâches, ces résultats vont à l'encontre de ceux présentés par [Cilibrasi & Vitanyi, 2007] dont l'approche "*Normalized Google Distance*" est résumée en section 6.3.2.1.

7.3.2.3 Les combinaisons

Nous proposons alors d'évaluer la qualité des approches HyPon et HybAd, afin de combiner nos deux précédentes approches de validations. Afin d'effectuer ces combinaisons, nous avons sélectionné les meilleures variantes de chaque approche. Pour les vecteurs sémantiques, nous avons opté pour l'approche contextualisée avec une distance de concordance ($s = 2$ et $w = 0,5$). Nous avons sélectionné l'approche IM^3 avec *nb_somme* pour la validation Web. Les résultats obtenus avec HyPon pour un paramètre k variant

HyPon	***valeur du paramètre k***										
Seuil	*0 (VW)*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10 (VS)*
5000	0.608	0.620	0.643	0.666	0.654	0.674	**0.722**	0.696	0.660	0.615	0.520
10000	0.661	0.666	0.666	0.680	**0.691**	0.637	0.586	0.571	0.546	0.539	0.516
15000	0.685	0.691	**0.702**	0.697	0.668	0.630	0.596	0.574	0.558	0.549	0.532
20000	0.711	0.721	**0.722**	0.707	0.684	0.649	0.617	0.588	0.572	0.561	0.551
25000	0.732	**0.741**	0.738	0.724	0.698	0.667	0.634	0.611	0.598	0.585	0.573
30000	0.748	**0.756**	0.754	0.740	0.718	0.687	0.656	0.628	0.608	0.599	0.591
35000	0.760	**0.767**	0.764	0.752	0.732	0.707	0.677	0.649	0.625	0.611	0.606
40000	0.773	**0.776**	0.774	0.764	0.747	0.725	0.700	0.673	0.648	0.627	0.620
45000	0.785	**0.786**	0.782	0.773	0.759	0.739	0.717	0.694	0.672	0.647	0.634
50000	**0.800**	0.799	0.795	0.787	0.773	0.756	0.737	0.716	0.695	0.673	0.651

Tab. 7.12 – AUC obtenues avec HyPon

par pas de 0,1 sont présentés dans le tableau 7.12. Notons que les résultats avec k=0 et k=1 sont respectivement ceux obtenus avec l'approche Validation Web (VW) et Vecteurs Sémantiques (VS). Les résultats obtenus sont significativement meilleurs pour de faibles seuils, pour k variant de 2 à 4. Par exemple, pour $k = 4$ et un seuil de 5000, l'AUC est améliorée de 5 points avec un score de 0,69. Notons cependant que cette dernière ne reste pas suffisante. Nous expérimentons alors l'approche HybAd dont les résultats sont présentés dans le tableau 7.13. Nous obtenons avec l'approche HybAd de meilleures AUC qu'avec les précédentes approches, quel que soit le seuil testé. Les améliorations sont encore plus significatives pour les premiers seuils. En effet, pour un seuil de 5 000, l'AUC passe de $0,61$ avec la validation Web à $0,81$ avec HybAd. Cette combinaison est l'approche fournissant de meilleurs résultats afin de répondre à notre problématique, la validation automatique des relations syntaxiques induites. Notons également que **les**

	Approches utilisées			
Seuil	*VS*	*VW*	*HyPon*	*HybAd*
5000	0.520	0.608	0.643	**0.813**
10000	0.516	0.661	0.666	**0.795**
15000	0.532	0.685	0.702	**0.788**
20000	0.551	0.711	0.722	**0.790**
25000	0.573	0.732	0.738	**0.789**
30000	0.591	0.748	0.754	**0.791**
35000	0.606	0.760	0.764	**0.793**
40000	0.620	0.773	0.774	**0.796**
45000	0.634	0.785	0.782	**0.799**
50000	0.651	0.800	0.795	**0.804**

TAB. 7.13 – AUC obtenues avec HYBAD, comparées aux autres approches

scores obtenus avec HybAd ne sont pas dépendants du choix du seuil car les AUC restent relativement constantes (AUC variant de $0,79$ à $0,81$).

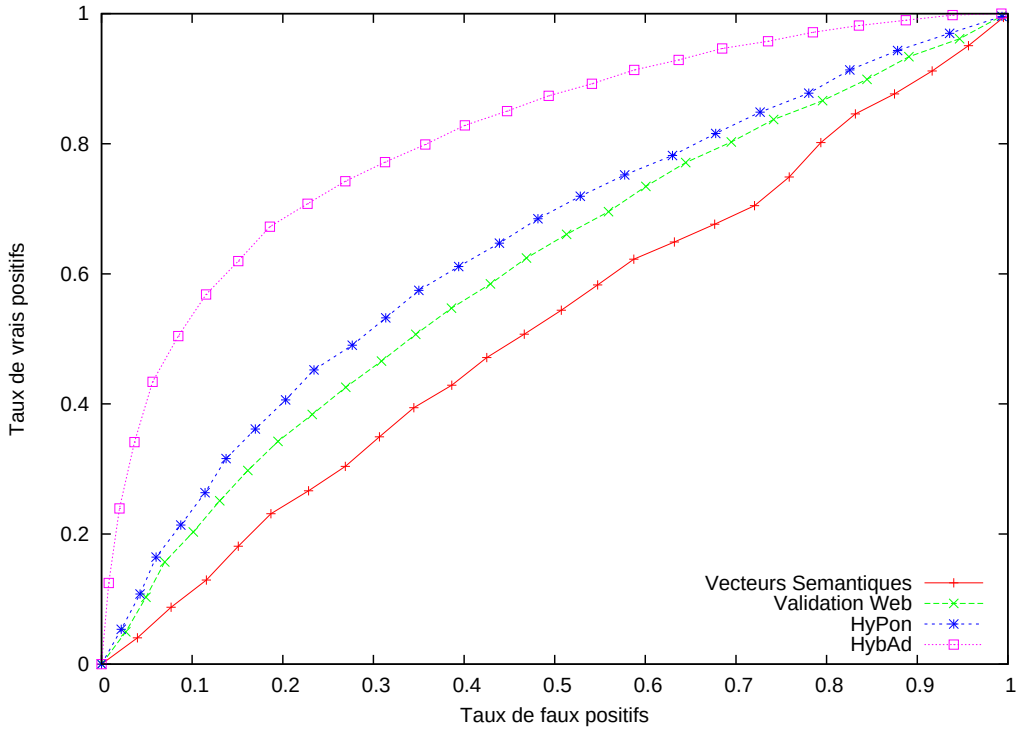

FIG. 7.6 – Courbes pour un seuil de 5 000

Les figures 7.6 et 7.7 présentent respectivement les courbes ROC représentant les résultats obtenus pour un seuil de 5000 et 50000 avec les approches VS, VW, HYPON et HYBAD. Elles nous permettent de conclure de la manière suivante.

– Validation Web : Approche mieux adaptée avec des seuils élevés (par exemple afin d'effectuer une tâche de classification de textes).

– HYBAD : Approche adaptée aux faibles seuils (par exemple dans le cadre de construction

de classes conceptuelles). Ces résultats confirment ceux obtenus lors de nos expérimentations visant à enrichir des classes conceptuelles.
Notons que l'approche HybAd pourrait également être utilisée avec des seuils importants. Cependant, l'approche Validation Web est moins coûteuse et fournit des résultats équivalents. Ainsi, mieux vaut utiliser cette dernière. Les résultats obtenus avec l'approche

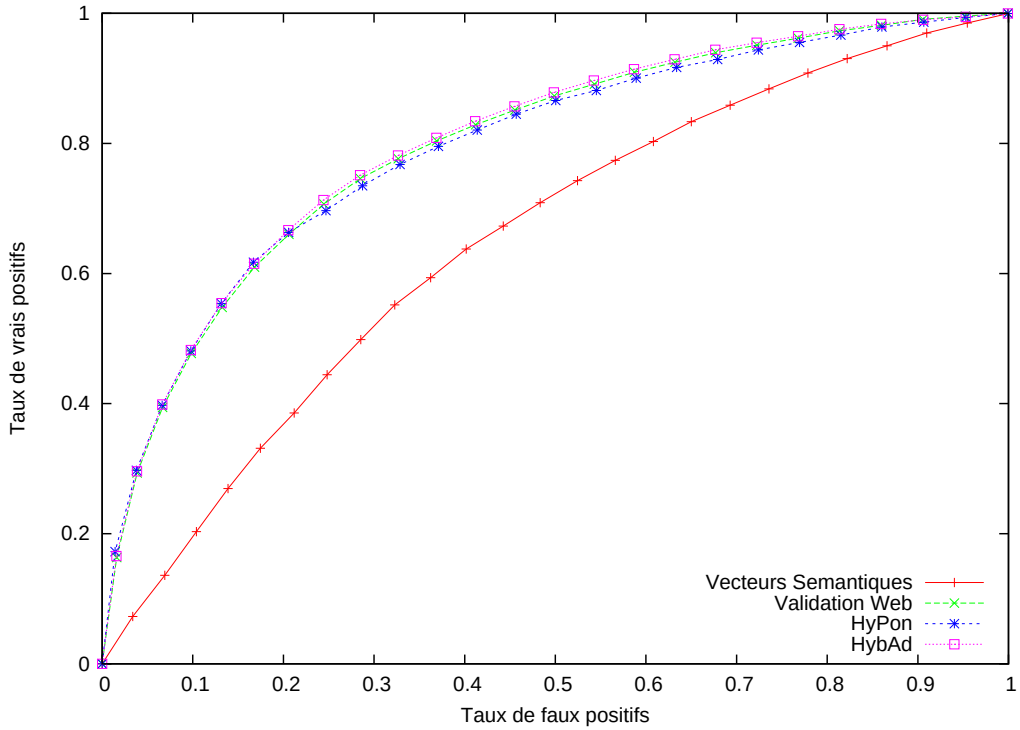

Fig. 7.7 – Courbes pour un seuil de 50 000

HybAd semblent être les plus prometteurs. Ils se fondent sur une première sélection effectuée avec l'approche VS, qui, comme nous l'avons précédemment mentionné, utilise des connaissances linguistiques afin de déterminer la proximité sémantique entre un verbe et un objet complémentaire. Cependant, les résultats obtenus avec l'approche VS, pour des faibles seuils, sont assez proches de ceux que nous obtiendrions avec une distribution aléatoire des relations induites (cf tableau 7.8 et 7.9). D'un point de vue purement statistique, les informations linguistiques n'introduisent pas de changement remarquable. Nous avons alors cherché à comprendre comment des résultats négatifs peuvent fournir avec l'approche HybAd de très bons scores. Ainsi, nous avons produit une distribution aléatoire des relations induites. Dès lors, nous appliquons l'approche HybAd en remplaçant la liste triée avec l'approche VS par notre distribution aléatoire. En d'autres termes, nous partons de la liste de relations syntaxiques induites triées de manière aléatoire pour finalement classer les n premières relations avec l'approche VW. Nous proposons de comparer les résultats de HybAd avec cette nouvelle approche dans le tableau 7.14. Les AUC obtenues pour

	HybAd		Aléatoire	
Seuil	AUC	+	AUC	+
5000	0.813	1362	0.801	750
10000	0.795	2675	0.808	1542
15000	0.788	3809	0.808	2323
20000	0.790	4790	0.810	3078
25000	0.789	5575	0.809	3863
30000	0.791	6248	0.809	4648
35000	0.793	6838	0.808	5438
40000	0.796	7332	0.808	6229
45000	0.799	7758	0.807	6982
50000	0.804	8070	0.806	7758

TAB. 7.14 – Comparaison de l'approche HybAd avec l'approche "aléatoire"

chacune des approches sont du même ordre. Cependant, le nombre de relations positives obtenu pour chaque approche est très différent (c.-à-d. le nombre de relations syntaxiques induites retrouvées dans le corpus du *Monde*). L'approche HYBAD permet presque d'obtenir deux fois plus de relations positives que l'approche "aléatoire" pour un même seuil. Nous montrons avec cette expérimentation que l'apport linguistique de l'approche VS n'est pas négligeable, cette dernière permettant d'extraire un nombre plus important de relations induites "positives".

7.3.3 Discussions

7.3.3.1 La qualité des résultats

L'aire sous la courbe ROC (AUC) est un bon indicateur de la qualité d'une mesure en permettant une évaluation globale des fonctions de rang. Nous proposons d'étudier plus finement la pertinence des premières relations en calculant la précision, car ce sont les premières relations qui pourront être prises en compte par un expert dans le cadre de l'enrichissement de classes conceptuelles. Nous proposons alors de calculer pour les 1 000 premières relations induites, la précision des approches *VW* et HYBAD. La précision qui calcule la proportion de relations induites correctes est définie comme suit :

$$\text{Précision} = \frac{\text{Nombre de relations syntaxiques induites positives}}{\text{Nombre de relations syntaxiques induites}} \tag{7.2}$$

La figure 7.8 montre les *courbes d'élévation* ou *courbes lift* (précision en fonction du nombre de relations syntaxiques) des 1 000 premières relations. Une telle courbe permet d'avoir une vue globale de la précision. Nous constatons que l'approche VW obtient une précision peu variante. En effet, la précision oscille entre $0,70$ et $0,75$, score étant par ailleurs de bonne qualité. La figure 7.8 montre également la qualité de l'approche HYBAD

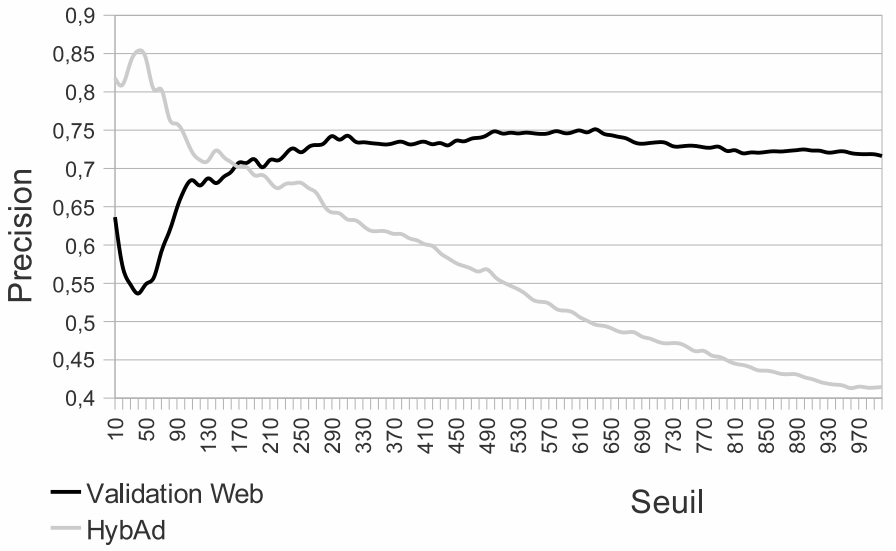

FIG. 7.8 – Courbe lift comparant le classement des approches VW et HYBAD.

pour les 200 premières relations. Ce résultat signifie que les relations pertinentes sont bien ordonnées en début de liste par cette combinaison et confirme donc les résultats obtenus avec les courbes ROC : l'approche HYBAD est mieux adaptée pour de faibles seuils.
Après avoir discuté de la précision des résultats des approches VW et HYBAD, nous avons alors cherché à savoir si les relations syntaxiques placées en tête de liste avec la combinaison HYBAD étaient les mêmes que celles classées en tête avec la validation Web. Autrement dit, est-ce que les relations privilégiées par HYBAD sont les relations les plus populaires sur le Web ?
Nous présentons dans le tableau 7.15 un extrait des 500 premières relations syntaxiques

Relations syntaxiques induites	*Classement avec HybAd*	*Classement avec VW*	*Positive / Négative*
mettre_note	0	7	+
éviter_récession	50	642	+
établir_sorte	100	1132	+
connaître_décision	150	1645	+
nommer_entraîneur	200	2142	-
réaliser_retour	250	2639	+
suspendre_traitement	300	3314	+
maintenir_ouverture	350	3854	+
accepter_déclaration	400	4452	+
provoquer_séparation	450	4878	-
gérer_ouverture	500	5424	-

TAB. 7.15 – Comparaison des relations classées avec VW et HYBAD

ordonnées avec l'approche HybAd, et leur classement respectif avec l'approche VW. Ces résultats montrent que les relations syntaxiques induites situées en tête du classement effectué par HybAd ne sont pas "populaires". En effet, ces dernières ne sont pas fréquemment employées sur le Web (c.-à-d. ces relations syntaxiques induites ne sont pas placées en tête de liste avec l'approche VW). Deux conclusions peuvent être émises pour interpréter ces résultats.
– La première vient du fait que le Web contient un très grand nombre d'informations, lesquelles sont rédigées dans un langage "de tous les jours". Ainsi, nos relations induites peuvent êtres des relations d'un langage plus littéraire.
– La seconde explication implique que nous déterminions et validions des pépites de connaissances, pouvant être plus discriminantes et plus intéressantes que des relations fréquentes qui n'apportent pas d'informations nouvelles. Une telle situation s'avère classique dans le domaine de la "Fouille de Données".

7.3.3.2 La taille minimum du corpus de validation

Après avoir discuté de la qualité des relations syntaxiques induites obtenues avec nos approche de validation, nous nous focalisons dans cette section sur le protocole expérimental automatique. Rappelons que ce dernier nécessite l'emploi d'un second corpus afin de déterminer si une relation qui est "induite" de notre corpus de test, existe dans ce second corpus. Nous avons montré lors des expérimentations menées afin de construire et d'enrichir des classes conceptuelles que ce protocole fournissait une bonne interprétation des résultats. Notons cependant que la qualité du protocole fut discutée avec l'ensemble des relations syntaxiques (sans considérer de seuil). Ce dernier est en effet limité à cause des faux négatifs qu'il génère.
Nous avons alors cherché à répondre à la question suivante. La taille du second corpus (que nous nommerons par la suite "corpus de validation") utilisé afin de valider l'existence ou non d'une relation induite influence-t-elle la qualité du protocole ? Ainsi, nous proposons d'effectuer l'expérimentation suivante. Le corpus de validation va être divisé en n parties. Chaque section de corpus va ensuite servir à valider les approches utilisant la validation Web, et ce en effectuant une validation croisée. Ainsi, pour $n = 1000$, 1000 expérimentations vont être effectuées afin de calculer une AUC moyenne correspondant à un corpus d'évaluation d'une taille d'environ 50000/1000 soit 50 articles en moyenne. Notons que la valeur $n = 1$ revient à considérer la totalité du corpus de validation. Cette dernière valeur pour le paramètre n nous servira alors de score AUC de référence.
La figure 7.9 présente les AUC obtenues pour $n \in [1, 10000]$. Nous ne présentons dans cette figure uniquement les résultats obtenus avec l'approche VW, pour les quatre mesures statistiques employées (avec *nb_somme*). Nous nous focalisons sur le seuil 50000

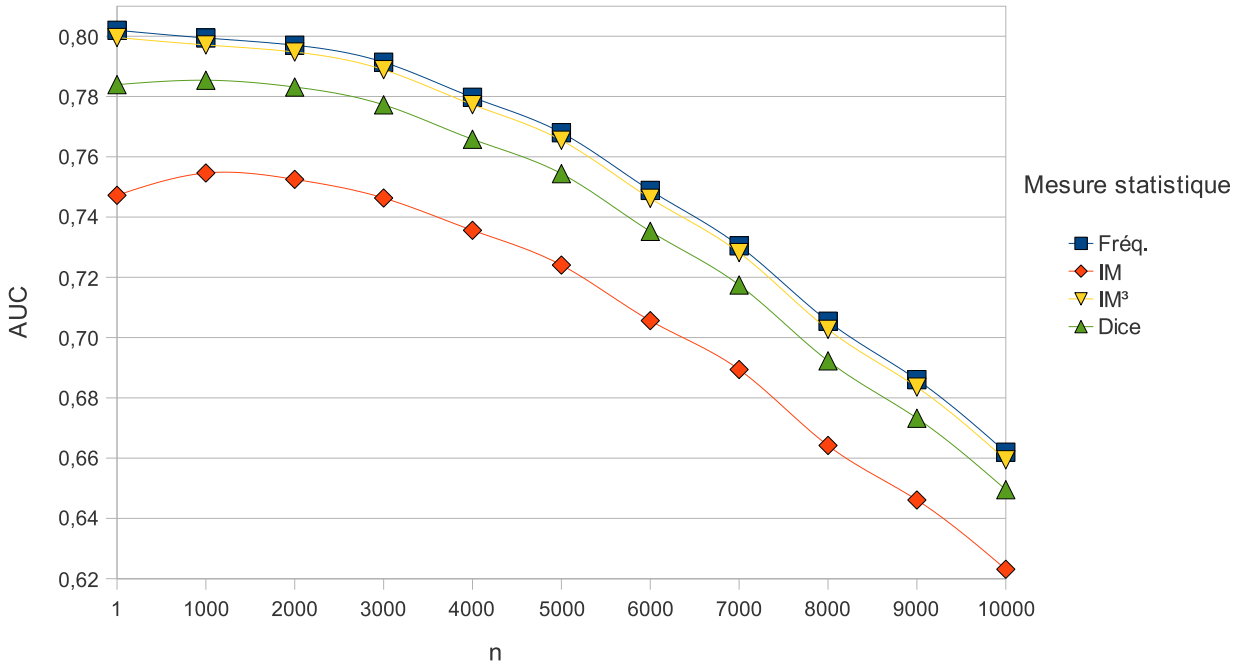

FIG. 7.9 – AUC obtenues en fonction de la taille du corpus de validation.

indiquant que nous considérons la totalité des relations syntaxiques, seuil où l'approche VW obtient de meilleurs résultats. Ces résultats expérimentaux montrent que la taille du corpus de validation peut être réduite de 4000 tout en conservant des résultats équivalents à ceux obtenus avec le corpus de validation dans son ensemble. En effet, pour l'intervalle $n \in [1, 4,000]$, les résultats obtenus en termes d'AUC sont du même ordre, variant de plus ou moins $0,2$ points. Au delà de 4000, les scores ne sont plus pertinents. Par exemple avec la *fréquence*, l'AUC passe de $0,80$ pour le corpus entier à $0,79$ pour $n = 3000$ puis à $0,67$ pour $n = 10000$. Cette baisse des AUC peut s'expliquer par une limite théorique due à un trop faible nombre de relations couvertes (nombre de relations syntaxiques induites retrouvées dans le corpus de validation) pour des valeurs de n trop grandes. En effet, un nombre trop faible de relations couvertes reflète un manque de finesse dans les AUC résultantes.

n	***Nb .Rel. Retrouvées***		
1	*8268*	*5000*	*4*
1000	*20*	*6000*	*3,33*
2000	*10*	*7000*	*2,8*
3000	*6,6*	*8000*	*2,5*
4000	*5*	*9000*	*2,2*
		10000	*2*

TAB. 7.16 – Nombre de relations syntaxiques retrouvées en fonction de la taille du corpus de validation.

Les résultats présentés dans le tableau 7.16 indiquent le nombre de relations couvertes en fonction de la taille du corpus considéré (la taille du corpus correspond à celle du corpus de validation divisée par n). Pour une valeur de n inférieure à 4000, les AUC sont assez similaires à celles obtenues avec la totalité du corpus de validation. Ainsi, nous montrons par les résultats du tableau 7.16 qu'avec seulement 5 relations syntaxiques couvertes (c.-à-d. retrouvées dans le "corpus de validation"), nous sommes en mesure d'appliquer efficacement le protocole d'évaluation automatique.

La figure 7.10 présente les AUC obtenues en fonction de la mesure statistique utilisée. Nous confirmons que pour un $n \in [1, 4000]$, les AUC résultantes sont similaires et que pour n supérieur à 4000, la qualité des AUC se dégrade. Les résultats de cette figure mettent également en avant le fait que l'allure des courbes, suivant la valeur de n, reste la même. Cela signifie que quel que soit la valeur de n (entre 1 et 10 000 ici) les différentes mesures statistiques conservent leur hiérarchie. Par exemple, les mesures *fréquence* et IM^3 restent celles qui, quel que soit la valeur de n, fournissent le meilleur classement des relations syntaxiques induites.

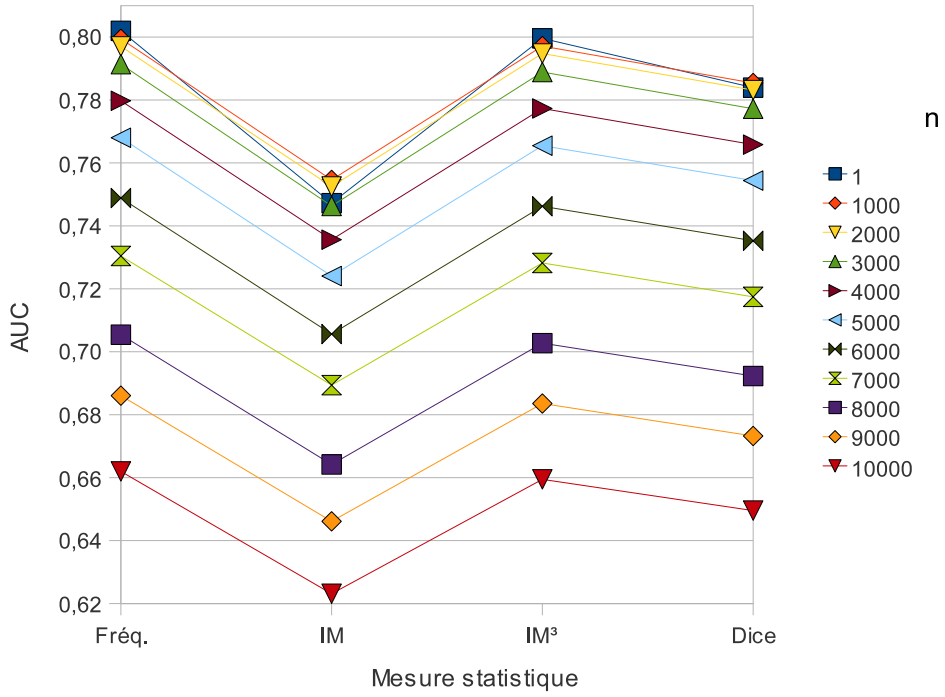

FIG. 7.10 – AUC obtenues en fonction de la mesure statistique utilisée avec l'approche VW.

7.3.3.3 La qualité du protocole d'évaluation

Nous avons montré dans la section précédente que le protocole d'évaluation automatique permettait d'utiliser un corpus de validation de faible taille. Il reste cependant important de mesurer la robustesse du protocole vis à vis des relations syntaxiques jugées non pertinentes. Cela correspond à des relations syntaxiques qui n'ont pas été retrouvées dans le corpus de validation. Certaines de ces relations peuvent être de faux négatifs, comme précédemment mentionné. Rappelons en effet que le corpus de validation n'est pas exhaustif en ce sens qu'il ne contient pas nécessairement toutes les relations syntaxiques induites d'un autre corpus. Une évaluation humaine a donc été effectuée afin de mesurer la quantité de faux négatifs, permettant également de confronter une évaluation automatique à une évaluation humaine tel que le montrent nos précédentes expérimentations (section 7.2). Notons cependant qu'une évaluation manuelle des 50000 relations induites n'est pas réalisable. Ainsi, nous proposons de suivre le protocole suivant.

Nous classons dans un premier temps la totalité des relations syntaxiques induites extraites de notre corpus (50 000) par l'approche VW avec la fréquence. Nous extrayons alors de manière homogène 100 relations parmi les 50 000. Ces relations sont alors évaluées par un expert les qualifiant de cohérentes ou non. En d'autres termes, l'expert indique si l'association d'un verbe v avec un objet o est sémantiquement pertinente. Notre objectif est alors de calculer une AUC "moyenne" correspondant à une extrapolation des résultats qui seraient obtenus avec la totalité des relations syntaxiques induites.

Nombre de relations	100		50000
Type d'évaluation	manuelle	automatique	automatique
AUC	0,88	0,82	0,81

FIG. 7.11 – Comparaison de l'évaluation manuelle et automatique

Le tableau 7.11 présente les AUC obtenues avec cette évaluation manuelle ainsi qu'avec l'évaluation automatique des 100 mêmes relations syntaxiques. Ils sont également comparés à l'AUC obtenue pour l'ensemble des relations syntaxiques avec le protocole automatique. Les points remarquables résultant de ces expérimentations sont les suivants.

• **Interpolation de qualité**. En effet, les AUC obtenues pour l'approche automatique avec les 100 relations et avec l'ensemble des relations syntaxiques sont très proches, confirmant l'homogénéité de l'échantillon sélectionné.

• **Quantité acceptable de faux négatifs**. Les résultats obtenus avec le proto-

cole automatique et ceux issus de l'approche manuelle restent en effet assez proches. Les faux négatifs sont donc existants avec le protocole d'évaluation automatique mais restent acceptables.

7.3.4 Synthèse

Cette seconde phase d'évaluations a permis d'évaluer la qualité des approches de validation du modèle SelDeF "à grande échelle". Nous avons en effet déjà montré lors des précédentes expérimentations la qualité de certaines de celles-ci pour un nombre réduit de relations syntaxiques induites. Ainsi, nous avons expérimenté dans cette section la qualité de nos approches sur 50 000 relations induites. Ce nombre de relations étant trop conséquent afin d'appliquer une validation manuelle, nous avons proposé un protocole d'évaluation automatique. Ce dernier considère une relation syntaxique induite comme "probable" si cette dernière existe dans un autre corpus du même domaine. Nous avons montré que ce protocole est de qualité, produisant un nombre acceptable de faux négatifs. Par ailleurs, nous avons également montré à partir de notre corpus que le recouvrement de cinq relations suffisait afin que le protocole soit applicable.
Nous avons évalué la qualité des approches de validation en calculant l'aire sous des courbes ROC (AUC). Plusieurs seuils ont également été testés afin de proposer la meilleure approche de validation en fonction de la tâche pour laquelle les relations induites seront utilisées. Nous sommes alors arrivés aux mêmes conclusions que lors de nos précédentes expérimentations : l'approche HybAd est adaptée afin de sélectionner un nombre réduit de relations. L'approche VW est quant à elle préconisée afin d'extraire un grand nombre de relations. Notons cependant que ces deux approches sont assez coûteuses en termes de temps. Par exemple, pour un ensemble de 50000 relations syntaxiques, il est nécessaire d'effectuer 350 000 requêtes en utilisant la mesure statistique IM^3 (1 requête pour le verbe, 1 pour l'objet et 5 pour les couples verbe-objet séparés par cinq articles : $50000 \times 7 = 350000$). Bien qu'assez coûteuse, ces approches restent cependant automatiques et plus rapides qu'une évaluation manuelle nécessitant un travail d'expertise trop important au regard de la qualité de descripteurs à analyser.

Chapitre 8

Conclusion et Perspectives

Sommaire

8.1 Synthèse

La notion de descripteurs est essentielle à toute tâche manipulant des données. Ces descripteurs sont le plus souvent fondés sur des approches statistiques. Citons par exemple la fouille de données qui privilégie ce type d'approches. Les langues naturelles que nous avons manipulées sont d'une nature plus complexe et peuvent être décrites par d'autres techniques, prenant en considération leurs fondements linguistiques. Ainsi, un type de descripteurs peut être une flexion, un lemme, etc. comme nous l'avons montré en section 2.1.1. Les approches de sélections vont également tenir compte des propriétés linguistiques. Nous pouvons alors sélectionner des descripteurs en fonction de leurs propriétés morphosyntaxiques ou encore nous référer à des modèles de connaissances spécifiques à un domaine. Nous avons défendu dans ce mémoire le fait que la syntaxe devait être considérée afin de décrire des données textuelles écrites en langues naturelles. Pour cela, nous avons proposé deux approches de sélection de descripteurs. Le principe est d'extraire les relations syntaxiques d'un corpus et d'utiliser les objets des verbes sémantiquement proches comme descripteurs. Notre objectif fut alors de présenter des méthodes indépendantes des tâches effectuées. La syntaxe reste à notre sens une information bénéfique à tout type de tâche nécessitant la manipulation de textes rédigés en langues naturelles.

Le premier modèle de sélection de descripteurs est le modèle SelDe. Nous avons proposé une méthode d'enrichissement de contextes, *ExpLSA*, afin d'évaluer la qualité

des descripteurs fournis par SelDe. Cette approche utilise le modèle de réduction LSA, afin de mettre en relation de nouveaux descripteurs du fait de l'enrichissement. Deux tâches ont alors été effectuées : la classification conceptuelle et la classification de textes. Ces expérimentations ont mis en évidence la qualité des descripteurs de SelDe. Les résultats de classification ont en effet montré que l'approche ExpLSA obtenait des résultats supérieurs ou équivalents à ceux obtenus sans enrichissement. Notons également que l'introduction de paramètres de sélections proposés par SelDe se révèle pertinente. Nous obtenons en effet la plupart du temps de meilleurs résultats de classification en les utilisant. Cependant, les performances de classification sont faiblement améliorées par rapport à l'utilisation d'un corpus sans enrichissement. Finalement, ces résultats expérimentaux nous amènent à conclure par les points suivants.

– Les descripteurs de SelDe **sont pertinents**, notamment avec l'appui de différents filtres statistiques proposés dans nos travaux qui permettent d'obtenir de meilleures classifications par rapport à une approche sans filtrage.

– L'approche ExpLSA est à améliorer. Les améliorations ne sont en effet pas assez significatives afin de proposer une approche efficace de classification. Ce type d'enrichissement nécessite un autre paradigme se fondant notamment sur des méthodes linguistiques. Nous avons en effet montré les limites des approches statistiques.

Les données textuelles extraites par SelDe doivent répondre à certaines contraintes. En effet, notre modèle d'extraction repose sur les informations syntaxiques contenues dans un corpus. Cependant, certains types de données sont assez pauvres syntaxiquement. Dans ce cas, le modèle SelDe ne peut être utilisé. Nous avons alors présenté un certain nombre d'alternatives afin de traiter des données "complexes", notamment en se fondant sur des descripteurs morphosyntaxiques : les catégories lexicales. Ces approches ont été pertinentes avec des données syntaxiquement mal formulées, bruitées et/ou incomplètes. Les données dépourvues de syntaxe comme des CV restent cependant assez difficiles à traiter par les approches présentées dans ce mémoire. Nous envisageons d'évaluer d'autres approches en nous fondant notamment sur la structure caractéristique de ces données, l'utilisation de la méthode LSA, la découverte de règles d'association pouvant aider à la compréhension des données, etc.

Nous avons montré la qualité des descripteurs sélectionnés par SelDe. Notons néanmoins que certaines connaissances apportées par ce modèle ont été écartées : celles résultantes des relations syntaxiques induites de la proximité sémantique de deux verbes. L'intérêt de ces relations vient du fait qu'elles ne sont pas originalement présentes dans

un corpus et qu'elles constituent dès lors de l'information nouvelle. Cette caractéristique est également problématique car ces relations produisent une quantité importante de bruit. Nous avons alors présenté un second modèle de sélection de descripteurs : le modèle SelDeF. Ce dernier fournit des méthodes de sélection adaptées aux relations syntaxiques induites. Nous avons expérimenté ce modèle pour une tâche de construction de classes conceptuelles utilisant SelDe et avec une méthode permettant d'enrichir ces classes se fondant sur les descripteurs fournis par SelDeF. Une seconde approche d'enrichissement de classes conceptuelles a également été proposée. Cette dernière utilise les ressources du Web afin de générer de nouveaux termes pour les classes conceptuelles. Nous avons finalement comparé ces deux approches par des évaluations manuelles et automatiques.

8.2 Perspectives

Un certain nombre de perspectives peuvent être envisagées afin de poursuivre les travaux présentés dans ce mémoire :
– prise en compte d'autres types de relations syntaxiques (comme sujet-verbe) dans SelDe suivant le domaine du corpus étudié,
– proposer des patrons caractéristiques des corpus en se fondant sur les relations syntaxiques à la manière de [Shen *et al.*, 2005] dont l'approche est évoquée en section 2.1.2.2,
– tester nos approches sur d'autres corpus de domaines plus spécifiques,
etc.

Nous proposons également de nous intéresser à la notion de **contexte** de manière plus précise dans nos perspectives. Cette dernière joue en effet un rôle essentiel dans la compréhension et l'exploitation d'une langue naturelle. Ainsi, nous partons du principe qu'un mot ne prend son sens que dans une phrase, donc dans un contexte donné. Cette notion provient de la théorie de l'*anvitabhidhana* datant du VII^{ème} siècle Indien telle que précisé dans [Rastier, 1996]. La définition même du contexte est assez discutée dans la littérature. [Adam, 1999] précise en effet que cette notion doit être clairement redéfinie. Il évoque le fait que le contexte soit toujours employé en linguistique afin de lever des ambiguïtés, alors que cette notion est présente dans toute interprétation de langage faisant intervenir une mémoire discursive[45]. Ainsi, de nombreuses définissions et théories sont issues de la littérature. Citons par exemple les théories de *Teun A. Van Dijk* détaillées dans [Van Dijk, 2008]. Ce dernier rapproche les notions de contexte et de

[45] [Courtine, 1981] défini en 1988 la mémoire discursive comme se référant à l'"existence historique de l'énoncé"

situation. Citons également *Emanuel A. Schegloff*. Pour ce dernier, il est impossible de définir comme pertinent un contexte précis sans savoir pour qui ce contexte est pertinent [Schegloff, 1992]. Outre les données textuelles, la prise en compte de contextes est assez répandue dans la littérature comme dans [Picard & Estraillier, 2008] qui propose de contextualiser une tâche de capture de mouvement en se référant aux contextes dans le discours. Par ailleurs, un certain nombre de conférences y sont dédiées comme CONTEXT[46] (*Conference on Modeling and Using Context*) où de nombreux travaux sur le sujet y sont publiés. Citons également le numéro spécial de la revue SCOLIA[47] (*Sciences COgnitives, Linguistique et Intelligence Artificielle*) de 1996 consacrée au contexte.

Cette notion importante de contexte nous a amené à le considérer dans les approches présentées dans ce manuscrit. Une première application a été envisagée : l'introduction de contexte dans le cadre de la validation automatique de relations syntaxiques induites. L'approche des vecteurs sémantiques contextualisées de mots propose en effet de considérer le contexte sémantique d'une phrase afin de pondérer les poids de ses composantes. Outre cette méthode, nous pouvons prendre en compte le contexte dans d'autres approches de ce mémoire.

8.2.1 Le contexte dans ExpLSA

L'approche devant à notre sens bénéficier le plus d'une contextualisation est la méthode d'enrichissement de contexte ExpLSA présentée en section 4.1. Plusieurs points peuvent motiver ce choix.
Comme évoqué à la fin du chapitre 4, la méthode ExpLSA effectue une expansion similaire pour chaque terme identique.
Nous proposons alors de prendre en compte le contexte dans la phase d'enrichissement de corpus. Rappelons que le contexte de la phrase est pris en compte lors de l'extraction des descripteurs avec la méthode SelDe mais il n'est pas pris en compte dans ExpLSA (c'est-à-dire au moment d'enrichir). Concrètement, si le terme "*légume*", rencontré dans un corpus, est enrichi avec les termes "*fruit*" et "*viande*", il restera enrichi avec ces termes tout au long du processus d'enrichissement. Rappelons qu'ExpLSA est définie pour une tâche de classification de documents textuels. Alors, la non contextualisation de cette approche revient à émettre l'hypothèse que chaque document d'un corpus contient un vocabulaire propre à sa classe. Cela se vérifie dans certains cas, comme pour une classe "sport" ou probablement les termes "*arbitre*", "*joueur*", "*compétition*", etc.

[46] http ://mainesail.umcs.maine.edu/Context/context-conferences
[47] http ://www.patrick-schmoll.com/contexte.html

sont très employés. Cependant, certains de ces termes peuvent être également employés dans d'autres classes comme "*people*", "politique[48]". Nous mettons ainsi en avant la nature polysémique des termes pouvant défavoriser l'approche ExpLSA. Il est ainsi indispensable afin d'améliorer cette approche de prendre en compte le contexte généré par les documents d'un corpus.

Une proposition d'approche prenant en compte le contexte avec ExpLSA est la suivante. Rappelons que l'approche ExpLSA propose d'enrichir un corpus par le biais de descripteurs (des relations syntaxiques), sélectionnés par une mesure statistique. Ainsi, plutôt que de sélectionner les descripteurs de manière statistique dans ExpLSA, nous nous inspirons de la validation Web du modèle SelDeF destinée à la sélection des *objets complémentaires* (cf. 6) afin d'effectuer cette sélection. Cependant, à la différence de la validation Web "classique", cette dernière va prendre en compte le contexte "document" d'un corpus et proposer de valider des *objets communs*. Nous proposons la méthode suivante dont le principe est illustré dans la figure 8.1.

Pour chaque couple de verbes jugés proches composés des verbes *v1* et *v2* avec

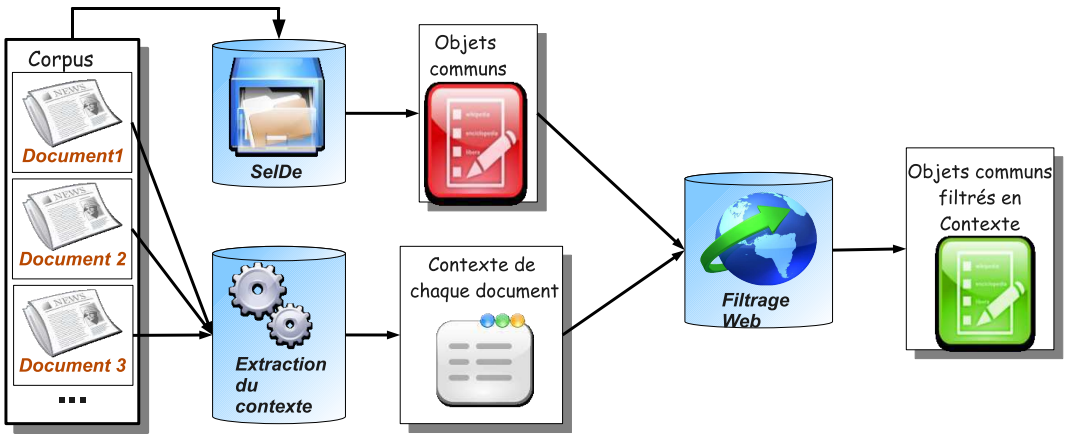

FIG. 8.1 – Proposition de contextualisation pour ExpLSA

$O_{i\in[1,nbObjet]}$ leurs objets communs, nous proposons de valider par le Web les relations syntaxiques par les requêtes suivantes dans un moteur de recherche :

"v1 O_1 *"* *contexte* – et – *"v2* O_1 *"* *contexte*

"v1 O_2 *"* *contexte* – et – *"v2* O_2 *"* *contexte*

...

"v1 $O_{nbObjet}$ *"* *contexte* – et – *"v2* $O_{nbObjet}$ *"* *contexte*

Le contexte, représentant le document dans lequel le terme à enrichir se situe, peut être défini par les termes les plus fréquents du document ou par des informations

[48] Nous pensons dans ce cas au "compétitions" dans le cadre de campagnes électorales par exemple.

morphosyntaxiques plus riches. Nous obtenons alors un score pour chaque objet du couple de verbes. À partir de la liste triée des objets en fonction de leurs scores obtenus, nous pouvons finalement ne sélectionner que les premiers objets par couple. Finalement, nous n'enrichissons plus chaque terme d'un corpus de manière identique (c'est-à-dire avec **tous** les objets communs d'un couple) mais uniquement avec les **objets les plus pertinents dans le contexte du document dans lequel apparait le terme**.
Notons pour finir les points suivants :
– Cette proposition est adaptable pour la validation d'objets complémentaires.
– Une approche similaire peut également intervenir dans le choix du couple de verbes. En effet, le document dans lequel le terme à enrichir apparait peut également influencer ce choix.
– Finalement, cette proposition peut de surcroit utiliser des mesures statistiques tels que le coefficient de Dice ou l'information mutuelle à l'instar de la validation Web "classique".

8.2.2 Le contexte pour l'enrichissement de classes conceptuelles

Les expérimentations menées afin de construire et d'enrichir des classes conceptuelles peuvent également être améliorées en considérant le contexte. Pour l'approche d'enrichissement se fondant sur SelDeF, nous pouvons utiliser la validation Web proposée dans la section précédente. Le contexte peut par exemple être défini par le nom de la classe conceptuelle comme l'exemple "objets symboliques" développé dans le chapitre 7. L'approche d'acquisition de nouveaux candidats utilisant le Web peut également prendre en compte le contexte de la classe conceptuelle à enrichir, en introduisant ces connaissances dans les requêtes Web.

8.2.3 Mesurer la proximité sémantique de verbes

Nous mesurons avec les approches SelDe et SelDeF la proximité des verbes décrits par leurs objets respectifs avec la mesure d'Asium. Le contexte est ici pris en compte par les relations syntaxiques. Cependant, l'information résultante des documents d'un corpus, quand documents il y a, n'est pas considérée. Une technique pouvant introduire le contexte d'un document peut s'inspirer du *tf-idf*. Ainsi, au lieu de définir la mesure d'Asium en prenant en compte uniquement le nombre d'occurrences des objets des verbes, nous proposons de pondérer ce score par la fréquence d'apparition de chaque relation syntaxique (c.-à-d. l'objet et le verbe considéré) dans les documents. Cette proposition reste cependant à approfondir du fait de sa complexité.

Une autre possibilité d'introduire le contexte "document" serait d'utiliser des vecteurs sémantiques globaux (cf. section 6.3.1.2) afin de mesurer la proximité des verbes d'un corpus. Le principe est de calculer le vecteur sémantique d'un verbe de la manière suivante.

1. Extraire du corpus chaque phrase dans laquelle le verbe à représenter apparait.
2. Représenter ces phrases sous forme de vecteurs sémantiques (implicitement contextualisés).
3. Calculer le barycentre de ces vecteurs afin d'obtenir le vecteur sémantique représentant notre verbe "en contexte".

Dès lors, nous pouvons mesurer la proximité de verbes d'un corpus en calculant la distance de concordance (définie en section 6.3.1.4) entre chacun des vecteurs "verbe". Cette approche ne décrit alors plus les verbes en fonction de leurs objets mais en fonction des phrases (et implicitement des documents) dans lesquels ils sont présents.

8.2.4 Vers une nouvelle problématique : les descripteurs dans les entrepôts de données

La notion d'entrepôts de données (*Data Warehouse* en anglais) apparue pour la première fois dans [Devlin & Murphy, 1988] dans lequel les auteurs présentent une alternative aux bases de données relationnelles. Ils proposent un modèle d'entrepôts de données (nommé EBIS) afin de traiter d'importantes masses d'informations fournies par les systèmes d'informations de la société IBM. Un entrepôt de données peut être défini comme une collection de données organisées par sujets ou thématiques. Ces bases de données sont principalement utilisées pour améliorer des systèmes d'aide à la décision. Les données contenues dans un entrepôt ont la particularité d'être persistantes (elles sont en lecture seule, donc stables et non modifiables) et temporelles (chaque donnée est datée) [Inmon, 1991]. Elles sont organisées suivant des axes d'analyses pouvant être l'année, le nombre d'habitants, le type de clientèles, etc. Ce caractère dimensionnel introduit le fait que ces bases de données sont modélisées par des objets multidimensionnels. Alors, le fait d'agréger les données d'un entrepôt par des fonctions d'agrégations permet de produire en sortie un nouvel objet multidimensionnel. Les paramètres de l'agrégation vont en quelque sorte limiter la vue de l'entrepôt à certaines contraintes, pouvant être un intervalle d'âge, de temps, etc.

De nombreux travaux ont été réalisés dans la littérature afin d'améliorer les techniques fondatrices des entrepôts de données comme ceux menés au sein du LIRMM [Choong *et al.*, 2006]. Une question demeure néanmoins, la notion de descripteurs. En

effet, la plupart des méthodes utilisant ou proposant des approches afin de modéliser des entrepôts utilisent des descripteurs statistiques, le plus souvent de type fréquentiels. Citons par exemple [Xide Lin *et al.*, 2008] où les auteurs proposent un nouveau modèle de représentation de type cube de données. Ce modèle, outre les bénéfices d'un modèle traditionnel de cube de données, inclut des méthodes de sélection statistique de descripteurs de données textuelles.

Nos futurs travaux, qui vont être menés à moyen terme, vont se focaliser sur l'étude précise des descripteurs pouvant être bénéfiques aux entrepôts de données manipulant des données textuelles et les descripteurs des documents. De telles données sont beaucoup plus complexes à traiter dans un contexte "entrepôt". Il est ainsi nécessaire de produire de nouvelles mesures d'agrégations spécifiques aux données textuelles.

Publications personnelles

Revues nationales et internationales

[1 - IJDEM] Laroum S., **Béchet N.**, Hamza H., Roche M. Hybred : An OCR document representation for the classification tasks. In *IJDEM (International Journal on Data Engineering and Management).*

[2 - RNTI] Laroum S., **Béchet N.**, Hamza H., Roche M. Classification automatique de documents bruités à faible contenu textuel. In *RNTI (Revue des Nouvelles Technologies de l'Information), numéro spécial fouille de données complexes.*

Conférences internationales

[3 - MAW'09] **Béchet N.**, Roche M., Chauché J. A Hybrid Approach to Validate Induced Syntactic Relations.
In *MAW'09 (the 2009 IEEE International Symposium on Mining and Web) - In conjunction with IEEE AINA 2009, May 2009, Bradford, UK.* p. 727–732.

[4 - ISMIS'09] Kessler R., **Béchet N.**, Torres-Moreno J.M., Roche M., El-Bèze M. Job Offer Management : How Improve the Ranking of Candidates. LNCS - ISMIS'09 (International Symposium on Methodologies for Intelligent Systems). p 431–441.

[5 - ECIR'09] **Béchet N.**, Roche M., Chauché J. Towards the Selection of Induced Syntactic Relations.
In *ECIR'09 (Springer-Verlag, LNCS (poster proceedings) - 31st European Conference on Information Retrieval), Toulouse, France, April 2009*, p. 786-790.

[6 - ICDIM'08] **Béchet N.**, Roche M., Chauché J. How the ExpLSA approach impacts the document classification tasks.
In *IEEE International Conference on Digital Information Management - ICDIM'08, University of East London, London, United Kingdom, November 2008*, p. 241–246.

[7 - IIP'08] Bayoudh I., **Béchet N.**, Roche M. Blog classification : Adding Linguistic Knowledge to Improve the K-NN Algorithm.
In *Springer IFIP - International Federation for Information Processing, Volume 288 - IIP'08, Beijing, China, October 2008*, p. 68–77.

[8 - CICLing'08] **Béchet N.**, Roche M. & Chauché J. ExpLSA : An approach based on syntactic knowledge in order to improve LSA for a conceptual classification task.
In *RCS volume (Research in Computing Science), CICLing 2008 (posters proceedings), February 2008, Haifa University, Israel*, p. 213–224.

[9 - JADT'08] **Béchet N.**, Roche M. & Chauché J. Utilisation de ExpLSA pour la classification de textes.
In *JADT'08 (Journées internationales d'Analyse statistique des Données Textuelles, Mars 2008, ENS Lettres et Sciences humaines, Lyon, France*, p. 167–178.

Workshops internationaux

[10 - QSI'08] Kessler R., **Béchet N.**, Roche M., El-Bèze M., Torres-Moreno J.M. E-Gen : automatic profiling system for ranking candidates answers in Human Resources.
In *Springer-Verlag, LNCS - On The Move federated Conferences and Workshops - QSI'08 - workshop - OTM'08, Monterrey, Mexico, November 2008*, p. 625-634.

[11 - CIR'07] **Béchet N.**, Roche M. & Chauché J. Improving LSA by expanding the contexts.
In *CIR'07 (Context-Based Information Retrieval) workshop - CONTEXT'07 (short paper), 20-24 August 2007, Roskilde University, Denmark*, p. 105-108.

Conférences nationales

[12 - TOTh'09] **Béchet N.**, Roche M., Chauché J. Corpus et Web : deux alliés pour la construction et l'enrichissement automatique de classes conceptuelles. In *TOTh'09 (Terminologie & Ontologie : Théories et Applications)*. A paraître.

[13 - TALN'09] Kessler R., **Béchet N.**, Torres-Moreno J.M., Roche M., El-Bèze M. Profilage de candidatures assisté par Relevance Feedback. In *TALN'09 - Traitement Automatique du Langage Naturel (poster)*. 10 p.

[14 - EGC'09] **Béchet N.**, Roche M. & Chauché J. Vers une approche hybride pour la validation de relations syntaxiques induites.
In *EGC'09 (Extraction et Gestion des Connaissances), 27 au 30 Janvier 2009 à Strasbourg.* p. 169-180.
Article nominé parmi les 9 meilleurs articles académiques d'EGC'09.

[15 - EGC'08] **Béchet N.**, Roche M. & Chauché J. ExpLSA : utilisation d'informations syntaxico-sémantiques associées à LSA pour améliorer les méthodes de classification conceptuelle.
In *EGC'08 (Extraction et Gestion des Connaissances), INRIA Sophia Antipolis, 29 janvier au 1er février 2008*, p. 589–600.

[16 - INFORSID'07] **Béchet N.** Utilisation d'informations syntaxico-sémantiques associées à LSA.
In *INFORSID'07 (Informatique des organisations et systèmes d'information et de décision) dans le cadre du Forum Jeunes Chercheurs,22 au 25 mai 2007, Perros-Guirec*, p. 555–556.

Ateliers nationaux

[17 - EvalECD'09] **Béchet N.** Description d'un protocole d'évaluation automatique comme alternative à l'évaluation humaine. Application à la validation de relations syntaxiques induites,
In *EvalECD'09 (Evaluation des Méthodes d'Extraction de Connaissances dans les Données), Atelier de EGC'09*, p. 525-534.

[18 - QDC'08] **Béchet N.**, Bayoudh I. Quelles connaissances linguistiques permettent d'améliorer la classification de blogs avec les k-ppv ?
In *QDC'08 workshop (Qualité des Données et des Connaissances) - EGC'08, Février 2008, Sophia Antipolis, France*, p. 71-80.

[19 - AtelierTALN'07] Wandmacher T., **Béchet N.**, Barhoumi Z., Poirier F., Antoine J.Y. Système Sibylle d'aide à la communication pour personnes handicapées : modèle linguistique et interface utilisateur.
In *Reconstruire la langue dans les communications alternatives et augmentées, TALN'07, Juin 2007, Toulouse, France*

Soumissions

[20 - NLE] **Béchet N.**, Chauché J., Prince V., Roche M., Validating Induced Syntactic Relations in a Semantic Knowledge Acquisition Process : An Automatic Procedure Reducing the Human Expert Effort. Soumis à la revue internationale *NLE (Natural Language Engineering).*

Table des figures

Liste des tableaux

Annexe A

Classification conceptuelle

Nous présentons ici les annexes relatives aux expérimentations du chapitre 4 concernant la tâche de classification conceptuelle.

A.1 Détail des expérimentations dans le choix des paramètres

Nous présentons les différentes valeurs des paramètres que nous avons expérimentés. Ces paramètres sont expérimentés avec un $SA = 0,8$.

Sélection Couple	NbOccMin	*MicroMoy*	*MacroMoy*
Asium	2	39,84%	37,97%
Asium	4	43,76%	42,45%
Asium	6	44,57%	42,95%
Asium	8	43,84%	42,10%
Occurrences	2	37,80%	35,82%
Occurrences	4	43,76%	42,26%
Occurrences	6	44,00%	42,30%
Occurrences	8	43,84%	42,10%

Tab. A.1 – Paramètre NbOccMin

Sélection Couple	NbOccMax	*MicroMoy*	*MacroMoy*
Asium	2	43,84%	42,10%
Asium	4	43,84%	42,46%
Asium	6	40,57%	38,88%
Asium	8	38,94%	37,07%
Occurrences	2	43,84%	42,10%
Occurrences	4	43,10%	41,69%
Occurrences	6	40,73%	38,90%
Occurrences	8	40,33%	38,36%

Tab. A.2 – Paramètre NbOccMax

Sélection Couple	NbOccMin	NbOccMax	*MicroMoy*	*MacroMoy*
Asium	2	4	43,84%	42,46%
Asium	2	5	42,53%	41,04%
Asium	2	6	40,57%	38,88%
Asium	3	6	41,39%	39,78%
Asium	4	6	42,61%	41,21%
Occurrences	2	4	43,10%	41,69%
Occurrences	2	5	41,80%	40,01%
Occurrences	2	6	40,73%	38,90%
Occurrences	3	6	42,20%	40,77%
Occurrences	4	6	42,45%	40,91%

Tab. A.3 – Paramètres NbOccMin et NbOccMax

Sélection Couple	*Order*	*NbObj*	*MicroMoy*	*MacroMoy*
Asium	croissant	2	39,18%	37,64%
Asium	croissant	4	40,57%	39,19%
Asium	croissant	6	39,67%	37,74%
Asium	croissant	8	39,84%	38,36%
Asium	croissant	10	39,02%	37,49%
Asium	décroissant	2	42,04%	40,54%
Asium	décroissant	4	40,16%	38,45%
Asium	décroissant	6	38,78%	36,70%
Asium	décroissant	8	38,45%	36,57%
Asium	décroissant	10	38,04%	36,06%
Occurrences	croissant	2	41,96%	40,51%
Occurrences	croissant	4	41,71%	40,40%
Occurrences	croissant	6	41,71%	40,01%
Occurrences	croissant	8	39,18%	37,55%
Occurrences	croissant	10	39,27%	37,23%
Occurrences	décroissant	2	41,22%	39,83%
Occurrences	décroissant	4	40,49%	39,02%
Occurrences	décroissant	6	39,02%	37,24%
Occurrences	décroissant	8	39,43%	38,00%
Occurrences	décroissant	10	37,47%	35,41%

Tab. A.4 – Paramètres NbObj et Order

A.2 Résultats expérimentaux

A.2.1 Résultats pour chaque concept deux à deux

Nous détaillons les résultats obtenus entre chaque concept deux à deux pour les concepts les plus significatifs.

Type de corpus	*MicroMoy*	*MacroMoy*
ExpLSA_1, SA=0,9	**67,04%**	**67,02%**
Corpus original	65,18%	65,25%
ExpLSA_2, SA=0,9	65,18%	65,25%
ExpLSA_1, SA=0,6	63,50%	63,52%
ExpLSA_2, SA=0,6	63,31%	63,42%

Tab. A.5 – Concepts “Activité - Comportement et attitude”

Type de corpus	*MicroMoy*	*MacroMoy*
ExpLSA_2, SA=0,9	**65,56%**	**64,69%**
ExpLSA_1, SA=0,9	**64,73%**	**63,90%**
ExpLSA_1, SA=0,6	**64,06%**	**63,17%**
Corpus original	63,56%	62,62%
ExpLSA_2, SA=0,6	57,24%	55,67%

Tab. A.6 – Concepts “Environnement - Activité”

Type de corpus	*MicroMoy*	*MacroMoy*
ExpLSA_1, SA=0,9	**70,13%**	**70,27%**
ExpLSA_2, SA=0,9	**69,84%**	**70,12%**
Corpus original	69,12%	69,25%
ExpLSA_2, SA=0,6	69,12%	69,34%
ExpLSA_1, SA=0,6	68,54%	68,65%

Tab. A.7 – Concepts “Environnement - Relationnel”

Type de corpus	*MicroMoy*	*MacroMoy*
ExpLSA_2, SA=0,9	**65,56%**	**64,69%**
ExpLSA_1, SA=0,9	**64,73%**	**63,90%**
ExpLSA_1, SA=0,6	**64,06%**	**63,17%**
Corpus original	63,56%	62,62%
ExpLSA_2, SA=0,6	57,24%	55,67%

TAB. A.8 – Concepts “Relationnel - Activité”

Type de corpus	*MicroMoy*	*MacroMoy*
ExpLSA_1, SA=0,9	**70,13%**	**70,27%**
ExpLSA_2, SA=0,9	**69,84%**	**70,12%**
Corpus original	69,12%	69,25%
ExpLSA_2, SA=0,6	69,12%	69,34%
ExpLSA_1, SA=0,6	68,54%	68,65%

TAB. A.9 – Concepts “Relationnel - Comportement et attitude”

Type de corpus	*MicroMoy*	*MacroMoy*
ExpLSA_2, SA=0,9	**66,06%**	**65,21%**
ExpLSA_2, SA=0,6	**65,57%**	**64,72%**
ExpLSA_1, SA=0,9	**63,92%**	**63,00%**
Corpus original	63,26%	62,16%
ExpLSA_1, SA=0,6	63,26%	62,16%

TAB. A.10 – Concepts “Environnement - Comportement et attitude”

Type de corpus	*MicroMoy*	*MacroMoy*
ExpLSA_2, SA=0,9	**67,91%**	**67,34%**
ExpLSA_1, SA=0,9	**67,71%**	**67,03%**
Corpus original	67,00%	66,30%
ExpLSA_1, SA=0,6	66,34%	65,63%
ExpLSA_2, SA=0,6	64,71%	63,90%

TAB. A.11 – Moyenne des résultats expérimentaux

A.2.2 Résultats pour tous les concepts

Résultats obtenus avec l'ensemble des concepts, en ne se limitant pas aux quatre concepts les plus représentatifs.

Type de corpus	*MicroMoy*	*MacroMoy*
ExpLSA_2, SA=0,9	**30,40%**	**11,51%**
ExpLSA_1, SA=0,9	**30,17%**	**9,89%**
Corpus original	29,84%	10,07%
ExpLSA_1, SA=0,6	29,45%	9,47%
ExpLSA_2, SA=0,6	28,17%	9,13%

TAB. A.12 – Résultats pour tous les concepts, incluant les non significatifs

Annexe B

Classification de textes

Nous présentons ici les annexes relatives aux expérimentations du chapitre 4 concernant la tâche de classification de textes.

B.1 Détail des expérimentations dans le choix des paramètres

Nous présentons les différentes valeurs des paramètres que nous avons expérimentés. Ces paramètres sont expérimentés avec un $SA = 0,8$.

Sélection Couple	NbOccMin	*MicroMoy*	*MacroMoy*
Asium	2	74,34%	71,33%
Asium	4	75,55%	72,47%
Asium	6	75,48%	72,68%
Asium	8	74,87%	71,23%
Occurrences	2	73,33%	69,59%
Occurrences	4	75,41%	72,74%
Occurrences	6	75,18%	72,20%
Occurrences	8	74,87%	71,23%

Tab. B.1 – Paramètre NbOccMin

Sélection Couple	NbOccMax	*MicroMoy*	*MacroMoy*
Asium	2	74,87%	71,23%
Asium	4	74,27%	71,13%
Asium	6	74,03%	70,97%
Asium	8	74,03%	70,58%
Occurrences	2	74,87%	71,23%
Occurrences	4	73,73%	70,43%
Occurrences	6	73,39%	70,13%
Occurrences	8	74,20%	70,90%

Tab. B.2 – Paramètre NbOccMax

Sélection Couple	NbOccMin	NbOccMax	*MicroMoy*	*MacroMoy*
Asium	2	4	74,27%	71,13%
Asium	2	5	74,03%	70,97%
Asium	2	6	74,03%	70,97%
Asium	3	5	74,50%	71,27%
Asium	3	6	74,50%	71,27%
Asium	4	6	74,87%	71,23%
Occurrences	2	4	73,73%	70,43%
Occurrences	2	5	73,39%	70,13%
Occurrences	2	6	73,39%	70,13%
Occurrences	3	5	74,94%	71,57%
Occurrences	3	6	74,94%	71,57%
Occurrences	4	6	74,87%	71,23%

Tab. B.3 – Paramètres NbOccMin et NbOccMax

Sélection Couple	*Order*	*NbObj*	*MicroMoy*	*MacroMoy*
Asium	croissant	2	**75,55%**	72,51%
Asium	croissant	4	75,18%	72,33%
Asium	croissant	6	74,84%	71,71%
Asium	croissant	8	74,44%	71,37%
Asium	croissant	10	73,97%	71,01%
Asium	décroissant	2	75,31%	72,30%
Asium	décroissant	4	75,51%	72,93%
Asium	décroissant	6	74,71%	71,53%
Asium	décroissant	8	74,74%	71,78%
Asium	décroissant	10	73,83%	70,77%
Occurrences	croissant	2	75,51%	72,63%
Occurrences	croissant	4	75,41%	72,66%
Occurrences	croissant	6	74,84%	71,79%
Occurrences	croissant	8	74,50%	71,58%
Occurrences	croissant	10	74,03%	71,20%
Occurrences	décroissant	2	75,38%	**72,95%**
Occurrences	décroissant	4	75,51%	72,93%
Occurrences	décroissant	6	75,31%	72,68%
Occurrences	décroissant	8	74,77%	71,95%
Occurrences	décroissant	10	73,76%	70,83%

Tab. B.4 – Paramètres NbObj et Order

B.2 Résultats expérimentaux avec les algorithmes NaiveBayes et k-ppv

Nous présentons les résultats expérimentaux obtenus pour la tâche de classification de textes avec l'approche Bayésienne Naïve (NaiveBayes) et les k plus proches voisins (k-ppv). Ces résultats sont obtenus avec $k = 300$ pour LSA.

	NaiveBayes		K-ppv	
Type	MicroMoy	MacroMoy	MicroMoy	MacroMoy
Asium 0,6	59,67%	57,56%	61,82%	57,79%
Asium 0,9	66,50%	63,96%	68,01%	67,10%
Corpus original	*70,90%*	***68,81%***	***71,51%***	***68,76%***
ExpLSA_1, SA=0,6	69,73%	67,59%	69,42%	66,53%
ExpLSA_1, SA=0,9	**71,24%**	68,85%	71,31%	68,97%
ExpLSA_2, SA=0,6	67,81%	65,86%	67,98%	64,61%
ExpLSA_2, SA=0,9	70,90%	68,81%	**71,51%**	**68,76%**

Tab. B.5 – Corpus de référence

	NaiveBayes		K-ppv	
Type	MicroMoy	MacroMoy	MicroMoy	MacroMoy
Asium 0,6	58,53%	52,67%	68,53%	62,35%
Asium 0,9	63,00%	58,15%	77,50%	72,44%
Corpus original	***76,48%***	***72,00%***	***79,28%***	*74,49%*
ExpLSA_1, SA=0,6	73,71%	69,48%	77,58%	72,26%
ExpLSA_1, SA=0,9	75,92%	71,64%	79,04%	**74,59%**
ExpLSA_2, SA=0,6	71,67%	66,83%	76,26%	70,58%
ExpLSA_2, SA=0,9	58,69%	57,11%	76,10%	71,73%

Tab. B.6 – Corpus de dépêches avec de grands articles

	NaiveBayes		K-ppv	
Type	MicroMoy	MacroMoy	MicroMoy	MacroMoy
Asium 0,6	45,84%	46,36%	56,40%	53,63%
Asium 0,9	59,75%	57,30%	69,48%	65,04%
Corpus original	***74,86%***	***73,54%***	***77,20%***	***74,95%***
ExpLSA_1, SA=0,6	71,04%	69,79%	74,64%	72,09%
ExpLSA_1, SA=0,9	70,96%	70,20%	75,85%	73,62%
ExpLSA_2, SA=0,6	64,38%	63,21%	70,74%	67,50%
ExpLSA_2, SA=0,9	58,95%	58,28%	72,69%	69,95%

Tab. B.7 – Grand corpus des dépêches

	NaiveBayes		K-ppv	
Type	MicroMoy	MacroMoy	MicroMoy	MacroMoy
Asium 0,6	53,99%	53,50%	62,97%	59,50%
Asium 0,9	44,32%	49,08%	70,76%	68,22%
Corpus original	***73,90%***	***72,16%***	***75,28%***	***72,67%***
ExpLSA_1, SA=0,6	71,02%	69,12%	72,08%	69,23%
ExpLSA_1, SA=0,9	72,53%	70,95%	74,27%	71,92%
ExpLSA_2, SA=0,6	69,80%	67,95%	71,67%	68,89%
ExpLSA_2, SA=0,9	51,04%	55,34%	70,48%	68,06%

Tab. B.8 – Corpus de dépêches avec des petits articles

	NaiveBayes		K-ppv	
Type	MicroMoy	MacroMoy	MicroMoy	MacroMoy
Asium 0,6	62,74%	61,29%	60,59%	60,75%
Asium 0,9	68,12%	65,95%	68,53%	65,26%
Corpus original	*68,80%*	*66,74%*	*68,73%*	*66,00%*
ExpLSA_1, SA=0,6	67,38%	65,92%	66,11%	63,00%
ExpLSA_1, SA=0,9	68,93%	**67,53%**	**69,27%**	**68,02%**
ExpLSA_2, SA=0,6	**69,60%**	67,26%	68,12%	62,77%
ExpLSA_2, SA=0,9	68,80%	66,74%	68,73%	66,00%

Tab. B.9 – Petit corpus des dépêches

	NaiveBayes		K-ppv	
Type	MicroMoy	MacroMoy	MicroMoy	MacroMoy
Asium 0,6	59,56%	59,25%	63,96%	60,92%
Asium 0,9	58,99%	61,37%	69,09%	65,77%
Corpus original	***72,96%***	***70,72%***	***74,56%***	***72,05%***
ExpLSA_1, SA=0,6	71,01%	68,90%	71,60%	69,25%
ExpLSA_1, SA=0,9	72,05%	69,70%	72,81%	70,42%
ExpLSA_2, SA=0,6	64,10%	63,10%	67,89%	64,72%
ExpLSA_2, SA=0,9	68,78%	67,75%	71,78%	68,84%

Tab. B.10 – Corpus d'opinion avec de grands articles

	NaiveBayes		K-ppv	
Type	MicroMoy	MacroMoy	MicroMoy	MacroMoy
Asium 0,6	58,52%	55,75%	59,88%	56,70%
Asium 0,9	57,86%	55,00%	59,23%	55,75%
Corpus original	***68,72%***	***67,35%***	***70,18%***	***68,17%***
ExpLSA_1, SA=0,6	65,66%	64,00%	67,92%	65,95%
ExpLSA_1, SA=0,9	66,43%	64,60%	68,53%	66,44%
ExpLSA_2, SA=0,6	61,23%	58,57%	63,26%	60,24%
ExpLSA_2, SA=0,9	58,77%	55,17%	64,03%	60,89%

Tab. B.11 – Grand corpus d'opinion

	NaiveBayes		K-ppv	
Type	MicroMoy	MacroMoy	MicroMoy	MacroMoy
Asium 0,6	58,33%	58,47%	58,50%	56,57%
Asium 0,9	60,22%	59,70%	61,27%	60,15%
Corpus original	*63,93%*	*62,95%*	***63,92%***	***62,41%***
ExpLSA_1, SA=0,6	63,23%	62,40%	60,87%	59,17%
ExpLSA_1, SA=0,9	64,33%	63,47%	63,45%	62,17%
ExpLSA_2, SA=0,6	61,22%	60,50%	61,61%	59,73%
ExpLSA_2, SA=0,9	**64,66%**	**63,72%**	61,95%	60,72%

TAB. B.12 – Corpus d'opinion avec des petits articles

	NaiveBayes		K-ppv	
Type	MicroMoy	MacroMoy	MicroMoy	MacroMoy
Asium 0,6	60,87%	58,17%	63,24%	59,13%
Asium 0,9	64,10%	62,30%	63,47%	59,28%
Corpus original	*65,66%*	*63,90%*	***65,84%***	***62,14%***
ExpLSA_1, SA=0,6	63,35%	61,75%	64,10%	60,07%
ExpLSA_1, SA=0,9	64,97%	62,95%	63,76%	59,60%
ExpLSA_2, SA=0,6	**65,95%**	**64,10%**	65,26%	61,60%
ExpLSA_2, SA=0,9	65,66%	63,90%	**65,84%**	**62,14%**

TAB. B.13 – Petit corpus d'opinion

Annexe C

Données complexes

C.1 Taille de l'espace de représentation d'HYBRED

Nous donnons l'espace de représentation d'HYBRED dans cette annexe.

	Espace de représentation Avec HYBRED								
	NV			NVA			NA		
	3-cara	4-cara	5-cara	3-cara	4-cara	5-cara	3-cara	4-cara	5-cara
Corpus A	1288	1603	1822	1497	2027	2369	1393	1836	2092
Corpus B	1995	3294	4030	2339	4083	5188	2277	3788	4683
Corpus C	846	876	832	1000	1101	1083	906	929	901

Tab. C.1 – Tableau de comparaison de l'espace de recherche avec HYBRED.

C.2 Résultats expérimentaux obtenus avec les corpus A et C

C.2.1 Évaluation des différents descripteurs

Nous donnons les résultats comparatifs des différents descripteurs obtenus sur les corpus A et C.

Algorithmes	k-ppv		SVM		NaiveBayes	
mesure	Fréquentiel	tf-idf	Fréquentiel	tf-idf	Fréquentiel	tf-idf
mot	**95.7**	**96.5**	**97.9**	**97.5**	**96.7**	**96.7**
2-mots	88.7	85.5	90.3	86.7	91.9	89.5
3-mots	77.1	76.3	74.2	74.6	77.1	73.0
2-caractères	94.3	77.9	95.9	85.9	87.9	77.9
3-caractères	**96.3**	94.7	**97.9**	97.9	96.3	93.5
4-caractères	94.3	**97.5**	**97.9**	**98.3**	96.3	94.3
5-caractères	95.5	96.7	97.5	**98.3**	**96.7**	**95.1**
Lemme	95.3	95.1	95.8	**96.7**	95.1	95.9
N	95.9	95.9	97.1	97.1	96.7	97.1
V	84.7	83.5	86.3	83.9	92.7	84.3
NV	**96.3**	**95.9**	97.1	**98.0**	96.7	**96.7**
NVA	95.9	**95.9**	**97.5**	**98.0**	**97.1**	**96.7**
NA	95.5	95.1	**97.5**	**98.0**	**97.1**	**96.7**
VA	93.1	92	95.5	95.0	95.1	91.0

Tab. C.2 – Résultats du corpus A avec les différents descripteurs (précision).

Algorithmes	k-ppv		SVM		NaiveBayes	
mesure	Fréquentiel	tf-idf	Fréquentiel	tf-idf	Fréquentiel	tf-idf
Mot	**96.8**	**95.3**	**98.4**	**98.4**	93.7	**98.4**
2-mots	90.6	89.0	96.8	93.7	**98.4**	93.7
3-mots	84.3	84.3	79.6	79.6	85.9	86.0
2-caractères	84.3	84.3	79.6	79.8	85.9	**85.9**
3-caractères	96.8	98.4	**100**	**100**	93.7	82.8
4-caractères	**98.4**	98.4	**100**	**100**	**95.3**	82.8
5-caractères	**98.4**	**100**	**100**	**100**	90.6	**85.9**
Lemme	90.6	90.6	98.2	**98.4**	92.1	95.3
N	91.1	**93.0**	**95.6**	95.1	93.6	94.6
V	88.2	87.5	88.4	87.8	85.2	84.9
NV	92.4	92.7	95.5	95.5	**94.1**	94.3
NVA	**93.3**	92.6	**95.6**	**95.8**	**94.1**	94.5
NA	92.8	92.4	95.6	95.4	93.9	**94.8**
VA	92.0	91.4	93.7	93.7	91.7	91.4

Tab. C.3 – Résultats du corpus C avec les différents descripteurs (précision).

C.2.2 Évaluation de l'approche HYBRED

Nous donnons les résultats obtenus avec l'approche HYBRED sur les corpus A et C.

Algorithme k-ppv						
Descripteurs	N	V	NV	NVA	NA	VA
2-caractères	75.6	78.4	72.8	71.2	71.2	77.6
3-caractères	92.0	**92.0**	94.0	95.2	94.0	90.4
4-caractères	**94.8**	86.0	**96.8**	**98.0**	**96.8**	**90.8**
5-caractères	94.4	85.2	95.2	96.4	94.0	90.4
Algorithme SVM						
Descripteurs	N	V	NV	NVA	NA	VA
2-caractères	90.0	81.6	88.8	84.4	86.4	87.2
3-caractères	**97.6**	91.2	98.0	97.6	**98.4**	96.4
4-caractères	**97.6**	**95.2**	**98.4**	**98.4**	98.0	**98.0**
5-caractères	96.8	94.0	98.0	**98.4**	98.0	96.4
Algorithme NaiveBayes						
Descripteurs	N	V	NV	NVA	NA	VA
2-caractères	77.6	76.4	78.4	76.8	76.8	77.6
3-caractères	93.6	88.8	90.4	92.0	91.6	91.6
4-caractères	**94.8**	**91.2**	**93.6**	**94.8**	96.0	**92.4**
5-caractères	92.8	90.8	92.0	94.0	**96.4**	91.6

Tab. C.4 – Précision obtenue avec l'approche HYBRED pour le corpus A

Algorithme k-ppv (tf-idf)						
Descripteurs	N	V	NV	NVA	NA	VA
2-caractères	87.5	67.1	84.3	90.6	84.3	79.6
3-caractères	96.8	**82.8**	96.8	98.4	96.8	**87.5**
4-caractères	**100**	75.0	**100**	**100**	**100**	78.1
5-caractères	**100**	81.2	**100**	**100**	**100**	81.2
Algorithme SVM (tf-idf)						
Descripteurs	N	V	NV	NVA	NA	VA
2-caractères	95.3	85.9	93.7	93.7	93.7	90.6
3-caractères	**100**	95.3	**100**	**100**	**100**	96.8
4-caractères	**100**	**96.8**	**100**	**100**	**100**	**100**
5-caractères	**100**	**96.8**	**100**	**100**	**100**	**100**
Algorithme NaiveBayes (tf-idf)						
Descripteurs	N	V	NV	NVA	NA	VA
2-caractères	87.5	67.1	**93.75**	**93.7**	**93.7**	**78.1**
3-caractères	76.5	70.3	84.3	79.6	84.3	76.5
4-caractères	**78.1**	70.3	79.6	73.4	79.6	71.8
5-caractères	76.5	**73.4**	71.8	75.0	71.8	65.6

Tab. C.5 – Précision obtenue avec l'approche HYBRED pour le corpus C

Annexe D

Construction classes conceptuelles

D.1 Détail des résultats pour Hypon

Nous détaillons ici les résultats obtenus avec l'approche Hypon pour différentes valeurs du paramètre k.

	HYPON (> 1,5)										
nb rel.	***K=0***	***0,1***	***0,2***	***0,3***	***0,4***	***0,5***	***0,6***	***0,7***	***0,8***	***0,9***	***K=1***
50	0,62	0,74	0,73	**0,79**	0,76	0,7	0,75	0,71	0,72	0,5	0,54
100	0,53	0,54	0,58	0,54	0,69	0,67	0,7	0,71	**0,71**	0,64	0,57
150	0,58	0,59	0,69	0,69	0,71	0,76	0,77	0,78	**0,78**	0,74	0,52
200	0,61	0,68	0,73	**0,75**	0,64	0,58	0,58	0,61	0,63	0,64	0,5
250	0,57	**0,68**	0,61	0,56	0,47	0,47	0,5	0,47	0,46	0,47	0,42
300	0,52	**0,61**	0,54	0,49	0,45	0,42	0,39	0,39	0,39	0,38	0,34
350	**0,56**	0,55	0,52	0,51	0,48	0,47	0,46	0,46	0,46	0,45	0,42
400	0,58	**0,6**	0,55	0,53	0,51	0,5	0,48	0,48	0,49	0,49	0,47
450	**0,6**	0,55	0,54	0,52	0,5	0,49	0,48	0,48	0,48	0,48	0,46
500	**0,57**	0,51	0,48	0,47	0,46	0,45	0,44	0,44	0,44	0,44	0,41
550	**0,53**	0,46	0,44	0,43	0,42	0,42	0,41	0,41	0,41	0,41	0,39

Tab. D.1 – Détail des résultats de l'approche Hypon pour la moyenne

	Hypon (score = 2 , 75% des experts)										
nb rel.	**K=0**	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**	**K=1**
50	0,29	0,41	0,51	0,64	0,55	0,63	0,67	0,72	0,64	0,56	0,5
100	0,46	0,46	0,45	0,53	0,55	0,7	0,67	0,7	0,71	0,65	0,66
150	0,42	0,42	0,41	0,55	0,52	0,65	0,74	0,76	0,77	0,74	0,55
200	0,45	0,52	0,52	0,64	0,63	0,65	0,7	0,67	0,68	0,67	0,57
250	0,52	0,61	0,62	0,57	0,59	0,55	0,52	0,52	0,58	0,56	0,53
300	0,53	0,59	0,58	0,52	0,51	0,43	0,45	0,42	0,39	0,38	0,35
350	0,57	0,53	0,51	0,46	0,44	0,41	0,41	0,41	0,43	0,44	0,42
400	0,6	0,56	0,54	0,5	0,5	0,45	0,46	0,46	0,46	0,46	0,46
450	0,58	0,57	0,56	0,53	0,52	0,5	0,49	0,49	0,49	0,5	0,46
500	0,52	0,48	0,48	0,45	0,45	0,43	0,42	0,42	0,42	0,42	0,41
550	0,48	0,44	0,43	0,42	0,41	0,4	0,4	0,4	0,4	0,4	0,39

Tab. D.2 – Détail des résultats de l'approche Hypon pour le vote

D.2 Résultats expérimentaux pour les autres critères pour le vote et la moyenne

Nous présentons ici les résultats obtenus avec différents critères pour le vote et la moyenne.

	Seuil de validation								
	>2			> 1,5			>1		
nb relations	**AUC**	**+**	**-**	**AUC**	**+**	**-**	**AUC**	**+**	**-**
50	0	0	50	0,54	7	43	0,52	11	39
100	0	0	100	0,57	11	89	0,48	24	76
150	0	0	150	0,52	16	134	0,45	38	112
200	0	0	200	0,5	22	178	0,44	54	146
250	0	1	249	0,42	33	217	0,44	70	180
300	0,17	1	299	0,34	53	247	0,37	101	199
350	0,29	1	349	0,42	57	293	0,42	115	235
400	0,38	1	399	0,47	61	339	0,43	134	266
450	0,45	1	449	0,46	71	379	0,43	153	297
500	0,5	1	499	0,41	88	412	0,43	173	327
550	0,55	1	549	0,39	104	446	0,42	196	354

Tab. D.3 – Résultats avec VS pour les critères de la moyenne et du vote

	Critères de validation											
	2 pour 50 %			1 pour 50%			2 pour 75%			1 pour 75%		
nb rel.	*AUC*	+	-	*AUC*	+	-	*AUC*	+	-	*AUC*	+	-
50	0,51	8	42	0,49	16	34	0,5	6	44	0,45	10	40
100	0,52	15	85	0,48	34	66	0,66	8	92	0,45	23	77
150	0,48	23	127	0,43	56	94	0,55	12	138	0,43	37	113
200	0,44	34	166	0,42	80	120	0,57	15	185	0,45	50	150
250	0,43	45	205	0,44	101	149	0,53	19	231	0,44	66	184
300	0,35	69	231	0,4	133	167	0,35	37	263	0,38	93	207
350	0,41	78	272	0,43	155	195	0,42	40	310	0,42	108	242
400	0,45	87	313	0,41	185	215	0,46	44	356	0,43	124	276
450	0,43	102	348	0,42	211	239	0,46	50	400	0,44	141	309
500	0,41	120	380	0,44	234	266	0,41	63	437	0,44	160	340
550	0,4	138	412	0,43	261	289	0,39	75	475	0,43	182	368

Tab. D.4 – Résultats avec VS pour les critères du vote

> 2	**Fréquence**			**IM**			**IM³**			**Dice**		
nb rel.	***AUC***	**+**	**-**	***AUC***	**+**	**-**	***AUC***	**+**	**-**	***AUC***	**+**	**-**
50	0	0	50	0	0	50	0	0	50	0	0	50
100	0,2	1	99	0,17	1	99	**0,29**	1	99	0,22	1	99
150	0,47	1	149	0,45	1	149	**0,53**	1	149	0,48	1	149
200	0,6	1	199	0,59	1	199	**0,65**	1	199	0,61	1	199
250	0,68	1	249	0,67	1	249	**0,72**	1	249	0,69	1	249
300	0,74	1	299	0,73	1	299	**0,77**	1	299	0,74	1	299
350	0,77	1	349	0,77	1	349	**0,8**	1	349	0,78	1	349
400	0,8	1	399	0,79	1	399	**0,82**	1	399	0,81	1	399
450	0,82	1	449	0,82	1	449	**0,84**	1	449	0,83	1	449
500	0,84	1	499	0,84	1	499	**0,86**	1	499	0,85	1	499
550	0,86	1	549	0,85	1	549	**0,87**	1	549	0,86	1	549

>1,5	**Fréquence**			**IM**			**IM³**			**Dice**		
nb rel.	***AUC***	**+**	**-**	***AUC***	**+**	**-**	***AUC***	**+**	**-**	***AUC***	**+**	**-**
50	0,47	2	48	**0,62**	14	36	0,28	8	42	0,48	10	40
100	0,28	16	84	**0,53**	26	74	0,42	18	82	0,51	19	81
150	0,41	25	125	**0,58**	35	115	0,4	30	120	0,44	34	116
200	0,44	35	165	**0,61**	41	159	0,44	41	159	0,5	43	157
250	0,5	40	210	**0,57**	51	199	0,5	48	202	0,51	52	198
300	0,48	50	250	0,52	65	235	**0,53**	54	246	0,53	60	240
350	0,47	60	290	**0,56**	70	280	0,53	62	288	0,55	67	283
400	0,49	67	333	**0,58**	75	325	0,54	69	331	0,57	72	328
450	0,47	78	372	**0,6**	79	371	0,54	77	373	0,57	79	371
500	0,46	90	410	**0,57**	90	410	0,5	90	410	0,54	90	410
550	0,44	104	446	**0,53**	104	446	0,47	104	446	0,51	104	446

> 1	**Fréquence**			**IM**			**IM³**			**Dice**		
nb rel.	***AUC***	**+**	**-**	***AUC***	**+**	**-**	***AUC***	**+**	**-**	***AUC***	**+**	**-**
50	**0,68**	9	41	0,55	26	24	0,46	16	34	0,53	20	30
100	0,32	33	67	**0,57**	45	55	0,43	36	64	0,52	38	62
150	0,42	51	99	**0,62**	60	90	0,44	55	95	0,49	59	91
200	0,44	69	131	**0,61**	76	124	0,46	75	125	0,5	78	122
250	0,49	82	168	**0,55**	97	153	0,49	92	158	0,47	100	150
300	0,47	103	197	0,5	123	177	**0,51**	108	192	0,5	117	183
350	0,45	125	225	**0,54**	136	214	0,5	127	223	0,5	136	214
400	0,48	139	261	**0,57**	147	253	0,51	144	256	0,55	146	254
450	0,48	159	291	**0,57**	161	289	0,52	159	291	0,56	159	291
500	0,48	176	324	**0,57**	176	324	0,52	176	324	0,55	176	324
550	0,47	197	353	**0,55**	197	353	0,5	197	353	0,53	197	353

Tab. D.5 – Résultats avec VW pour les critères de la moyenne

2,50%	Fréquence			IM			IM³			Dice		
nb rel.	*AUC*	+	-	*AUC*	+	-	*AUC*	+	-	*AUC*	+	-
50	**0,63**	4	46	0,55	19	31	0,33	11	39	0,45	14	36
100	0,29	22	78	**0,54**	34	66	0,41	25	75	0,5	27	73
150	0,38	36	114	**0,6**	44	106	0,41	40	110	0,47	44	106
200	0,42	49	151	**0,62**	53	147	0,44	54	146	0,51	56	144
250	0,5	56	194	**0,57**	67	183	0,5	64	186	0,5	70	180
300	0,47	70	230	0,51	86	214	**0,52**	74	226	0,52	81	219
350	0,46	84	266	0,53	96	254	0,51	87	263	**0,53**	92	258
400	0,49	94	306	0,56	104	296	0,53	96	304	**0,56**	100	300
450	0,48	108	342	**0,58**	110	340	0,52	108	342	0,56	110	340
500	0,47	122	378	**0,57**	122	378	0,51	122	378	0,55	122	378
550	0,46	139	411	**0,53**	139	411	0,49	139	411	0,52	139	411

1,50%	Fréquence			IM			IM³			Dice		
nb rel.	*AUC*	+	-	*AUC*	+	-	*AUC*	+	-	*AUC*	+	-
50	**0,65**	16	34	0,54	34	16	0,56	21	29	0,64	25	25
100	0,37	46	54	**0,62**	56	44	0,45	47	53	0,49	49	51
150	0,43	71	79	**0,62**	78	72	0,42	75	75	0,48	77	73
200	0,46	95	105	**0,6**	100	100	0,47	99	101	0,47	104	96
250	0,51	113	137	**0,54**	128	122	0,47	125	125	0,47	131	119
300	0,48	141	159	0,49	161	139	**0,5**	147	153	0,49	154	146
350	0,45	172	178	**0,52**	181	169	0,49	173	177	0,49	181	169
400	0,49	191	209	**0,56**	198	202	0,5	197	203	0,53	198	202
450	0,49	216	234	**0,56**	218	232	0,52	217	233	0,54	218	232
500	0,5	237	263	**0,57**	237	263	0,53	237	263	0,55	237	263
550	0,5	261	289	**0,56**	261	289	0,52	261	289	0,54	261	289

2,75%	Fréquence			IM			IM³			Dice		
nb rel.	*AUC*	+	-	*AUC*	+	-	*AUC*	+	-	*AUC*	+	-
50	0,47	2	48	**0,64**	10	40	0,29	7	43	0,49	8	42
100	0,29	14	86	0,5	20	80	0,46	14	86	**0,53**	14	86
150	0,46	19	131	**0,62**	24	126	0,42	23	127	0,47	25	125
200	0,45	28	172	**0,61**	29	171	0,45	32	168	0,53	31	169
250	0,52	31	219	**0,56**	37	213	0,52	36	214	0,51	39	211
300	0,5	38	262	0,51	47	253	0,53	42	258	**0,53**	45	255
350	0,48	46	304	0,57	49	301	0,57	45	305	**0,58**	47	303
400	0,54	47	353	0,59	52	348	0,6	47	353	**0,6**	50	350
450	0,52	54	396	**0,61**	55	395	0,58	53	397	0,6	55	395
500	0,49	64	436	**0,56**	64	436	0,52	64	436	0,55	64	436
550	0,46	75	475	**0,52**	75	475	0,48	75	475	0,51	75	475

1,75%	Fréquence			IM			IM³			Dice		
nb rel.	*AUC*	+	-	*AUC*	+	-	*AUC*	+	-	*AUC*	+	-
50	0,55	7	43	**0,6**	23	27	0,39	13	37	0,53	16	34
100	0,3	31	69	**0,56**	42	58	0,41	32	68	0,4	37	63
150	0,41	48	102	**0,59**	58	92	0,4	53	97	0,43	58	92
200	0,45	64	136	**0,58**	74	126	0,43	73	127	0,48	75	125
250	0,49	76	174	**0,54**	94	156	0,48	88	162	0,46	96	154
300	0,46	97	203	**0,5**	117	183	0,5	103	197	0,5	112	188
350	0,45	118	232	**0,54**	130	220	0,51	118	232	0,51	128	222
400	0,49	129	271	**0,57**	139	261	0,52	134	266	0,55	137	263
450	0,48	148	302	**0,59**	150	300	0,53	148	302	0,56	150	300
500	0,48	164	336	**0,58**	164	336	0,52	164	336	0,56	164	336
550	0,48	183	367	**0,56**	183	367	0,51	183	367	0,54	183	367

Tab. D.6 – Résultats avec VW pour les critères du vote

Bibliographie

[Abney & Abney, 1990] Steven Abney et Steven P. Abney. "Rapid Incremental Parsing with Repair". Dans les actes de *University of Waterloo*, pp 1–9, 1990.

[Abney, 1991] S. Abney. "Parsing by Chunks". 1991.

[Abney, 1996] Steven Abney. "Partial parsing via finite-state cascades". *Natural Language Engineering*, pp 337–344, 1996.

[Adam, 1999] J.-M. Adam. *Linguistique textuelle. Des genres de discours aux textes*. Nathan, Paris, 1999.

[Aggarwal *et al.*, 1998] Charu C. Aggarwal, Zheng Sun, et Philip S. Yu. "Online Algorithms for Finding Profile Association Rules". Dans les actes de *CIKM*, pp 86–95, 1998.

[Aït-Mokhtar & Chanod, 1997] Salah Aït-Mokhtar et Jean-Pierre Chanod. "Incremental finite-state parsing". Dans les actes de *Proceedings of the fifth conference on Applied natural language processing*, pp 72–79, Morristown, NJ, USA, 1997. Association for Computational Linguistics.

[Aseervatham, 2008] Sujeevan Aseervatham. "A Local Latent Semantic Analysis-based Kernel for Document Similarities". Dans les actes de *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on Neural Networks IEEE International Joint Conference on Neural Networks*, pp 214–219, 2008.

[Aussenac-Gilles & Jacques, 2006] Nathalie Aussenac-Gilles et Marie-Paule Jacques. "Designing and Evaluating Patterns for Ontology Enrichment from Texts". Dans les actes de *EKAW*, pp 158–165, 2006.

[Bacchin *et al.*, 2005] Michela Bacchin, Nicola Ferro, et Massimo Melucci.

"A probabilistic model for stemmer generation". *Inf. Process. Manage.*, pp 121–137, 2005, Pergamon Press, Inc.

[Bagein *et al.*, 2001] Michel Bagein, Thierry Dutoit, Nawfal Tounsi, Fabrice Malfrere, Alain Ruelle, et Dominique Wynsberghe. "Le projet Euler : Vers une synthèse de parole générique et multilingue". Dans les actes de *TAL, Traitement automatique des langues*, pp 275–296. Association pour le traitement automatique des langues, Paris, 2001.

[Baroni & Bernardini, 2004] Marco Baroni et Silvia Bernardini. "BootCaT : Bootstrapping Corpora and Terms from the Web". Dans les actes de *In Proceedings of LREC 2004*, pp 1313–1316, 2004.

[Basili *et al.*, 2000] Roberto Basili, Alessandro Moschitti, Ro Moschitti, et Maria Teresa Pazienza. "Language Sensitive Text Classification". Dans les actes de *In In proceeding of 6th RIAO Conference (RIAO 2000), Content-Based Multimedia Information Access, Coll ge de*, 2000.

[Basili *et al.*, 2001] Roberto Basili, Alessandro Moschitti, et Maria Teresa Pazienza. "A Hybrid Approach to Optimize Feature Selection Process in Text Classification". Dans les actes de *AI*IA 01 : Proceedings of the 7th Congress of the Italian Association for Artificial Intelligence on Advances in Artificial Intelligence*, pp 320–326, London, UK, 2001. Springer-Verlag.

[Baum & Petrie, 1966] Leonard E. Baum et Ted Petrie. "Statistical Inference for Probabilistic Functions of Finite State Markov Chains". *The Annals of Mathematical Statistics*, pp 1554–1563, 1966, Institute of Mathematical Statistics.

[Bayes, 1763] T. Bayes. "An essay towards solving a problem in the doctrine of chances.". *Philosophical Transactions of the Royal Soc. of London*, pp 370–418, 1763.

[Béchet *et al.*, 2007] Nicolas Béchet, Mathieu Roche, et Jacques Chauché. "Improving LSA by expanding the contexts". Dans les actes de *Context-Based Information Retrieval (CIR) workshop - CONTEXT'07*, pp 105–108, 2007.

[Bellot & El-Bèze, 2001] Patrice Bellot et Marc El-Bèze. Classification et segmentation de textes par arbres de décision. Dans les actes

de *Technique et Science Informatiques (TSI)*, volume 20, pp 107–134. Hermès, 2001.

[Benamara *et al.*, 2007] F. Benamara, C. Cesarano, A. Picariello, D. Reforgiato, et V.S Subrahmanian. "Sentiment Analysis : Adjectives and Adverbs are better than Adjectives Alone". Dans les actes de *IADIS Applied Computing, Boulder, Colorado, U.S.A, 26/03/07-28/03/07*, pp 203–206. ACM, 2007.

[Bergo, 2001] Alexander Bergo. "Text Categorization and Prototypes". 2001.

[Besançon & Chalendar, 2005] R. Besançon et G. De Chalendar. "L'analyseur syntaxique de LIMA dans la campagne d'évaluation EASy". Dans les actes de *Proceedings of TALN'05*, 2005.

[Besançon, 2001] Romaric Besançon. "*Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles de textes*". PhD thesis, 2001.

[Bezdek, 1981] James C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA, 1981.

[Bizer & Rainer, 2005] Ralf Heese Bizer et Eckstein Rainer. "Impact of Semantic web on the job recruitment Process". *International Conference Wirtschaftsinformatik*, 2005.

[Borkowski, 1969] Casimir Borkowski. "Structure effectiveness and uses of The Citation Identifier an operational computer program for automatic identification of case citations in legal literature". Dans les actes de *Proceedings of the 1969 conference on Computational linguistics*, pp 1–22. Association for Computational Linguistics, 1969.

[Bouquiaux, 1987] L. Bouquiaux. *Enquête et description des langues à tradition orale, I-II-III*. J.M.C.T. Thomas (eds.), Paris, 1987.

[Bourigault & Fabre, 2000] D. Bourigault et C. Fabre. "Approche linguistique pour l'analyse syntaxique de corpus". *Cahiers de Grammaires*, pp 131–151, 2000.

[Bourigault & Lame, 2002] D. Bourigault et G. Lame. "Analyse distributionnelle et structuration de terminologie. Application à la con-

struction d'une ontologie documentaire du Droit". Dans les actes de *TAL*, pp 43–51, 2002.

[Bourigault, 1993] D. Bourigault. "Analyse syntaxique locale pour le repérage de termes complexes dans un texte". *T.A.L.*, pp 105–118, 1993.

[Bourigault, 1994] D. Bourigault. "*LEXTER un Logiciel d'EXtraction de TERminologie. Application à l'extraction des connaissances à partir de textes.*". Thèse en mathématiques, informatique appliquée aux sciences de l'homme, École des hautes Études en sciences sociales, Paris, 1994.

[Bourigault, 2002] D. Bourigault. "UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus". Dans les actes de *Actes de TALN, Nancy*, pp 75–84, 2002.

[Bourse *et al.*, 2004] Michel Bourse, Michel Leclère, Emmanuel Morin, et Francky Trichet. "Human Resource Management and Semantic Web Technologies". Dans les actes de *ICTTA*, pp 641–642, 2004.

[Bowers, 2001] John Bowers. "Syntactic Relations. Manuscript". 2001.

[Boyd *et al.*, 1993] Richard Boyd, James R. Driscoll, et Inien Syu. "Incorporating Semantics Within a Connectionist Model and a Vector Processing Model". Dans les actes de *TREC*, pp 291–302, 1993.

[Braschler & Ripplinger, 2004] Martin Braschler et Bärbel Ripplinger. "How Effective is Stemming and Decompounding for German Text Retrieval ?". *Inf. Retr.*, pp 291–316, 2004, Kluwer Academic Publishers.

[Breiman *et al.*, 1984] Leo Breiman, Jerome Friedman, Charles J. Stone, et R. A. Olshen. *Classification and Regression Trees*. Chapman & Hall/CRC, 1984.

[Brill, 1994] E. Brill. "Some Advances in Transformation-Based Part of Speech Tagging". Dans les actes de *AAAI, Vol. 1*, pp 722–727, 1994.

[Brunet, 2002] E. Brunet. "Le Lemme comme on l'aime". Dans les actes de *JADT 2002, 6e Journées internationales d'analyse des*

données textuelles, pp 221–232. Morin A. et Sébillot P. (éd.), 2002.

[Cardie, 1997] Claire Cardie. "Empirical Methods in Information Extraction". *AI magazine*, pp 65–79, 1997.

[Cederberg & Widdows, 2003] Scott Cederberg et Dominic Widdows. "Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction". Dans les actes de *In Proceedings of CoNLL*, pp 111–118, 2003.

[Chauché & Prince, 2007] Jacques Chauché et Violaine Prince. "Classifying texts through natural language parsing and semantic filtering". Dans les actes de *3rd International Language and Technology Conference, Poznan, Pologne*, 2007.

[Chauché, 1984] Jacques Chauché. "Un outil multidimensionnel de l'analyse du discours". Dans les actes de *Proceedings of Coling, Standford University, California*, pp 11–15, 1984.

[Chauché, 1990] J. Chauché. "Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance". Dans les actes de *TA Information*, pp 17–24, 1990.

[Chauché *et al.*, 2003] J. Chauché, V. Prince, S. Jaillet, et M. Teisseire. "Classification automatique de textes à partir de leur analyse syntaxico-sémantique". *TALN'03*, pp 45–55, 2003.

[Chauché, 2007] Jacques Chauché. "*SYGMART : Manuel de référence Version 5.2*", 2007.

[Choi, 2000] F. Y. Y. Choi. "Advances in domain independent linear text segmentation". Dans les actes de *the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, USA*, 2000.

[Choong *et al.*, 2006] Yeow Wei Choong, Anne Laurent, et Dominique Laurent. "Pixelizing Data Cubes : A Block-Based Approach". Dans les actes de *VIEW*, pp 63–76, 2006.

[Church & Hanks, 1989] K. W. Church et P. Hanks. "Word Association Norms, Mutual Information, and Lexicography". Dans les actes de *Proceedings of the 27th Annual Conference of the Associ-*

ation of Computational Linguistic, volume 16, pp 76–83, 1989.

[Church *et al.*, 1991] Kenneth Church, William Gale, Patrick Hanks, et Donald Hindle. "Using statistics in lexical analysis". Dans les actes de *Lexical Acquisition : Exploiting On-Line Resources to Build a Lexicon*, pp 115–164. Erlbaum, 1991.

[Cilibrasi & Vitanyi, 2007] Rudi Cilibrasi et Paul M. B. Vitanyi. "The Google Similarity Distance". *IEEE Transactions on Knowledge and Data Engineering*, page 370, 2007.

[Cornish *et al.*, 2005] Francis Cornish, Alan Garnham, Wind H. Cowles, Marion Fossard, et Virginie André. "Indirect anaphora in English and French : A cross-linguistic study of pronoun resolution.". *Journal of Memory and Language*, pp p363–376, 2005.

[Cornuéjols & Miclet, 2002] A. Cornuéjols et L. Miclet. *Apprentissage artificiel, Concepts et algorithmes*. Eyrolles, 2002.

[Courtine, 1981] J.-J. Courtine. *Analyse du discours politique*. Langages 62, Larousse, 1981.

[Cover & Hart, 1967] T. Cover et P. Hart. "Nearest neighbor pattern classification". *Information Theory, IEEE Transactions on*, pp 21–27, 1967.

[Cybenko, 1989] G. Cybenko. "Approximation by superpositions of a sigmoidal function". *Mathematics of Control, Signals, and Systems (MCSS)*, pp 303–314, 1989.

[Daille, 1994] B. Daille. "*Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*". PhD thesis, Université Paris 7, 1994.

[Daille, 1996] B. Daille. "Study and Implementation of Combined Techniques for Automatic Extraction of Terminology". Dans les actes de *P. Resnik and J. Klavans (eds). The Balancing Act : Combining Symbolic and Statistical Approaches to Language, MIT Press*, pp 49–66, 1996.

[David & Plante, 1990] S. David et P. Plante. "De la nécessité d'une approche morpho syntaxique dans l'analyse de textes". Dans les actes de *Intelligence Artificielle et Sciences Cognitives au Quebec*, volume 3, pp 140–154, 1990.

[de Saussure, 1916] Ferdinand DE SAUSSURE. "Cours de linguistique générale". 1916.

[Deerwester *et al.*, 1990] S.T. DEERWESTER, G.W. DUMAIS, T.K. LAUNDER, et R. HARSHMANN. "Indexing by Latent Semantic Analysis". Dans les actes de *Journal of the American Society for Information Science 41*, pp 391–407, 1990.

[Dempster *et al.*, 1977] A. P. DEMPSTER, N. M. LAIRD, et D. B. RUBIN. "Maximum likelihood from incomplete data via the EM algorithm". *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, pp 1–38, 1977.

[Denoyer, 2004] Ludovic DENOYER. "*Apprentissage et Inférence Statistique dans les bases de documents structurés*". Phd thesis, University of Paris VI, LIP6, 8 rue du capitaine Scott, 75015 PARIS, dec 2004.

[Desrosiers-Sabbath, 1984] R. DESROSIERS-SABBATH. *Comment enseigner les concepts*. 1984.

[Devlin & Murphy, 1988] B. A. DEVLIN et P. T. MURPHY. "An architecture for a business and information system". *IBM Syst. J.*, pp 60–80, 1988, IBM Corp.

[Dini & Mazzini, 2002] Luca DINI et Giampaolo MAZZINI. "Opinion classification through Information Extraction". Dans les actes de *Proceedings of the Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields (Data Mining)*, pp 299–310, 2002.

[Dodge, 2007] Yadolah DODGE. *Statistique : dictionnaire encyclopédique*. Springer, Paris, 2007.

[Dongbo, 2001] Pang Jianfengand Bu DONGBO. "Research and implementation of text categorization system based on VSM". Dans les actes de *Application Research of Computers*, pp 23–26, 2001.

[Dorier, 1996] Jean-Luc DORIER. "Genèse des premiers espaces vectoriels de fonctions". Dans les actes de *Revue d'histoire des mathématiques, vol. 2*, pp 265–307. Société mathématique de France, Paris, 1996.

[Duda *et al.*, 2001] Richard O. DUDA, Peter E. HART, et David G. STORK. *Pattern Classification*. John Wiley & Sons, Inc., 2001.

[Dumais, 1991] Susan T. Dumais. "Improving the retrieval of information from external sources". Dans les actes de *Revue d'histoire des mathématiques, vol. 2*, pp 229–236. Behavior Research Methods, Instruments and Computers, 1991.

[Dunning, 1993] Ted Dunning. "Accurate Methods for the Statistics of Surprise and Coincidence". *Computational Linguistics*, pp 61–74, 1993.

[Dunning, 1994] Ted Dunning. "Statistical Identification of Language". 1994.

[Dutoit, 1997] Thierry Dutoit. *An introduction to text-to-speech synthesis*. Kluwer Academic Publishers, 1997.

[Enguehard, 1993] C. Enguehard. "Acquisition de terminologie à partir de gros corpus". Dans les actes de *Informatique & Langue Naturelle, ILN'93*, pp 373–384, 1993.

[Enguehard, 2001] C. Enguehard. "Apprentissage de schémas lexicaux pour l'acquisition de candidats termes". Dans les actes de *Actes des Journée Applications, Apprentissage et Acquisition de Connaissances à partir de Textes électroniques (A3CTE)*, pp 17–25, 2001.

[Fabre & Bourigault, 2006] C. Fabre et D. Bourigault. "Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques". Dans les actes de *TALN'06, 10-13 avril 2006*, pp 121–129, 2006.

[Fano & Hawkins, 1961] Robert M. Fano et David Hawkins. *Transmission of Information : A Statistical Theory of Communications*. London : M.I.T. Press & Wiley, 1961.

[Faure & Nedellec, 1999] D. Faure et C. Nedellec. "Knowledge acquisition of predicate argument structures from technical texts using machine learning : The system ASIUM". Dans les actes de *Proceedings of the 11th European Workshop, Knowledge Acquisition, Modelling and Management, number 1937 in LNAI*, pp 329–334, 1999.

[Faure, 2000] D. Faure. "*Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM*". PhD thesis, Université Paris-Sud, 20 Décembre 2000.

[Fei *et al.*, 2004] Zhongchao Fei, Jian Liu, et Gengfeng Wu. "Sentiment Classification Using Phrase Patterns". *Computer and Information Technology, International Conference on*, pp 1147–1152, 2004.

[Ferri *et al.*, 2002] C. Ferri, P. Flach, et J. Hernandez-Orallo. "Learning decision trees using the area under the ROC curve". Dans les actes de *Proceedings of ICML'02*, pp 139–146, 2002.

[Fisher *et al.*, 2004] Michelle Fisher, Jonathan E. Fieldsend, et Richard M. Everson. "Precision and Recall Optimisation for Information Access Tasks". Dans les actes de *The 1st Workshop on ROC Analysis in Artificial Intelligence (ROCAI-2004), part of ECAI-2004*, pp 45–54, 2004.

[Fluhr, 1997] C. Fluhr. "SPIRIT.W3 : A distributed cross-lingual indexing and search engine". Dans les actes de *Proceedings of the INET 97, The seventh annual conference of the internet society*, June 24-27, Kuala Lumpur, Malaysia, 1997.

[Foltz & Dumais, 1992] Peter W. Foltz et Susan T. Dumais. "Personalized information delivery : an analysis of information filtering methods". *Communications of the ACM*, pp 51–60, 1992.

[Foltz, 1996] Peter W. Foltz. "Latent Semantic Analysis for Text-Based Research". *Behavior research methods instruments and computers*, pp 197–202, 1996.

[Frakes & Baeza-Yates, 1992] William Frakes et Ricardo Baeza-Yates. *Information Retrieval : Data Structures & Algorithms*. Prentice-Hall, 1992.

[Frath *et al.*, 2000] P. Frath, R. Oueslati, et F. Rousselot. "*Genetic programming applied to model identification*", pp 291–304. Eyrolles, Paris, 2000.

[Fürnkranz, 1998] Johannes Fürnkranz. "A Study Using n-gram Features for Text Categorization". 1998.

[Gabrilovich & Markovitch, 2005] Evgeniy Gabrilovich et Shaul Markovitch. "Feature generation for text categorization using world knowledge". Dans les actes de *In IJCAI'05*, pp 1048–1053, 2005.

[Gabrilovich & Markovitch, 2006] Evgeniy Gabrilovich et Shaul Markovitch. "Overcoming the brittleness bottleneck using Wikipedia :

enhancing text categorization with encyclopedic knowledge". Dans les actes de *Twenty-First AAAI Conference on Artificial Intelligence*, 2006.

[Galy & Bourigault, 2005] E. Galy et D. Bourigault. "Analyse distributionnelle de corpus de langue générale et synonymie". Dans les actes de *JLC'05*, Lorient, 2005.

[Garvin, 1967] Paul L. Garvin. "The fulcrum syntactic analyzer for Russian". Dans les actes de *Proceedings of the 1967 conference on Computational linguistics*, pp 1–10, Morristown, NJ, USA, 1967. Association for Computational Linguistics.

[Genereux & Santini, 2007] M. Genereux et M. Santini. "Defi : classification de textes Francais subjectifs". Dans les actes de *In : 3eme DEfi fouille de textes*, Grenoble, Switzerland, 2007.

[Gonçalves & Quaresma, 2005] T. Gonçalves et P. Quaresma. *Evaluating preprocessing techniques in a Text Classification problem.* SBC - Sociedade Brasileira de Computação, São Leopoldo, RS, Brasil, July 2005.

[Goodman & Smyth, 1988] Rodney M. Goodman et Padhraic Smyth. "Information-Theoretic Rule Induction.". Dans les actes de *ECAI*, pp 357–362, 1988.

[Greevy & Smeaton, 2004] E. Greevy et Alan F. Smeaton. "Classifying racist texts using a support vector machine". Dans les actes de *SIGIR '04 : Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp 468–469, New York, NY, USA, 2004. ACM.

[Grefenstette, 1995] G. Grefenstette. "Comparing Two Language Identification Schemes". Dans les actes de *the 3rd International Conference on the Statistical Analysis of Textual Data (JADT'95)*, 1995.

[Grefenstette, 1999] Gregory Grefenstette. "Light parsing as finite state filtering". pp 86–94, 1999, Cambridge University Press.

[Guarino, 1998] Nicola Guarino. "Formal Ontology and Information Systems". pp 3–15. IOS Press, 1998.

[Guo & Murphey, 2000] Hong Guo et Yi Lu Murphey. "Automatic Feature Selection - A Hybrid Statistical Approach". *Pattern Recogni-*

tion, International Conference on, page 2382, 2000, IEEE Computer Society.

[Habert & Nazarenko, 1996] Benoît Habert et Adeline Nazarenko. "La syntaxe comme marche-pied de l'acquisition des connaissances : bilan critique d'une expérience". Dans les actes de *Journées sur l'acquisition des connaissances*, pp 137–142, Sète, mai 1996. AFIA.

[Harris, 1951] Zellig Harris. *Structural Linguistics*. The University of Chicago Press, 1951.

[Harris, 1968] Z. Harris. *Mathematical Structures of Language*. John Wiley & Sons, New-York, 1968.

[Hausmann, 1989] F. J. Hausmann. "Le dictionnaire de collocations". Dans les actes de *Hausmann F.J., Reichmann O., Wiegand H.E., Zgusta L. (eds), Wörterbücher : ein internationales Handbuch zur Lexicographie. Dictionaries. Dictionnaires. Berlin/New-York : De Gruyter*, pp 1010–1019, 1989.

[Hawking & Thistlewaite, 1998] David Hawking et Paul Thistlewaite. "Overview of TREC-6 Very Large Collection Track". Dans les actes de *In Proceedings of the Sixth Text REtrieval Conference (TREC-6*, pp 93–106, 1998.

[Hebb, 1961] D. Hebb. *Organization of behavior*. Science Edition, 1961.

[Hignette *et al.*, 2007] Gaëlle Hignette, Patrice Buche, Juliette Dibie-Barthélemy, et Ollivier Haemmerlé. "Annotation sémantique floue de tableaux guidée par une ontologie.". Dans les actes de *EGC*, pp 587–598, 2007.

[Hjelmslev, 1968] Louis Hjelmslev. *Prolégomènes à une théorie du langage*. Éditions de Minuit (Paris), 1968.

[Hobbs, 1993] Jerry R. Hobbs. "The generic information extraction system". Dans les actes de *MUC5 '93 : Proceedings of the 5th conference on Message understanding*, pp 87–91. Association for Computational Linguistics, 1993.

[Holland, 1975] J. H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI, USA, 1975.

[Ibekwe-sanjuan & sanjuan, 2003] F. Ibekwe-sanjuan et E. sanjuan. "Cartographie

de réseaux de termes". Dans les actes de *5ème Journées, Terminologie et Intelligence Artificielle (TIA'03)*, 2003.

[Inmon, 1991] W. H. Inmon. *Building the Data Warehouse*. 1991.

[Jalam & Chauchat, 2002] R. Jalam et J.H. Chauchat. "Pourquoi les n-grammes permettent de classer des textes ? Recherche de mots-clefs pertinents à l'aide des n-grammes caractéristiques". Dans les actes de *JADT'02, Journées Internationales d'Analyse des Données Textuelles*, pp 381–390. A. Morin & P. Sébillot, 2002.

[Jarmasz & Szpakowicz, 2003] Mario Jarmasz et Stan Szpakowicz. "Roget's thesaurus and semantic similarity". Dans les actes de *Conference on Recent Advances in Natural Language Processing*, pp 212–219, 2003.

[Jolliffe, 1986] I. T. Jolliffe. *Principal component analysis*. Springer series in statistics. Springer, 1986.

[Jordan, 1986] M. I. Jordan. "An introduction to linear algebra in parallel distributed processing". pp 365–422, 1986, MIT Press.

[Joshi & Hopely, 1996] Aravind K. Joshi et Phil Hopely. "A parser from antiquity". *Nat. Lang. Eng.*, pp 291–294, 1996, Cambridge University Press.

[Joshi *et al.*, 1975] Aravind K. Joshi, Leon S. Levy, et Masako Takahashi. "Tree Adjunct Grammars". *J. Comput. Syst. Sci.*, pp 136–163, 1975.

[Junker & Hoch, 1997] M. Junker et R. Hoch. "Evaluating OCR and Non-OCR Text Representations for Learning Document Classifiers". Dans les actes de *ICDAR '97 : Proceedings of the 4th International Conference on Document Analysis and Recognition*, pp 1060–1066, Washington, DC, USA, 1997. IEEE Computer Society.

[Kanejiya *et al.*, 2003] D. Kanejiya, A. Kumar, et S. Prasad. "Automatic Evaluation of Students' Answers using Syntactically Enhanced LSA". Dans les actes de *Proceedings of the Human Language Technology Conference (HLT-NAACL 2003) Workshop on Building Educational Applications using NLP*, 2003.

[Kearns, 1988] M. Kearns. “Thoughts on hypothesis boosting, ML class project”. 1988.

[Kelly & Stone, 1975] Edward F. Kelly et Philip J. Stone. *Computer Recognition of English Word Senses*. 1975.

[Kessler *et al.*, 2007] Rémy Kessler, Juan Manuel Torres-Moreno, et Marc El-Bèze. “E-Gen : Automatic Job Offer Processing system for Human Ressources”. *MICAI 2007, Aguscalientes, Mexique, pp 985-995*, 2007.

[Kessler *et al.*, 2008a] Rémy Kessler, Nicolas Béchet, Mathieu Roche, Marc El-Bèze, et Juan Manuel Torres Moreno. “Automatic Profiling System for Ranking Candidates Answers in Human Resources”. Dans les actes de *OTM Workshops*, pp 625–634, 2008.

[Kessler *et al.*, 2008b] Rémy Kessler, Juan Manuel Torres-Moreno, et Marc El-Bèze. “ E-Gen : Profilage automatique de candidatures”. *TALN 2008, Avignon, France*, pp 370–379, 2008.

[Kessler *et al.*, 2009] Rémy Kessler, Nicolas Béchet, Juan Manuel Torres Moreno, Mathieu Roche, et Marc El-Bèze. “Job Offer Management : How Improve the Ranking of Candidates”. Dans les actes de *ISMIS'09 (International Symposium on Methodologies for Intelligent Systems*, 2009.

[Kießling, 2002] Werner Kiessling. “Foundations of preferences in database systems”. Dans les actes de *VLDB'02 : Proceedings of the 28th international conference on Very Large Data Bases*, pp 311–322. VLDB Endowment, 2002.

[Kim, 2008] M. Y. Kim. “Detection of gene interactions based on syntactic relations.”. *Journal of biomedicine & biotechnology*, 2008.

[Kittler, 1978] J. Kittler. “Pattern Recognition and Signal Processing, Chapter Feature Set Search Algorithms Sijthoff and Noordhoff”. *Alphen aan den Rijn, Netherlands*, pp 41–60, 1978.

[Knapen & Briot, 1999] E. Knapen et B. Briot. “Test de gestionnaire de thesaurus pour la terminologie”. Dans les actes de *La banque des mots*, pp 31–50, 1999.

[Kohomban & Lee, 2007] Upali Sathyajith KOHOMBAN et Wee Sun LEE. "Optimizing Classifier Performance in Word Sense Disambiguation by Redefining Sense Classes". Dans les actes de *IJCAI*, pp 1635–1640, 2007.

[Kongovi *et al.*, 2002] Madhusudhan KONGOVI, Juan Carlos GUZMAN, et Venu DASIGI. "Text Categorization An Experiment Using Phrases.". Dans les actes de *ECIR*, pp 213–228, 2002.

[Koutrika & Ioannidis, 2004] Georgia KOUTRIKA et Yannis IOANNIDIS. "Personalization of Queries in Database Systems". Dans les actes de *ICDE '04 : Proceedings of the 20th International Conference on Data Engineering*, page 597. IEEE Computer Society, 2004.

[Kumps *et al.*, 2004] Nicolas KUMPS, Pascal FRANCQ, et Alain DELCHAMBRE. "Création d'un espace conceptuel par analyse de données contextuelles". Dans les actes de *Proceedings of JADT 2004*, 3 2004.

[Labadié & Prince, 2008] A. LABADIÉ et Violaine PRINCE. "Lexical and Semantic Methods in Inner Text Topic Segmentation : A Comparison between C99 and Transeg". Dans les actes de *NLDB*, pp 347–349, 2008.

[Labadié, 2008] Alexandre LABADIÉ. "*Segmentation thématique de texte linéaire et non-supervisée : Détection active et passive des frontières thématiques en Français*". PhD thesis, Université Montpellier II - Sciences et Techniques du Languedoc, 12 2008.

[Lafourcade & Prince, 2001] Mathieu LAFOURCADE et Violaine PRINCE. "Relative Synonymy and Conceptual Vectors". Dans les actes de *in Proceedings of NLPRS2001, Tokyo*, pp 127–134, 2001.

[Lafourcade, 2006] Mathieu LAFOURCADE. "Conceptual vector learning - comparing bootstrapping from a thesaurus or induction by emergence". 2006.

[Lallich & Teytaud, 2004] Stéphane LALLICH et Olivier TEYTAUD. "Evaluation et validation de l'intérêt des règles d'association". *Revue des Nouvelles Technologies de l'Information*, pp 193–217, 2004.

[Landauer & Dumais, 1997] T. LANDAUER et S. DUMAIS. "A Solution to Plato's Problem : The Latent Semantic Analysis Theory of Acquisition,

Induction and Representation of Knowledge". *Psychological Review*, pp 211–240, 1997.

[Landauer *et al.*, 1997] T. Landauer, D. Laham, B. Rehder, et M. E. Schreiner. "How well can passage meaning be derived without using word order ? A comparison of Latent Semantic Analysis and humans". Dans les actes de *Proceedings of the 19th annual meeting of the Cognitive Science Society*, pp 412–417, 1997.

[Larkey *et al.*, 2000] L.S. Larkey, P. Ogilvie, M.A. Price, et B. Tamilio. "Acrophile : An automated Acronym Extractor and Server". Dans les actes de *Proceedings of the Fifth ACM International Conference on Digital Libraries*, pp 205–214, 2000.

[Laroum *et al.*, 2009] Sami Laroum, Nicolas Béchet, Hatem Hamza, et Mathieu Roche. "Classification automatique de documents bruités à faible contenu textuel". *RNTI : Revue des Nouvelles Technologies de l'Information*, page 25, 2009.

[Larousse, 1992] Thésaurus Larousse. *Thésaurus Larousse - des idées aux mots, des mots aux idées.* Ed.Larousse, Paris, 1992.

[Larson *et al.*, 1987] Richard Larson, Steven Paul Abney, et Steven Paul Abney. "*The English Noun Phrase in its Sentential Aspect*". PhD thesis, 1987.

[Laueufer, 1992] C. Laueufer. "Syllabification and resyllabification in French". Dans les actes de *Theoretical analyses in romance linguistics*, pp 18–36. J. Benjamins Pub. Co : Amsterdam, 1992.

[Le Ny, 1979] J. F. Le Ny. *La sémantique psychologique.* Presses Universitaires de France, Paris, 1979.

[Lehr & Pong, 2003] Robert G. Lehr et Annpey Pong. "ROC Curve". Dans les actes de *Encyclopedia of Biopharmaceutical Statistics, Taylor & Francis Group, Dekker, New York*, pp 884–891. Chow, S.C. (Ed.), 2003.

[Lei & Mirghafori, 2007] H. Lei et N. Mirghafori. "Word-conditioned phone N-grams for speaker recognition". *Proc. ICASSP, Honolulu*, 2007.

[Lewis, 1998] David D. Lewis. "Naive (Bayes) at forty : The independence assumption in information retrieval". pp 4–15. Springer Verlag, 1998.

[Lewis *et al.*, 2004] David D. Lewis, Yiming Yang, Tony G. Rose, et Fan Li. "RCV1 : A New Benchmark Collection for Text Categorization Research". *J. Mach. Learn. Res.*, pp 361–397, 2004, MIT Press.

[L'Homme, 1998] M. C. L'Homme. "Le statut du verbe en langue de spécialité et sa description lexicographique". Dans les actes de *Cahiers de Lexicologie 73*, pp 61–84, 1998.

[Li & Park, 2007] Cheng Hua Li et Soon Cheol Park. "An Efficient Document Categorization Model Based on LSA and BPNN". Dans les actes de *ALPIT '07 : Proceedings of the Sixth International Conference on Advanced Language Processing and Web Information Technology (ALPIT 2007)*, pp 9–14. IEEE Computer Society, 2007.

[Liao *et al.*, 2007] C. Liao, S. Alpha, et P. Dixon. "Feature preparation in text categorization". 2007.

[Lin, 1998] Dekang Lin. "Extracting collocations from text corpora". Dans les actes de *In First Workshop on Computational Terminology*, pp 57–63, 1998.

[Lopez, 1999] Patrice Lopez. "*Analyse d'énoncés oraux pour le Dialogue Homme-Machine à l'aide de Grammaires Lexicalisées d'Arbres*". Thèse de doctorat à l'université de nancy 1, 1999.

[Lovins, 1968] Julie B. Lovins. "Development of a Stemming Algorithm.". 1968.

[Macqueen, 1967] J. B. Macqueen. "Some methods of classification and analysis of multivariate observations". Dans les actes de *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp 281–297, 1967.

[Maisonnasse *et al.*, 2008] Loïc Maisonnasse, Catherine Berrut, et Jean-Pierre Chevallet. "L'expressivité des modèles de recherche d'informations précises". Dans les actes de *INFORSID*, 2008.

[Maniez, 1999] J. Maniez. “Des classifications aux thésaurus : Du bon usage des facettes”. Dans les actes de *Documentaliste*, volume 36, pp 249–260. Association française des documentalistes et des bibliothécaires spécialisés, Paris, FRANCE, 1999.

[Mansur *et al.*, 2006] M. Mansur, N. UzZaman, et M. Khan. “Analysis of N-gram based text categorization for Bangla in a newspaper corpus”. Dans les actes de *Proc. of 9th International Conference on Computer and Information Technology (ICCIT 2006)*, Dhaka, Bangladesh, 2006.

[Mayaffre, 2005] D. Mayaffre. “De la lexicométrie à la logométrie”. 2005.

[Mcculloch & Pitts, 1943] Warren Mcculloch et Walter Pitts. “A logical calculus of the ideas immanent in nervous activity”. *Bulletin of Mathematical Biology*, pp 115–133, 1943.

[McHale, 1998] Michael L. McHale. “A Comparison of WordNet and Roget’s Taxonomy for Measuring Semantic Similarity”. Dans les actes de *Proc COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems. Montreal, Canada*, pp 115–120, 1998.

[Mclachlan & Krishnan, 2007] Geoffrey J. Mclachlan et Thriyambakam Krishnan. *The EM Algorithm and Extensions (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 2007.

[Mel’čuk, 1998] Igor Mel’čuk. Collocations and Lexical Functions. Dans les actes de *Phraseology : Theory, Analysis, and Applications*, pp 23–54. Oxford : Clarendon Press, 1998.

[Miller, 1985] George A. Miller. “WordNet : a dictionary browser”. Dans les actes de *the First International Conference on Information in Data*, 1985.

[Miller *et al.*, 2000] Ethan Miller, Dan Shen, Junli Liu, et Charles Nicholas. “Performance and Scalability of a Large-scale N-gram Based Information Retrieval System”. *Journal of Digital Information*, 2000.

[Mitchell, 1997] T. Mitchell. *Machine Learning*. McGraw-Hill Education (ISE Editions), 1997.

[Morris & Hirst, 1991] Jane Morris et Graeme Hirst. “Lexical cohesion computed by thesaural relations as an indicator of the struc-

ture of text". *Comput. Linguist.*, pp 21–48, 1991, MIT Press.

[Moulinier, 1996] Isabelle MOULINIER. "*Une approche de la categorisation de textes par l'apprentissage symbolique*". Phd thesis, University of Paris VI, LIP6, 8 rue du capitaine Scott, 75015 PARIS, 1996.

[Nazarenko *et al.*, 2001] A. NAZARENKO, P. ZWEIGENBAUM, B. HABERT, et J. BOUAUD. "Corpus-based Extension of a Terminological Semantic Lexicon". Dans les actes de *Recent Advances in Computational Terminology*, pp 327–351, 2001.

[Paliouras *et al.*, 1999] Georgios PALIOURAS, Vangelis KARKALETSIS, Christos PAPATHEODOROU, et Constantine D. SPYROPOULOS. "Exploiting Learning Techniques for the Acquisition of User Stereotypes and Communities". Dans les actes de *UM99 User Modeling : Proceedings of the Seventh International Conference*, pp 169–178. Springer-Verlag, 1999.

[Pallier, 1994] C. PALLIER. "*Rôle de la syllabe dans la perception de la parole : Etudes attentionelles*". PhD thesis, Ecole des Hautes Etudes en Sciences Sociales, Paris, 1994.

[Pang & Lee, 2002] Bo PANG et Lillian LEE. "Thumbs up ? Sentiment Classification using Machine Learning Techniques". Dans les actes de *In Proceedings of EMNLP*, pp 79–86, 2002.

[Paolo *et al.*, 2004] Wordnet Senses PAOLO, Paolo ROSSO, Edgardo FERRETTI, Daniel JIMÉNEZ, et Vicente VIDAL. "Text Categorization and Information Retrieval Using". Dans les actes de *In Proceedings of the 2nd Global Wordnet Conference (GWC'04*, pp 299–304. Springer-Verlag, 2004.

[Paradis & Nie, 2005] F. PARADIS et J.Y. NIE. "Filtering Contents with Bigrams and Named Entities to Improve Text Classification". Dans les actes de *AIRS*, pp 135–146, 2005.

[Paroubek *et al.*, 2005] P. PAROUBEK, I. ROBBA, A. VILNAT, et L. G. POUILLOT. "EASy : Campagne d'évaluation des analyseurs syntaxiques". *Ateliers de la 12e Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2005)*, 2005.

[Pelleg & Moore, 2000] Dan PELLEG et Andrew W. MOORE. "X-means : Extending K-means with Efficient Estimation of the Number of

Clusters". Dans les actes de *ICML '00 : Proceedings of the Seventeenth International Conference on Machine Learning*, pp 727–734, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[Peng *et al.*, 2003] Fuchun Peng, Dale Schuurmans, et Shaojun Wang. "Augmenting Naive Bayes Classifiers with Statistical Language Models". 2003.

[Picard & Estraillier, 2008] Francois Picard et Pascal Estraillier. "Motion capture system contextualization". Dans les actes de *ACE '08 : Proceedings of the 2008 International Conference on Advances in Computer Entertainment Technology*, pp 147–150, New York, NY, USA, 2008. ACM.

[Pisetta *et al.*, 2006] V. Pisetta, H. Hacid, F. Bellal, et G. Ritschard. "Traitement automatique de textes juridiques". Dans les actes de *Semaine de la Connaissance (SdC 06), Nantes*. Rémi Lehn and Mounira Harzallah and Nathalie Aussenac-Gilles and Jean Charlet, Juin 2006.

[Platt, 1999] J. Platt. "Fast training of support vector machines using sequential minimal optimization". Dans les actes de *B. Scholkopf, C. J. C. Burges, and A. J. Smola, editors, Advances in Kernel Methods – Support Vector Learning, pages 185-208. MIT Press, Cambridge, MA*, 1999.

[Porter, 1980] M. F. Porter. "An algorithm for suffix stripping". *Program*, pp 130–137, 1980, Morgan Kaufmann Publishers Inc.

[Pottier, 1964] B. Pottier. "Vers une sémantique moderne". Dans les actes de *Travaux de sémantique et de littérature*, pp 107–137, 1964.

[Poudat *et al.*, 2006] C. Poudat, G. Cleuziou, et V. Clavier. "Catégorisation de textes en domaines et genres. Complémentarité des indexations lexicale et morphosyntaxique". *Document Numérique*, pp 61–76, 2006, Lavoisier-Hermes.

[Pratt, 1939] Fletcher Pratt. *Secret and urgent : the story of codes and ciphers*. R. Hale, London, 1939.

[Quinlan, 1986] J. R. Quinlan. "Induction of decision trees". *Machine Learning*, pp 81–106, 1986.

[Quinlan, 1993] J. Ross Quinlan. *C4.5 : Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning)*. Morgan Kaufmann, 1993.

[Rafter *et al.*, 2000] Rachael Rafter, Barry Smyth, et Keith Bradley. "Inferring Relevance Feedback from Server Logs : A Case Study in Online Recruitment". 2000.

[Rastier, 1996] F. Rastier. "Le problème épistémologique du contexte et le statut de l'interprétation dans les sciences du langage". Dans les actes de *Troisième congrès européen de systémique, Rome*, pp 397–402, 1996.

[Rehder *et al.*, 1998] B. Rehder, M. Schreiner, M. Wolfe, D. Laham, T. Landauer, et W. Kintsch. "Using Latent Semantic Analysis to assess knowledge : Some technical considerations". Dans les actes de *Discourse Processes*, volume 25, pp 337–354, 1998.

[Riegel *et al.*, 1999] M. Riegel, JC Pellat, et R. Rioul. *Grammaire méthodique du français*. PUF, 1999.

[Riloff, 1995] Ellen Riloff. "Little Words Can Make a Big Difference for Text Classification". Dans les actes de *In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp 130–136. ACM Press, 1995.

[Robertson & Jones, 1976] S. E. Robertson et Sparck K. Jones. "Relevance weighting of search terms". *Journal of the American Society for Information Science*, pp 129–146, 1976.

[Robertson *et al.*, 1996] S.E. Robertson, S. Walker, M.M. Beaulieu, M. Gatford, et A. Payne. "Okapi at TREC-4". 1996.

[Roche & Chauché, 2006] Mathieu Roche et Jacques Chauché. "LSA : les limites d'une approche statistique". Dans les actes de *Actes de l'atelier FDC'06 (Fouille de Données Complexes), conférence EGC'2006*, pp 95–106, 2006.

[Roche & Prince, 2008] Mathieu Roche et Violaine Prince. "Managing the Acronym/Expansion Identification Process for Text-Mining Applications". *International Journal of Software and Informatics, Special issue on Data Mining*, pp 163–179, 2008.

[Roche, 2005] C. Roche. “Terminologie et ontologie”. Dans les actes de *Revue Langages v. 157, Éditions Larousse*, pp 48–62, 2005.

[Roche *et al.*, 2004] M. Roche, T. Heitz, O. Matte-Tailliez, et Y. Kodratoff. “Exit : Un système itératif pour l’extraction de la terminologie du domaine à partir de corpus spécialisés”. Dans les actes de *Proceedings of JADT’04*, volume 2, pp 946–956, 2004.

[Rogati & Yang, 2002] M. Rogati et Y. Yang. “High performing and scalable feature selection for text classification”. Dans les actes de *the Eleventh International Conference on Information and Knowledge Management*, pp 659–661, 2002.

[Roget, 1852] P. Roget. *Thesaurus of English Words and Phrases*. Longman, London, 1852.

[Rosenblatt, 1958] Frank Rosenblatt. “The perceptron : a probabilistic model for information storage and organization in the brain”. *Psychological Review*, pp 386–408, 1958.

[Rosenblatt, 1988] Frank Rosenblatt. “The perceptron : a probabilistic model for information storage and organization in the brain”. *Neurocomputing : foundations of research*, pp 89–114, 1988, MIT Press.

[Roussanaly *et al.*, 2005] Azim Roussanaly, Benoît Crabbé, et Jérôme Perrin. “Premier bilan de la participation du LORIA à la campagne d’évaluation EASY”. Dans les actes de *12e Conférence annuelle sur le Traitement Automatique des Langues Naturelles - TALN 2005*, Dourdan, France, 2005. ATALA.

[Saeys *et al.*, 2007] Yvan Saeys, Inaki Inza, et Pedro Larranaga. “A review of feature selection techniques in bioinformatics”. *Bioinformatics*, pp 2507–2517, 2007.

[Sager, 1981] Naomi Sager. *Natural Language Information Processing : A Computer Grammmar of English and Its Applications*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1981.

[Sager, 1990] Juan C. Sager. *A Practical Course in Terminology Processing*. John Benjamins Publishing Co, 1990.

[Salton & Lesk, 1965] Gerard Salton et M. E. Lesk. "The SMART automatic document retrieval systems an illustration". *Communications of the ACM*, pp 391–398, 1965.

[Salton & Yang, 1973] Gerard Salton et C.S. Yang. "On the Specification of Term Values in Automatic Indexing". 1973, Cornell University.

[Salton, 1971] Gerard Salton. *The SMART Retrieval System—Experiments in Automatic Document Processing.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.

[Salton *et al.*, 1975a] G. Salton, C. S. Yang, et C. T. Yu. "A theory of term importance in automatic text analysis.". *Journal of the American Society for Information Science*, pp 33–44, 1975.

[Salton *et al.*, 1975b] Gerard Salton, A. Wong, et C. S. Yang. "A Vector Space Model for Automatic Indexing". *Commun. ACM*, pp 613–620, 1975.

[Salton *et al.*, 1983] Gerard Salton, Edward A. Fox, et Harry Wu. "Extended Boolean information retrieval". *Commun. ACM*, pp 1022–1036, November 1983, ACM.

[Saneifar *et al.*, 2009] Hassan Saneifar, Stéphane Bonniol, Anne Laurent, Pascal Poncelet, et Mathieu Roche. "Terminology Extraction from Log Files". Dans les actes de *DEXA*, pp 769–776, 2009.

[Schegloff, 1992] Emanuel A. Schegloff. "In Another Context". Dans les actes de *Rethinking Context. Language as an Interactive Phenomenon*, pp 191–227, Cambridge, 1992. Cambridge University Press.

[Schmid, 1995] H. Schmid. "Improvements in part-of-speech tagging with an application to German". Dans les actes de *Proceedings of the ACL SIGDAT-Workshop, Dublin*, 1995.

[Schreiner *et al.*, 1998] M.E. Schreiner, Bob Rehder, Darrell Laham, Peter W. Foltz, Walter Kintsch, Thomas K L, Thomas K. Landauer, Michael B. W. Wolfe, et Michael B. W. Wolfe. "Learning from text : Matching readers and texts by Latent Semantic Analysis". 1998.

[Schröder *et al.*, 2003] Marc Schröder, Marc Schröder, Marc Schröder, Jürgen Trouvain, et Jürgen Trouvain. "The German Text-to-Speech Synthesis System MARY : A Tool for Research, Development and Teaching". Dans les actes de *International Journal of Speech Technology*, pp 365–377, 2003.

[Schwab, 2005] Didier Schwab. "*Approche hybride lexicale et thématique pour la modélisation, la détection et l'exploitation des fonctions lexicales en vue de l'analyse sémantique de texte*". PhD thesis, 2005.

[Sebag & Schoenauer, 1988] M. Sebag et M. Schoenauer. "Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases". *European Knowledge Acquisistion Workshop, EKAW'88*, 1988, Boose, M. L. J., Gaines,. B. (eds.).

[Shannon, 1948] C. E. Shannon. "A mathematical theory of communication". *Bell system technical journal*, 1948.

[Shen *et al.*, 2005] Dan Shen, Geert jan M. Kruijff, et Dietrich Klakow. "Exploring Syntactic Relation Patterns for Question Answering". Dans les actes de *In Proc. Of IJCNL'05*, 2005.

[Shen *et al.*, 2006] Wade Shen, William Campbell, Terry Gleason, Doug Reynolds, et Elliot Singer. "Experiments with Lattice-based PPRLM Language Identification". Dans les actes de *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006*, pp 1–6, 2006.

[Sheynin, 1977] O. B. Sheynin. "Early History of the Theory of Probability". *Archive for History of Exact Sciences*, pp 201–259, 1977.

[Silberztein, 1993] Max Silberztein. *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*. Masson : Paris, 1993.

[Silverstein *et al.*, 1998] Craig Silverstein, Sergey Brin, et Rajeev Motwani. "Beyond Market Baskets : Generalizing Association Rules to Dependence Rules". *Data Min. Knowl. Discov.*, pp 39–68, 1998, Kluwer Academic Publishers.

[Sjöblom, 2002] M.K. Sjöblom. "Le choix de la lemmatisation. Différentes méthodes appliquées à un même corpus". Dans les

actes de *JADT : 6es Journées internationales d'Analyse statistique des Données Textuelles*, 2002.

[Slonim *et al.*, 2002] N. Slonim, N. Friedman, et N. Tishby. "Unsupervised document classification using sequential information maximization". Dans les actes de *SIGIR*, pp 129–136, 2002.

[Smadja *et al.*, 1996] F. Smadja, K. R. McKeown, et V. Hatzivassiloglou. "Translating collocations for bilingual lexicons : A statistical approach". *Computational Linguistics*, pp 1–38, 1996.

[Small, 1980] Steven Lawrence Small. "*Word expert parsing : a theory of distributed word-based natural language understanding*". PhD thesis, College Park, MD, USA, 1980.

[Sokal, 1977] R. Sokal. "Clustering and classification : background and current directions". pp 1–15. J.V. Ryzin (Ed.), Classification and Clustering, Academic Press, New York, 1977.

[Solso, 1979] Robert L. Solso. "Bigram and trigram frequencies and versatilities in the English language". Dans les actes de *Behavior Research Methods & Instrumentation*, volume 11(5), pp 475–484, 1979.

[Song *et al.*, 2008] Young-In Song, Kyoung-Soo Han, Sang-Bum Kim, So-Young Park, et Hae-Chang Rim. "A novel retrieval approach reflecting variability of syntactic phrase representation". *J. Intell. Inf. Syst.*, pp 265–286, 2008, Kluwer Academic Publishers.

[Spiteri, 2005] Louise F. Spiteri. "Word association testing and thesaurus construction : A pilot study". Dans les actes de *Cataloging & classification quarterly*, pp 55–78. Haworth Press, Binghamton, NY, ETATS-UNIS, 2005.

[Spärck-Jones, 1970] Karen Spärck-Jones. "Some thoughts on classification for retrieval". *Journal of Documentation*, pp 89–101, 1970.

[Spärck-Jones, 1999] Karen Spärck-Jones. "Information Retrieval and Artificial Intelligence". *Artif. Intell.*, pp 257–281, 1999.

[Sutton, 1988] Richard S. Sutton. "Learning to Predict by the Methods of Temporal Differences". *Machine Learning*, pp 9–44, 1988.

[Tan *et al.*, 2002] C.M. Tan, Y.F. Wang, et C.D. Lee. "The use of bigrams

to enhance text categorization". *Inf. Process. Manage.*, pp 529–546, 2002, Pergamon Press, Inc.

[Thanopoulos *et al.*, 2002] A. Thanopoulos, N. Fakotakis, et G. Kokkianakis. "Comparative Evaluation of Collocation Extraction Metrics". Dans les actes de *Proceedings of LREC'02*, volume 2, pp 620–625, 2002.

[Turney, 2001] P.D. Turney. "Mining the Web for Synonyms : PMI–IR versus LSA on TOEFL". Dans les actes de *Proceedings of ECML'01, Lecture Notes in Computer Science*, pp 491–502, 2001.

[Van Dijk, 2008] Teun A. Van Dijk. *Discourse and context : a sociocognitive approach.* Cambridge Univ. Press, 2008.

[Vardhan *et al.*, 2007] B. Vishnu Vardhan, L. Pratap Reddy, et A. Vinay-Babu. "Text categorization using trigram technique for Telugu script". *Journal of Theoretical and Applied Information Technology*, pp 9–14, 2007.

[Vergne, 1990] J. Vergne. "*Etude et modélisation de la syntaxe des langues à l'aide de l'ordinateur*". Dossier d'habilitation à diriger des recherches, 1990.

[Verma *et al.*, 2007] R. Verma, P. Chen, et W. Lu. "A Semantic Free-text Summarization System Using Ontology Knowledge". Dans les actes de *Proceedings of the Document Understanding Conference 2007*, 2007.

[Vernier *et al.*, 2009] Matthieu Vernier, Laura Monceaux, Béatrice Daille, et Estelle Dubreil. "Catégorisation des évaluations dans un corpus de blogs multi-domaine". *Revue des Nouvelles Technologies de l'Information*, 2009.

[Wan & Tong, 2008] Yuan Wan et Hengqing Tong. "Categorization and Monitoring of Internet Public Opinion Based on Latent Semantic Analysis". *Business and Information Management, International Seminar on*, pp 121–124, 2008, IEEE Computer Society.

[Wang & Domeniconi, 2008] Pu Wang et Carlotta Domeniconi. "Building semantic kernels for text classification using wikipedia". Dans les actes de *KDD '08 : Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2008.

[Wang & Vergyri, 2006] Wen WANG et Dimitra VERGYRI. "The use of word n-grams and parts of speech for hierarchical cluster language modeling". Dans les actes de *ICASSP 2006*, pp 1057–1060, 2006.

[Watkins, 1989] C. WATKINS. "*Learning from Delayed Rewards*". PhD thesis, University of Cambidge, England, 1989.

[Wermter & Hahn, 2004] Joachim WERMTER et Udo HAHN. "Collocation extraction based on modifiability statistics". Dans les actes de *COLING '04*, page 980, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[Widrow & Hoff, 1960] Bernard WIDROW et Marcian E. HOFF. "Adaptive Switching Circuits". *1960 IRE WESCON Convention Record*, pp 96–104, 1960, IRE.

[Widrow & Hoff, 1988] Bernard WIDROW et Marcian E. HOFF. "Adaptive switching circuits". pp 123–134, 1988, MIT Press.

[Widrow, 1985] B. WIDROW. *Adaptative signal processing*. Pentice Hall, 1985.

[Wiemer-Hastings & Zipitria, 2001] Peter WIEMER-HASTINGS et Iraide ZIPITRIA. "Rules for syntax, vectors for semantics". Dans les actes de *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society*, 2001.

[Wiemer-Hastings *et al.*, 1999] Peter WIEMER-HASTINGS, Katja WIEMER-HASTINGS, et Arthur C. GRAESSER. "Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis". Dans les actes de *Artificial Intelligence in Education*, pp 535–542. IOS Press, 1999.

[Wilbur & Sirotkin, 1992] W. John WILBUR et Karl SIROTKIN. "The automatic identification of stop words". *J. Inf. Sci.*, pp 45–55, 1992, Sage Publications, Inc.

[Wilks, 1998] Yorick WILKS. "Language Processing and the Thesaurus". Dans les actes de *National Language Research Institute*, 1998.

[Witten *et al.*, 1999] I. WITTEN, E. FRANK, L. TRIGG, M. HALL, G. HOLMES, et S. CUNNINGHAM. "Weka : Practical machine learning tools and techniques with java implementations". pp 192–196. ICONIP/ANZIIS/ANNES'99 Int. Workshop :

Emerging Knowledge Engineering and Connectionist-Based Info. Systems, 1999.

[Xide Lin *et al.*, 2008] Cindy Xide Lin, Bolin Ding, Jiawei Han, Feida Zhu, et Bo Zhao. "Text Cube : Computing IR Measures for Multidimensional Text Database Analysis". Dans les actes de *Proc. 2008 Int. Conf. on Data Mining (ICDM'08)*, December 2008.

[Yan *et al.*, 2003] L. Yan, R.H. Dodier, M. Mozer, et R.H. Wolniewicz. "Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic". Dans les actes de *Proceedings of ICML'03*, pp 848–855, 2003.

[Yang & Liu, 1999a] Y. Yang et X. Liu. "A Re-Examination of Text Categorization Methods". Dans les actes de *SIGIR*, pp 42–49, 1999.

[Yang & Liu, 1999b] Yiming Yang et Xin Liu. "A re-examination of text categorization methods". Dans les actes de *SIGIR '99 : Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp 42–49. ACM Press, 1999.

[Yang & Pedersen, 1997] Yiming Yang et Jan O. Pedersen. "A Comparative Study on Feature Selection in Text Categorization". Dans les actes de *ICML '97 : Proceedings of the Fourteenth International Conference on Machine Learning*, pp 412–420. Morgan Kaufmann Publishers Inc., 1997.

[Yarowsky, 1992] David Yarowsky. "Word-Sense Disambiguation using Statistical Models of Roget's Categories Trained on Large Corpora". Dans les actes de *Proc of COLING-92*, pp 454–460, July 1992.

[Zampa & Lemaire, 2002] Virginie Zampa et Benoît Lemaire. "Latent Semantic Analysis for User Modeling". *J. Intell. Inf. Syst.*, pp 15–30, 2002.

[Zelikovitz & Hirsh, 2000] S. Zelikovitz et Haym Hirsh. "Improving Short Text Classification Using Unlabeled Background Knowledge". Dans les actes de Pat Langley, , *Proceedings of ICML-00, 17th International Conference on Machine Learning*, pp 1183–1190, Stanford, US, 2000. Morgan Kaufmann Publishers, San Francisco, US.

[Zelikovitz, 2004] S. Zelikovitz. "Transductive LSI for Short Text Classification Problems". Dans les actes de *Proceedings of the 17th International FLAIRS Conference*, 2004.

[Zhang *et al.*, 2005] Kai Zhang, Jian Sun, et Bin Wang. "A WordNet-based approach to feature selection in text categorization". pp 475–484, 2005, Springer-Verlag.

[Zhu, 2007] Xiaojin Zhu. "Semi-Supervised Learning Literature Survey". 2007.

[Zighed & Loudcher, 2004 à 2007] Djamel Abdelkader Zighed et Sabine Loudcher. "Rapport d'activité du laboratoire ERIC". 2004 à 2007.

[Zipf, 1941] G. K. Zipf. "National unity and disunity. The nation as a bio-social organism". *Princeton Press, Bloomington*, 1941.

Résumé :
Les mots constituent l'un des fondements des langues naturelles de type indo-européenne. Des corpus rédigés avec ces langues sont alors naturellement décrits avec des mots. Cependant, l'information qu'ils véhiculent seuls est assez réduite d'un point de vue sémantique. Il est en effet primordial de prendre en compte la complexité de ces langues comme par exemple leurs propriétés syntaxiques, lexicales et sémantiques. Nous proposons dans cette thèse de prendre en considération ces propriétés en décrivant un corpus par le biais d'informations syntaxiques permettant de découvrir des connaissances sémantiques.

Nous présentons dans un premier temps un modèle de sélection de descripteurs SelDe. Ce dernier se fonde sur les objets issus des relations syntaxiques d'un corpus. Le modèle SelDe a été évalué pour des tâches de classification de données textuelles. Pour cela, nous présentons une approche d'expansion de corpus, nommée *ExpLSA*, dont l'objectif est de combiner les informations syntaxiques fournies par SelDe et la méthode numérique *LSA*.

Le modèle SelDe, bien que fournissant des descripteurs de bonne qualité, ne peut être appliqué avec tous types de données textuelles. Ainsi, nous décrivons dans cette thèse un ensemble d'approches adaptées aux données textuelles dites *complexes*. Nous étudions la qualité de ces méthodes avec des données syntaxiquement mal formulées et orthographiées, des données bruitées ou incomplètes et finalement des données dépourvues de syntaxe.

Finalement un autre modèle de sélection de descripteurs, nommé SelDeF, est proposé. Ce dernier permet de *valider* de manière automatique des relations syntaxiques dites "induites". Notre approche consiste à combiner deux méthodes. Une première approche fondée sur des vecteurs *sémantiques* utilise les ressources d'un thésaurus. Une seconde s'appuie sur les connaissances du Web et des mesures statistiques afin de valider les relations syntaxiques. Nous avons expérimenté SelDeF pour une tâche de construction et d'enrichissement de classes conceptuelles. Les résultats expérimentaux montrent la qualité des approches de validation et reflètent ainsi la qualité des classes conceptuelles construites.

Abstract :
Words are one of the grounds of European languages. Corpora written with these languages are normally describe by words. However, extracted information given by words is semantically poor. Actually, to take into account the complexity of European languages are really important. As a result, we propose in this thesis to feature the characteristic of European languages by using syntactic informations in order to discover new semantic knowledge from corpora.

First, we present SELDE, a model of feature selection. This one is based on objects extracted from syntactic relations of a corpus. We experiment SELDE on textual classification tasks by proposing ExpLSA, an approach used to make a corpus expansion by using the SELDE features. The goal of ExpLSA is to combine the SELDE features with the statistic method LSA.

The SELDE model gives relevant features but cannot be apply with all kinds of textual data. Thus, we propose different approaches adapted to specific textual data, called complex textual data. We experiment our approaches with noised data, bad written data, and data without syntactic informations.

Finally, we propose the SELDEF model. It introduce the automatic validation of syntactic relations called induced. Two validation approaches are proposed : a Semantic-Vector-based approach and a Web Validation system. The Semantic Vectors approach is a Roget-based method which computes a syntactic relation as a vector. Web Validation uses a search engine to determine the relevance of a syntactic relation. Then, we propose approaches to combine both in order to rank induced syntactic relations. We experiment SELDEF in a conceptual classes building task. Obtained results confirm the quality of validation approaches and quality of built classes.

Discipline : Informatique

Laboratoire : Laboratoire d'Informatique, de Robotique et de Micro-électronique de Montpellier (LIRMM) ; UMR 5506 ; 161 rue Ada, 34392 Montpellier Cedex 5, France

Mots clés : TAL, fouille de textes, descripteur, syntaxe, classification.