# Fouille de Textes : de l'extraction des descripteurs linguistiques à leur induction

Mathieu Roche


**UNIVERSITÉ MONTPELLIER 2**


École Doctorale I2S

Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier

# HABILITATION À DIRIGER LES RECHERCHES

présentée par :

**Mathieu ROCHE**

soutenue le : 09 décembre 2011

Discipline/ Spécialité : Informatique

## Fouille de Textes : De l'extraction des descripteurs linguistiques à leur induction

| | |
|---|---|
| **Rapporteurs :** | |
| GAUSSIER Éric | Professeur, Université de Grenoble |
| LAPALME Guy | Professeur, Université de Montreal (Québec, Canada) |
| MATWIN Stan | Professeur, Université d'Ottawa (Ontario, Canada) |
| **Examinateurs :** | |
| AUSSENAC-GILLES Nathalie | Directrice de Recherche CNRS |
| GALLINARI Patrick | Professeur, Université Paris 6 |
| KODRATOFF Yves | Directeur de Recherche CNRS (en retraite) |
| PRINCE Violaine | Professeur, Université Montpellier 2 |
| TEISSEIRE Maguelonne | Directrice de Recherche Cemagref, UMR TETIS |

# Résumé

Les masses de données textuelles aujourd'hui disponibles engendrent un problème difficile lié à leur traitement automatique. Dans ce cadre, des méthodes de Fouille de Textes (FT) et de Traitement Automatique du Langage (TAL) peuvent, en partie, répondre à une telle problématique. Elles consistent à modéliser puis mettre en ?uvre des méthodologies appliquées aux données textuelles afin d'en déterminer le sens et/ou découvrir des connaissances nouvelles. Dans ce processus, le descripteur linguistique constitue un élément pivot.

Après une présentation des méthodes de traitement des descripteurs en eux-mêmes, ces derniers seront étudiés en contexte, c'est-à-dire en corpus. L'identification des descripteurs est souvent difficile à partir de corpus bruités et à faible contenu textuel sur lesquels nous concentrons nos efforts (par exemple, corpus issus du Web 2.0 ou du traitement OCR). Outre les mots considérés comme des descripteurs linguistiques pertinents en FT, nous nous sommes également intéressés à l'étude des syntagmes complexes à partir de corpus classiques puis d'une terminologie classique à partir de corpus complexes (par exemple, données logs ou corpus en français médiéval).

Dans la suite, les syntagmes étudiés ne se situent plus à proprement parler dans les textes mais ils seront induits à partir des mots issus des corpus. Les méthodes proposées permettent de mettre en relief des syntagmes originaux tout à fait utiles pour l'identification d'Entités Nommées, le titrage automatique ou la construction de classes conceptuelles. Contrairement au raisonnement déductif, le raisonnement inductif est dit hypothétique. Dans ce cadre, l'utilisation de méthodes de validation automatique des relations induites par le biais d'approches de Fouille du Web se révèle déterminant.

Les perspectives à ce travail se concentreront sur l'extraction de nouveaux descripteurs. Ces derniers seront associés à de nouvelles représentations sous forme d'entrepôts de données textuelles. Enfin, les travaux que nous souhaitons développer se focaliseront sur l ?analyse des textes dans un contexte plus vaste lié au multimédia que le paradigme du Web 2.0 a mis en exergue ces dernières années.

**Mots clés :** Fouille de Textes, Traitement Automatique du Langage, Descripteurs linguistiques, Terminologie, Entités Nommées, Titrage automatique, Construction de classes conceptuelles, Corpus, Web 2.0, Entrepôts de données textuelles

# Table des matières

# Introduction

Les masses de données textuelles aujourd'hui disponibles engendrent un problème difficile lié à leur traitement automatique. Des méthodes de Fouille de Textes (FT) et de Traitement Automatique du Langage (TAL) peuvent en partie répondre à une telle problématique. Elles consistent à modéliser puis mettre en œuvre des méthodologies appliquées aux données textuelles afin d'en déterminer le sens et/ou découvrir des connaissances nouvelles. La plupart des méthodologies proposées s'appuient sur des approches linguistiques et/ou statistiques.

Les processus de Fouille de Textes sont souvent composés de deux phases successives (cf. figure 1).

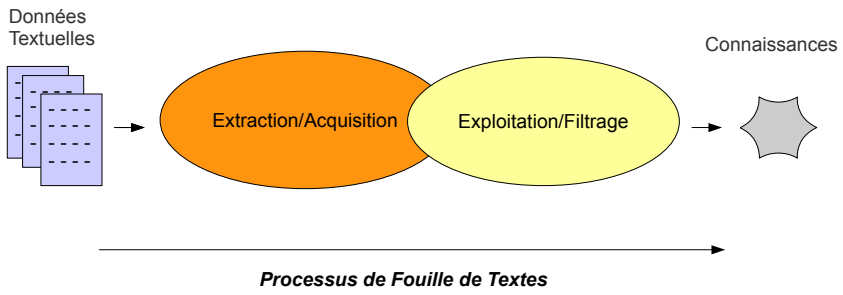


FIGURE 1 – Processus de Fouille de Textes et ses traitements

La première représente l'*extraction* des descripteurs tels que les mots-clés les plus significatifs. Ces derniers doivent être extraits, éventuellement traités afin de synthétiser au mieux le contenu des textes. Par exemple, si les mots *élection*, *2012* et *Elysée* sont mis en relief dans un article journalistique, il est aisé pour un humain de déterminer que ce texte traite de la prochaine élection présidentielle en France. La question que nous pouvons nous poser est : *De quelle manière extraire de tels descripteurs ?* Une fois cette problématique traitée automatiquement, semi-automatiquement, voire manuellement, nous pouvons nous interroger sur la manière d'exploiter ces mots-clés afin de les associer à une thématique pertinente. Ceci constitue la seconde phase d'un processus de Fouille de Textes. Dans ce cadre, les algorithmes utilisés consistent à *exploiter* les descripteurs et les données numériques associées (par exemple, la fréquence des mots dans

les textes). Les algorithmes peuvent alors associer la thématique "Élection présidentielle en France" à l'article journalistique évoqué précédemment. Outre les informations intrinsèques aux textes pour cette phase d'exploitation/filtrage, il peut être utile d'utiliser des connaissances externes (dictionnaires, taxonomies, etc).

### Synthèse des travaux menés

Les travaux que j'ai pu mener ces dernières années s'appuient sur ce processus dont une vision synthétique est proposée dans la figure 1. Ils se déclinent dans le cadre de collaborations académiques, industrielles ainsi que d'encadrement de stages de Master Recherche et de Thèses (cf. tableau 1). Dans ce tableau, les différents encadrements liés à des projets de recherche clairement identifiés sont représentés de couleur plus foncée [1]. Une partie de ce tableau sera rappelée en en-tête de chaque chapitre avec la liste des travaux concernés.

| *Thèmes de Recherche* | *Types de travaux* | *Années* | *Chapitres* |
|---|---|---|---|
| Extraction (80%) | | Exploitation (20%) | |
| Extraction de connaissances en Français Médiéval | **Projet STICS/UM2 – *TSAL***<br>porteur, co-responsable scientifique<br>*Stage Ingénieur CNAM* | 2006-2007 | Chap 2 |
| Analyse de sentiments | **Collaboration académique** – LIRMM/LGI2P (Nîmes)<br>*Stage M2 Recherche* | 2007-2011 | Chap 1 |
| Analyse de données SMS | **Projet CS/MSH-M/UM3 – *SMS4SCIENCES***<br>membre | 2011-2012 | Chap 4 |
| Extraction (70%) | | Exploitation (30%) | |
| Classification de données OCR | **Collaboration industrielle** – ITESOFT (Aimargues)<br>responsable scientifique<br>*Stage M2 Recherche* | 2007-2008 | Chap 4 |
| Classification de Blogs | **Collaboration industrielle** – PAPERBLOG (Paris)<br>responsable scientifique<br>*Stage Ingénieur* | 2007-2008 | Chap 2 |
| Classification en Ressources Humaines | **Collaboration académique** – LIRMM/LIA (Avignon) | 2008-2010 | X |
| Classification de Tweets | **Collaboration industrielle** – WEBREPORT (Bordeaux)<br>co-responsable scientifique<br>*Stage M2 Recherche* | 2010-2011 | Chap 2 |
| Recherche de gloses | **Projet CNRS PEPS – *RESENS***<br>porteur, co-responsable scientifique | 2010-2011 | Chap 2 |
| Recherche d'info. et cartographie de chercheurs | **Collaboration industrielle** – EXPERNOVA (Montpellier)<br>co-responsable scientifique | 2009-2011 | X |
| Extraction (50%) | | Exploitation (50%) | |
| Acquisition de classes conceptuelles | **THÈSE** (Région/CNRS) – N. Béchet | 2006-2009 | Chap 3 |
| Extraction d'information dans les fichiers logs | **THÈSE** CIFRE – H. Saneifar (en collaboration avec *Satin Technologies*) | 2008-2011 | Chap 2 |
| Cube de textes | **Collaboration académique** – LIRMM/CEMAGREF<br>*Stage Ingénieur CNAM* | 2009-2011 | Chap 4 |
| Visualisation de documents textuels | **Collaboration académique** – LIRMM/CEMAGREF/LABRI (Bordeaux) | 2009-2011 | X |
| Titrage automatique | **Collaboration industrielle** – Open-S/EvalAccess (Montpellier)<br>co-responsable scientifique<br>*Stages M2 Recherche*<br>**THÈSE** (Région/UM2) – C. Lopez | 2008-2009<br><br><br>2009-2012 | Chap 2 et 3 |
| Extraction (20%) | | Exploitation (80%) | |
| Mise en correspondance de schémas | **Projet ANR – *Forum***<br>membre<br>*Stage M2 Recherche* | 2005-2008 | Chap 1 |
| Extraction et gestion des sigles/expansions | **Projet CS/UM2 – *ProSigles***<br>porteur, responsable scientifique | 2007-2008 | Chap 1 |

TABLE 1 – Synthèse des travaux menés au LIRMM depuis 2005

1. D'autres collaborations et encadrements moins significatifs ont été menés depuis 2005. Ces derniers ne sont pas présents dans ce tableau 1 afin de conserver une vision synthétique de mes principaux travaux.

De manière générale, mes contributions appartiennent plus spécifiquement à une des phases du processus. Par exemple, les premiers travaux notés dans le tableau 1, s'intéressent plus spécifiquement à la phase d'extraction et d'acquisition d'un vocabulaire spécifique. Ainsi, je me suis intéressé à l'extraction de descripteurs complexes utiles pour les algorithmes classiques de classification de textes (cf. lignes 4 à 6). De la même manière, dans nos travaux liés à l'analyse de sentiments menés en collaboration avec le LGI2P (cf. ligne 2), nous avons concentré nos efforts sur la construction de dictionnaires de sentiments par rapport à un thème donné. Par exemple, l'adjectif *commercial* est associé à un sentiment *neutre* dans un cadre général. Pourtant, ce mot véhicule un sentiment plutôt *négatif* dans un contexte cinématographique.

Notons que les thèses que j'ai encadrées ces dernières années s'intéressent à des thèmes précis (titrage, extraction d'information dans les logs, classification conceptuelle) qui demandent néanmoins des contributions importantes pour chacune des phases. Ceci est crucial dans le but de proposer des systèmes globaux de bonne qualité.

Ce manuscrit synthétise chaque thème selon un fil directeur détaillé ci-dessous. La dernière colonne du tableau 1 précise dans quels chapitres du mémoire seront abordés les différents thèmes.

### Organisation du mémoire

Dans ce manuscrit d'Habilitation à Diriger les Recherche (HDR), les différentes phases seront abordées sous l'angle du matériau propre aux données textuelles qui sera appelé *descripteur linguistique*. Ce dernier est une chaîne de caractères (mot, syntagme, suite de caractères, etc) qui constitue une *Entrée* significative pour les algorithmes de Fouille de Textes. Dans ce document, nous aborderons différentes manières d'extraire et d'utiliser ces descripteurs. Ceci est illustré dans la figure 2.

Dans un premier temps, les informations sémantiques peuvent êtres issues des descripteurs en eux-même. De telles informations seront appelées des connaissances *endogènes* (section 1.1 du chapitre 1). Par exemple, dans le domaine biomédical, les suffixes sont souvent des indicateurs d'états pathologiques (*ite* désigne souvent une inflammation comme la pancréatite, appendicite, gastrite, *algie* ou *odynie* est associé à la douleur, etc.) ou des gestes techniques (*centèse* signifiant une ponction, *ectomie* associé à une ablation, *plastie* est propre à une réparation). Pour cela, plusieurs mesures fondées sur les chaînes de caractères peuvent être utilisées afin de déterminer les informations issues des descripteurs eux-mêmes.

Cependant, les informations endogènes ne sont pas toujours suffisantes, c'est la raison pour laquelle il peut se révéler utile de prendre en considération des facteurs *exogènes* (section 1.2 du chapitre 1). Dans ce manuscrit, nous exploiterons des "informations Web" (fonctions de rang qui s'appuient sur des méthodes de fouille du Web) pour obtenir des informations sémantiques comme la synonymie entre termes. Bien sûr, de manière plus globale, d'autres types de connaissances exogènes peuvent également être utilisées (par exemple, les taxonomies).

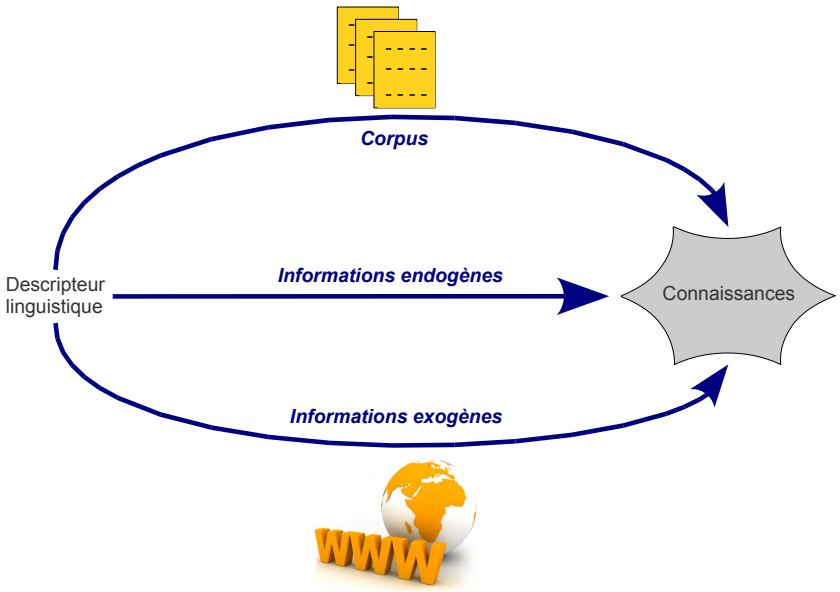


FIGURE 2 – Informations utilisées dans un processus de Fouille de Textes.

Toutes ces informations endogènes et exogènes sont propres à un descripteur linguistique donné. Elles seront abordées dans le chapitre 1 de ce mémoire. Dans le chapitre suivant, le descripteur sera placé dans un cadre plus vaste, c'est-à-dire au sein d'un corpus qui est ensemble de textes homogènes.

Dans un premier temps, l'utilisation d'un descripteur simple (*mot*) en corpus sera traité pour des tâches de classification de documents. Pour de telles tâches, l'identification des descripteurs est souvent difficile à partir de corpus bruités et à faible contenu textuel sur lesquels nous avons concentré nos efforts (corpus issus du Web 2.0 ou d'OCR [2] par exemple). Nous avons abordé ces travaux dans le cadre de plusieurs collaborations industrielles et académiques (cf. lignes 4 à 6 du tableau 1). Une partie de ces travaux sera traitée dans la section 2.1 du chapitre 2.

Outre les mots considérés comme des descripteurs linguistiques pertinents en Fouille de Textes, nous nous sommes également intéressés à l'étude des syntagmes qui sont des groupes de mots. Ceci fait suite aux travaux menés durant ma thèse effectuée à l'Université Paris-Sud (2001-2004). Depuis 2005, je me suis focalisé sur l'extraction d'une *terminologie complexe* [3] à partir de *corpus classiques* (corpus journalistique par exemple) et d'une *terminologie classique* [4] à partir de *corpus complexes* (fichiers logs, textes écrits en français médiéval). Ces deux types d'extraction de la terminologie seront abordés en section 2.2 du chapitre 2.

2. Optical Character Recognition.
3. Par exemple, syntagmes considérés comme des titres.
4. Bigrammes ou trigrammes de mots respectant des patrons syntaxiques définis.

Le chapitre 3 ne traite pas de l'extraction de syntagmes en eux même dans les corpus mais de leur *induction*. Ceci signifie que les syntagmes traités ne se situent pas à proprement parler dans les textes mais sont construits à partir des mots du corpus. Après leur construction (phase 1 de la figure 1) des processus de validation sont appliqués (phase 2 de la figure 1). Ces derniers sont fondés, en partie, sur les fonctions de rang proposées dans le chapitre 1.

Les différents types de raisonnement (déductif, abductif et inductif) issus du domaine de la logique sont décrits en section 3.1. Le raisonnement déductif s'appuie sur des règles solides d'implication logique. A contrario, les raisonnements abductifs et inductifs sont dits *hypothétiques*, c'est la raison pour laquelle des méthodes de validation automatique des relations induites se révèlent cruciales dans un processus de Fouille de Textes.

Ces étapes d'induction de relations syntagmatiques sont utilisées pour résoudre trois problèmes difficiles : identification d'Entités Nommées (section 3.2), construction de classes conceptuelles (section 3.3), titrage automatique (section 3.4).

Enfin, le chapitre 4 présente les perspectives à moyen et long terme que nous proposons. Ces dernières se situent à un niveau de granularité très fin (recherche de nouveaux descripteurs linguistiques) suivi d'une étude à un niveau plus vaste (corpus). Dans ce dernier contexte, nous proposons de nouvelles représentations ainsi que de nouvelles méthodes de stockage et de navigation dans les données textuelles. Enfin, nous présentons des pistes de travail afin d'analyser les données textuelles dans un contexte plus vaste lié au multimédia que le paradigme du Web 2.0 a mis en exergue ces dernières années.

*Dans la suite de ce manuscrit, les sections ou sous-sections propres aux travaux de recherche du tableau 1 seront contextualisées en y associant les articles de référence publiés (cf. textes encadrés en bleu). Par ailleurs, les références issues de mes propres publications porteront un numéro afin de les différencier des références bibliographiques notées avec le nom des premiers auteurs.*

# Chapitre 1

# Traitement du mot/terme

| *Thèmes de Recherche* | *Types de travaux* | *Années* |
|---|---|---|
| Extraction (80%) | | Exploitation (20%) |
| Analyse de sentiments | **Collaboration académique** – LIRMM/LGI2P (Nîmes)<br>*Stage M2 Recherche* | 2007-2011 |
| Extraction (20%) | Exploitation (80%) | |
| Mise en correspondance de schémas | **Projet ANR – *Forum***<br>membre<br>*Stage M2 Recherche* | 2005-2008 |
| Extraction et gestion des sigles/expansions | **Projet CS/UM2 – *ProSigles***<br>porteur, responsable scientifique | 2007-2008 |

Ce chapitre décrit les travaux menés sur le traitement des mots/termes. Pour cela, nous proposons des approches permettant de mesurer la proximité entre différentes chaînes de caractères. La première méthode utilise les *informations endogènes* des mots et/ou termes (cf. section 1.1). Les informations lexicales du contexte peuvent également être prises en compte afin de proposer une approche plus globale. Pourtant, ces seules informations sont, dans certains cas, insuffisantes. C'est la raison pour laquelle, la seconde approche proposée s'appuie sur l'utilisation de *connaissances exogènes* liées aux informations du Web (cf. section 1.2). Dans ce contexte, nous proposons de nouvelles fonctions de rang (cf. section 1.2.1) qui seront notamment utilisées pour la *désambiguisation d'acronymes* (cf. section 1.2.2) et l'*analyse de sentiments* (cf. section 1.2.3). Pour cette dernière application, notre principale contribution a trait à la construction de dictionnaires d'opinion spécialisés à un domaine. Ces fonctions de rang peuvent être adaptées à d'autres applications décrites dans les chapitres suivants (chapitres 2 et 3).

## 1.1 Le mot/terme, un descripteur endogène

Un mot ou un terme est constitué d'une chaîne de caractères qui peut contenir, intrinsèquement, des informations sémantiques. Par exemple, le mot *écoemballage* fait référence à deux concepts-clés : *l'emballage* et *le respect de l'environnement*. De manière générale, les mots ayant un préfixe *éco* peuvent faire référence au concept de l'écologie : *écocitoyen*, *écotourisme*, *écoterrorisme*[1], etc.

1. L'ensemble des termes présentés ici peuvent connaître une variation lexicale avec la présence d'un tiré après la chaîne *éco*. Cette présence permet d'associer ces termes à un lexique écologiste de manière plus explicite encore.

### 1.1.1 Mesures lexicales

Afin de calculer la similarité sémantique entre deux étiquettes (labels), de nombreuses mesures peuvent être utilisées [Euzenat et al., 2004, Maedche and Staab, 2002]. Nous décrivons ci-dessous deux mesures terminologiques classiques utilisées dans l'approche *Bmatch*. Ces mesures retournent des valeurs comprises dans l'intervalle [0, 1] : la valeur 1 signifie une similarité parfaite entre deux étiquettes, 0 dénote l'absence de similarité.

#### 1.1.1.1 n-grammes de caractères

La technique des n-grammes est utilisée pour calculer le nombre de n caractères consécutifs de différentes chaînes de caractères. Généralement, la valeur de n varie entre 2 et 5 et elle est souvent fixée à 3 [Kefi, 2006]. Par exemple, les tri-grammes entre les chaînes de caractères *"chat"* et *"chaton"* sont respectivement $tr(chat)$=$\{cha,\ hat\}$ et $tr(chaton)$ =$\{cha,\ hat,\ ato,\ ton\}$. Dans cet exemple, il y a donc deux tri-grammes communs : *cha* et *hat*. Pour calculer le taux de similarité entre deux éléments, la formule (1.1) issue des travaux de [Lin, 1998] est utilisée :

$$Tri(ch1, ch2) = \frac{1}{1 + |tr(ch1)| + |tr(ch2)| - 2 \times |tr(ch1) \cap tr(ch2)|} \in ]0, 1] \qquad (1.1)$$

Ainsi, dans notre cas, nous avons $Tri(chat, chaton) = \frac{1}{1+2+4-2\times2} = 0.33$.

Après avoir présenté la mesure des tri-grammes, nous pouvons utiliser une autre mesure appelée le String Matching décrite dans la section suivante.

#### 1.1.1.2 String Maching

Le String Matching, mesure proposée par [Maedche and Staab, 2002], s'appuie sur la distance d'édition (notée $E$) [Levenshtein, 1966]. Cette distance correspond à la somme minimale du coût des opérations à effectuer pour transformer deux chaînes de caractères. Les opérations prises en compte sont la suppression, l'insertion et le remplacement de caractères. Par exemple, nous pouvons relever deux opérations d'insertion (les caractères "o" et "n") entre les chaînes de caractères *"chat"* et *"chaton"*. Ainsi, nous avons $E(chat, chaton) = 2$. Le String Matching (noté $Str$) qui prend en compte la distance d'édition est donné par la formule (1.2).

$$Str(ch1, ch2) = \max\{0, \frac{\min\{|ch1|, |ch2|\} - E(ch1, ch2)}{\min\{|ch1|, |ch2|\}}\} \in [0, 1] \qquad (1.2)$$

Dans notre cas, nous avons $Str(chat, chaton) = \max\{0, \frac{4-2}{4}\} = 0.5$.

#### 1.1.1.3 Discussion

[Navarro, 1999] passe en revue d'autres mesures de la littérature qui permettent de comparer les chaînes de caractères. Outre les deux approches détaillées dans cette section et qui seront utilisées dans la suite, d'autres mesures s'appuient sur le simple calcul

du nombre de caractères qui diffèrent entre deux chaînes (distance de Hamming) ou la recherche de la plus plus grande sous-séquence commune. Comme le précise [Navarro, 1999], ce type de mesure est aussi bien utilisé en TAL qu'en traitement du signal ou en bioinformatique.

Les mesures terminologiques peuvent cependant avoir des limites dans le cas des :
- *Éléments polysémiques*. Par exemple, le mot "souris" peut avoir un sens très différent si le schéma traite d'informatique ou des Sciences du Vivant.
- *Éléments synonymes ou sémantiquement proches*. Par exemple, les animaux "souris" et "mulot" sont sémantiquement proches bien que les chaînes de caractères soient totalement distinctes.

Une approche plus globale doit alors être mise en place. Cette dernière dépend, en général, du domaine étudié, comme la *mise en correspondance de schémas* présentée en section suivante.

### 1.1.2 Mesures lexicales pour la mise en correspondance de schémas

La possibilité d'interroger des sources de données sémantiquement liées dépend uniquement de la capacité du système à trouver des correspondances entre leurs structures (ou schémas) et/ou leur contenu. Cette mise en correspondance s'effectue en grande partie manuellement ou semi-automatiquement. Par conséquent, les problèmes d'intégration sémantique sont devenus le goulot d'étranglement principal dans le déploiement des systèmes d'intégration à large échelle où le nombre de schémas à mettre en correspondance est très grand [Rahm and Bernstein, 2001, Doan et al., 2002, Tranier et al., 2004].

Notons que les approches consistant à mesurer la proximité sémantique entre les éléments d'arbres sont souvent essentielles pour la construction automatique d'ontologies [Aussenac-Gilles and Bourigault, 2003] ou leur alignement [Euzenat and Shvaiko, 2007].

Le projet ANR *Forum* [6] auquel j'ai participé à partir de 2005 consiste, entre autres, à proposer des méthodes de mise en correspondance de schémas à partir de données très spécialisées (épandage) sans informations lexicales et sémantiques disponibles. L'approche *Bmatch* développée dans le cadre du projet *Forum* s'appuie sur une combinaison de mesures terminologiques et d'informations contextuelles pour découvrir des correspondances entre schémas [8, 9]. Le principe mis en place qui n'utilise ni dictionnaire, ni connaissance spécifique aux langues étudiées est synthétisé ci-dessous.

Pour pallier les problèmes discutés en section 1.1.1.3, *Bmatch* propose de construire un contexte pour chaque élément à prendre en compte. Celui-ci est représenté par un vecteur formé avec les noms des éléments (labels) voisins. Le nombre d'éléments à considérer constitue un des paramètres de la mesure.

De manière plus précise l'approche *Bmatch* se décline en deux étapes :

- Une première étape consiste à remplacer par un même terme les éléments lexicalement proches du vecteur constitué. Ainsi, si deux termes sont proches, ils seront remplacés par celui ayant la taille la plus faible qui fournit en général une forme plus générique (comme dans le cas des flexions singulier/pluriel). Pour mesurer la proximité lexicale entre les termes, nous avons utilisé le String Maching et la mesure fondée sur les 3-grammes. La moyenne et le calcul du maximum entre ces deux mesures ont été expérimentés. Le choix du seuil de proximité lexicale choisi afin de remplacer les termes reste crucial. Celui-ci et l'ensemble des paramètres sont décrits, discutés et expérimentés dans [8, 9].

- L'étape suivante consiste à mesurer, via une mesure nommée $CM$, la proximité entre les vecteurs contextuels constitués. Pour cela nous avons appliqué la mesure de cosinus très répandue en Recherche d'Information. Cette mesure (formule (1.3)) a pour objectif de mesurer l'angle formé par les deux vecteurs. Le cosinus est défini comme la division du produit scalaire des vecteurs par le produit de leurs normes. Si deux vecteurs ont de nombreux descripteurs communs, la valeur retournée par le cosinus sera proche de 1.

$$CM(v_1, v_2) = \frac{v_1 \cdot v_2}{\sqrt{(v_1 \cdot v_1)(v_2 \cdot v_2)}} \tag{1.3}$$

  Notons que la mesure de cosinus est utilisée dans de nombreux travaux en fouille de textes présentés dans ce manuscrit. Par ailleurs, des études complémentaires fondées sur des méthodes d'apprentissage supervisé proposent des mesures de cosinus dites généralisées [Qamar and Gaussier, 2009].

Le principe global mis en place dans le cadre de *Bmatch* consiste à combiner cette mesure $CM$ fondée sur le contexte des éléments de l'arbre et l'approche $SM$ qui mesure la proximité lexicale entre les éléments. Différentes combinaisons entre $CM$ et $SM$ ont été expérimentées. Le calcul du maximum ($max(SM, CM)$) a donné les meilleurs résultats à partir de données réelles [8, 9].

Après avoir défini les mesures lexicales en elles-même, cette section s'intéresse à l'utilisation de connaissances exogènes afin de déterminer des liens sémantiques entre les mots/termes. Par exemple, l'utilisation d'informations sémantiques contenues dans les ontologies et taxonomies permet d'enrichir les méthodes de Recherche d'Information [Nyberg et al., 2010].

## 1.2 Le mot/terme associé à des connaissances exogènes

Dans cette section, des fonctions de rang initialement proposées dans [26] sont présentées. Les facteurs exogènes pris en compte sont liés aux informations issues du Web.

### 1.2.1 De nouvelles fonctions de rang

Les mesures que nous proposons de type fouille du Web (Web-Mining) s'appuient sur des information exogènes qui sont de deux ordres.

Dans un premier temps, nous mesurons la **présence d'associations** de mots sur le Web. Par exemple, les requêtes[2] : *"LIRMM Montpellier"* et *"LIMSI Montpellier"* via le moteur de recherche google[3] retournent des résultats dénotant la validité des associations entre les mots *LIRMM* (resp. *LIMSI*) et *Montpellier*. A contrario la requête *"LABRI Montpellier"* ne retourne aucun résultat montrant que l'association des deux mots composant cette requête est en fait non pertinente.

Cependant, cette seule information de présence/absence n'est pas toujours suffisante car la quantité de pages Web indexées est extrêmement importante[4] et engendre, inévitablement, la présence de documents bruités. Ainsi, nous devons aussi prendre en compte l'**intensité de cette association** qui repose sur le nombre de pages retournées par les moteurs de recherche avec les associations recherchées. Par exemple, les mêmes requêtes énoncées précédemment (*"LIRMM Montpellier"* et *"LIMSI Montpellier"*) retournent un nombre de pages différent (respectivement 11400 et 1) montrant une intensité de l'association qui n'est clairement pas du même ordre. L'intensité doit prendre en considération différentes informations (par exemple, informations contextuelles) et paramètres (par exemple, types d'opérateurs). De plus, pour mesurer la force de l'association, nous nous appuyons sur des critères statistiques (cf. section 1.2.1.2).

Nous avons appliqué et adapté notre mesure qui classe les associations possibles (fonction de rang) à deux types d'applications : la *désambiguisation d'acronymes* et la *fouille de données d'opinion*. Ces deux applications seront décrites en sections 1.2.2 et 1.2.3.

Pour illustrer notre mesure dans les sections suivantes, nous allons nous appuyer sur la tâche de désambiguisation des acronymes/définitions. Pour cela, à partir d'une liste d'expansions possibles d'acronymes, notre fonction de rang consiste à déterminer quelle expansion peut être considérée comme la plus pertinente. Par exemple, si l'acronyme "JO" est présent dans un texte, notre fonction de rang consiste à déterminer quelle définition (par exemple, "Jeux Olympiques" ou "Journal Officiel") est la plus adaptée.

#### 1.2.1.1 Origine des fonctions de rang proposées

Les fonctions de rang que nous proposons s'appuient sur les travaux de [Turney, 2001] et son algorithme PMI-IR (Pointwise Mutual Information and Information Retrieval). Ce dernier consiste à interroger le Web via le moteur de recherche AltaVista pour déterminer des synonymes appropriés. À partir d'un terme donné noté *mot*, l'objectif de PMI-IR est de choisir un synonyme parmi une liste donnée. Ces choix, notés $choix_i$, correspondent aux questions du TOEFL (Test of English as a Foreign Language). Ainsi, le but est de

2. En utilisant les guillemets pour effectuer une recherche exacte.
3. http ://www.google.com
4. Le nombre de pages web indexées par google est estimé vingt mille milliards de documents [Cacheda et al., 2010].

calculer, pour chaque *mot*, le synonyme $choix_i$ qui donne le meilleur score. Pour ce faire, l'algorithme PMI-IR utilise différentes mesures fondées sur la proportion de documents dans lesquels les deux termes sont présents. Nous donnons ci-dessous (formule (1.4)) une des mesures de base utilisée dans les travaux de [Turney, 2001]. Cette mesure s'appuie sur l'Information Mutuelle qui sera décrite dans la section 1.2.1.2.

$$score(\ choix_i\ ) = \frac{nb(\ mot\ NEAR\ choix_i\ )}{nb(\ choix_i\ )} \tag{1.4}$$

- $nb(x)$ calcule le nombre de documents contenant le mot $x$,
- $NEAR$ (utilisé dans la rubrique "recherche avancée" d'Altavista) est un opérateur qui précise si deux mots sont présents ensemble dans une fenêtre de 10 mots.

Ainsi, la formule (1.4) calcule la proportion de documents contenant *mot* et $choix_i$ dans une fenêtre de 10 mots par rapport au nombre de documents contenant le mot $choix_i$. Plus la proportion de documents contenant ces deux mots dans une même fenêtre est importante et plus *mot* et $choix_i$ sont considérés comme synonymes. D'autres formules plus élaborées ont également été appliquées. Elles utilisent les informations sur la présence de négations dans les fenêtres de 10 mots. Par exemple, les mots `"grand"` et `"petit"` ne sont pas synonymes si, dans une même fenêtre, la présence d'une négation associée à un des deux mots est relevée. D'autres approches utilisent le Web dans la littérature comme les travaux de [Cilibrasi and Vitanyi, 2007] qui proposent de mesurer la similarité de termes en utilisant, entre autres, le moteur de recherche Google (cf. section 1.2.1.3).

Notre approche possède des différences majeures par rapport à la méthode de [Turney, 2001]. Dans un premier temps, nous avons décidé de ne pas mesurer la dépendance entre les termes mais, de manière similaire aux travaux de [Daille, 1994], d'étudier la dépendance entre chacun des mots composant les syntagmes. Ceci permet de mesurer, de manière générique, la pertinence des syntagmes (par exemple, les *expansions des acronymes* dans ce chapitre ou les *titres* dans les deux prochains chapitres). De plus, l'Information Mutuelle utilisée par [Turney, 2001] est une mesure qui a des limites comme nous le montrerons par la suite. Ainsi, nos travaux s'appuient sur d'autres mesures de qualité. Enfin, l'utilisation d'un contexte spécifique permet d'améliorer significativement les mesures de base.

Précisons que notre approche n'utilise pas de corpus d'apprentissage ; les seules ressources utilisées sont les statistiques issues des moteurs de recherche. Notons enfin que contrairement à de nombreux travaux liés à la désambiguïsation sémantique [Audibert, 2003], notre approche n'utilise aucune connaissance linguistique telles que les informations lexicales et/ou syntaxiques. Cependant les informations grammaticales sont souvent des critères pertinents qui peuvent se révéler intéressants à associer aux mesures statistiques décrites dans la section suivante.

#### 1.2.1.2 Mesures statistiques utilisées

Dans la littérature, de nombreuses mesures de qualité sont utilisées afin d'effectuer un classement par intérêt décroissant. Ces mesures sont issues de domaines variés comme par exemple la recherche de règles d'associations [Azé, 2003, Lallich and Teytaud, 2004] ou l'extraction de la terminologie ([Daille, 1994] et [21]).

**Information Mutuelle**

Une des mesures couramment utilisée pour calculer une certaine forme de dépendance entre chacun des mots composant une co-occurrence est l'Information Mutuelle [Church and Hanks, 1990] :

$$I(x,y) \quad = \quad \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{1.5}$$

Une telle mesure a tendance à extraire des co-occurrences rares et spécifiques [Daille, 1994, Thanopoulos et al., 2002] et [21].

Notons que, dans la formule (1.5), l'utilisation de la fonction $log_2$ n'est pas nécessaire. En effet, la fonction $log_2$ est strictement croissante, l'ordre des co-occurrences donné par la mesure n'est donc pas affecté avec l'application ou non de la fonction $log_2$. Dans notre cas, $P(x,y)$ permet d'estimer la probabilité d'apparition des couples de mots ($x$,$y$) où $x$ et $y$ sont voisins dans cet ordre. Par exemple, avec l'acronyme JO, $x$ peut représenter le mot "Journal" et $y$ le mot "Officiel". Après diverses approximations [21], la formule (1.5) peut s'écrire de la manière suivante où $nb$ représente le nombre d'occurrences des mots et des couples de mots :

$$IM(x,y) \quad = \quad \log_2 \frac{nb(x,y)}{nb(x)nb(y)} \tag{1.6}$$

Cette mesure peut être adaptée aux co-occurrences ternaires de manière similaire aux travaux de [Jacquemin, 1997]. Ainsi, une extension naturelle consiste à appliquer cette mesure à des syntagmes formés de $n$ mots (formule (1.7)).

$$IM(x_1,...,x_n) \quad = \quad \log_2 \frac{nb(x_1,...,x_n)}{nb(x_1) \times ... \times nb(x_n)} \tag{1.7}$$

**Information Mutuelle au Cube**

L'Information Mutuelle au Cube [Daille, 1994] est une mesure empirique qui s'appuie sur l'Information Mutuelle mais en privilégiant davantage les co-occurrences fréquentes. Une telle mesure est définie par la formule (1.8).

$$IM3(x,y) \quad = \quad \log_2 \frac{nb(x,y)^3}{nb(x)nb(y)} \tag{1.8}$$

L'Information Mutuelle au Cube est utilisée dans de nombreux travaux liés à l'extraction des termes nominaux ou verbaux [Vivaldi et al., 2001, Claveau and Sébillot, 2003] ou des entités nommées [Downey et al., 2007] dans les textes. [Vivaldi et al., 2001] ont d'ailleurs estimé que l'Information Mutuelle au Cube était la mesure qui avait le meilleur comportement. De la même manière que l'Information Mutuelle, une telle mesure peut être étendue de la manière suivante :

$$IM3(x_1, ..., x_n) = \log_2 \frac{nb(x_1, ..., x_n)^3}{nb(x_1) \times ... \times nb(x_n)} \quad (1.9)$$

**Coefficient de Dice**

Dans la suite nous allons présenter une autre mesure de qualité appelée coefficient de Dice [Smadja et al., 1996]. Cette mesure est définie par la formule (1.10).

$$D(x, y) = \frac{2 \times P(x, y)}{P(x) + P(y)} \quad (1.10)$$

De manière similaire à l'Information Mutuelle au Cube, le coefficient de Dice privilégie moins les co-occurrences rares souvent non pertinentes [23].

La formule (1.10) permet de directement déduire [5] la formule (1.11) qui s'appuie sur le nombre d'occurrences $nb$ des mots et des couples de mots.

$$Dice(x, y) = \frac{2 \times nb(x, y)}{nb(x) + nb(y)} \quad (1.11)$$

Dans les travaux de [Petrovic et al., 2006], les auteurs présentent une extension de la formule d'origine de Dice à trois éléments :

$$Dice(x, y, z) = \frac{3 \times nb(x, y, z)}{nb(x) + nb(y) + nb(z)} \quad (1.12)$$

Nous pouvons, de la même manière, proposer une extension naturelle à $n$ éléments :

$$Dice(x_1, ..., x_n) = \frac{n \times nb(x_1, ..., x_n)}{nb(x_1) + ... + nb(x_n)} \quad (1.13)$$

Les deux sections suivantes (sections 1.2.1.3 et 1.2.1.4) présentent la mesure $DeMT$ que nous proposons (mesure de base et mesure contextuelle) qui s'appuie sur l'utilisation du coefficient de Dice.

#### 1.2.1.3 Fouille du Web et mesures statistiques

Notre contexte de travail est lié à l'élaboration d'une mesure de qualité qui s'appuie sur les ressources du Web. Dans ce cas, la fonction $nb$ utilisée dans les différentes mesures qui ont été détaillées dans la section précédente représente le nombre de pages retournées par des moteurs de recherche (Exalead [6] ou Google [7]).

5. en posant, $P(x) = \frac{nb(x)}{nb\_total}$, $P(y) = \frac{nb(y)}{nb\_total}$, $P(x, y) = \frac{nb(x,y)}{nb\_total}$
6. http ://www.exalead.fr/
7. www.google.com/

À partir de la formule (1.13), nous pouvons en déduire la formule (1.14) qui représente la mesure $DeMT$ (Dépendance entre les Mots du Terme) de base.

$$DeMT_{Dice}(a^j) = \frac{\left|\{a_i^j; a_i^j \notin M_{outils}\}_{i\in[1,n]}\right| \times nb(\bigcap_{i=1}^{n} a_i^j)}{\sum_{i=1}^{n} nb(a_i^j; a_i^j \notin M_{outils})} \text{ où } n \geq 2 \tag{1.14}$$

où

- $\bigcap_{i=1}^{n} a_i^j$ désigne la suite de mots $a_i^j$ ($i \in [1, n]$) que l'on considère comme une chaîne de caractères (utilisation des *guillemets* avec Exalead ou Google que l'on peut illustrer de la manière suivante : "$a_1^j ... a_n^j$").
- $M_{outils}$ représente une liste de mots outils (prépositions, articles, etc). Le but est de ne pas considérer le nombre de pages possédant ces mots outils qui ne sont pas porteurs de sens.
- $|.|$ désigne le nombre de mots de l'ensemble.

Avec l'acronyme $a =$`JO`, deux définitions sont possibles (voir http ://www.sigles.net/) : $a^1$ : `Jeux Olympiques` et $a^2$ : `Journal Officiel`

Les scores obtenus avec la mesure donnée par la formule (1.14) sont très proches [8].

- $DeMT_{Dice}(\texttt{JO}^1) = \frac{2\times nb(\texttt{Jeux} \cap \texttt{Olympiques})}{nb(\texttt{Jeux})+nb(\texttt{Olympiques})} = \frac{2\times 366508}{116929964+1207545} = 0.0062$
- $DeMT_{Dice}(\texttt{JO}^2) = \frac{2\times nb(\texttt{Journal} \cap \texttt{Officiel})}{nb(\texttt{Journal})+nb(\texttt{Officiel})} = \frac{2\times 603036}{178302348+28140994} = 0.0058$

Dans la pratique, le premier exemple revient à effectuer les trois requêtes suivantes : `"Jeux Olympiques"`, `Jeux`, `Olympiques`. Notons que dans ce cas, davantage de pages sont retournées avec la requête `"Journal Officiel"`, pourtant le score le plus élevé est obtenu avec `"Jeux Olympiques"`.

D'autres mesures Web de la littérature peuvent aussi être utilisées comme "Google Similarity Distance" [Cilibrasi and Vitanyi, 2007] donnée par la formule (1.15).

$$NGD(x, y) = \frac{max\{log(nb(x)), log(nb(y))\} - log(nb(x, y))}{log(N) - min\{log(nb(x)), log(nb(y))\}} \tag{1.15}$$

L'avantage de la mesure de Dice comparativement à la distance de [Cilibrasi and Vitanyi, 2007] tient au fait que, dans la mesure finale, nous ne prenons pas en compte le nombre total de pages Web indexées par les moteurs de recherche (noté $N$ dans la formule (1.15)). L'utilisation de cette valeur prise en compte dans [Cilibrasi and Vitanyi, 2007] peut se révéler problématique car elle évolue, elle est dépendante des moteurs de recherche et est souvent approximative [Cacheda et al., 2010] voire contestée. Dans le cas de la mesure de Dice, nous ne rencontrons pas ce type de situation. En effet, à partir de la formule de base de Dice (formule (1.10)), cette valeur est simplifiée (l'inverse de ce nombre de pages étant situé au numérateur et dénominateur avant simplification).

8. Requêtes effectuées en décembre 2006.

#### 1.2.1.4 Contextualisation de l'approche

La mesure de base proposée a une limite majeure liée au fait que le score ne prend pas en compte le contexte. Ainsi, dans le cas de la désambiguisation des acronymes, nous proposons de considérer ce dernier pour effectuer un choix de définition plus pertinent. Nous définissons le contexte comme des mots caractéristiques présents dans la page dans laquelle l'acronyme à définir est présent.

Plusieurs contextes $C$ peuvent être utilisés :
- $n$ mots les plus fréquents (sauf les mots outils).
- $n$ noms propres les plus fréquents.
- $n$ mots les plus rares.
- Utilisation d'informations grammaticales (noms, verbes, etc) [Brill, 1994] et/ou de la terminologie [Daille, 1994, Bourigault and Jacquemin, 1999] et [21].

Une combinaison de ces contextes peut également être envisagée. Notons que les expérimentations présentées dans ce manuscrit s'appuient sur le contexte représenté par les mots les plus fréquents des textes dans lesquels l'acronyme doit être désambiguisé. En effet, un tel contexte retourne des résultats tout à fait satisfaisants.

L'ajout d'informations contextuelles à la mesure $DeMT$ (formule (1.14)) permet la construction de la mesure $DeMTC$ (Dépendance entre les Mots du Terme Contextualisée) (formule (1.16)). Le principe de cette mesure contextuelle est d'appliquer des approches statistiques sur un ensemble qui est propre au domaine étudié. La dépendance des mots de la définition de l'acronyme est alors calculée à partir des seules pages partageant un contexte proche.

$$DeMTC_{Dice}(a^j) = \frac{\left|\{a_i^j + C; a_i^j \notin M_{outils}\}_{i\in[1,n]}\right| \times nb((\bigcap_{i=1}^{n} a_i^j) + C)}{\sum_{i=1}^{n} nb(a_i^j + C; a_i^j \notin M_{outils})} \tag{1.16}$$

Dans la formule (1.16) où $n \geq 2$, $a_i^j + C$ représente le mot $a_i^j$ avec tous les mots du contexte $C$. $nb(a_i^j + C)$ retourne le nombre de pages données par le moteur de recherche avec la requête $a_i^j + C$ (utilisation de l'opérateur $AND$ d'Exalead).

Reprenons l'exemple de l'acronyme $a =$`JO`, qui possède deux définitions possibles (`Jeux Olympiques` et `Journal Officiel`). Rappelons qu'avec la mesure de base, la définition privilégiée est toujours `Jeux Olympiques` :

$$DeMT_{Dice}(\texttt{JO}^1) = 0.0062 \text{ et } DeMT_{Dice}(\texttt{JO}^2) = 0.0058$$

Considérons le contexte $C = \{$`loi`$\}$. Dans ce cas, nous avons :
- $DeMTC_{Dice}(\texttt{JO}^1) = \frac{2\times nb((\texttt{Jeux} \cap \texttt{Olympiques}) + \texttt{loi})}{nb(\texttt{Jeux} + \texttt{loi}) + nb(\texttt{Olympiques} + \texttt{loi})} = 0.018$
- $DeMTC_{Dice}(\texttt{JO}^2) = \frac{2\times nb((\texttt{Journal} \cap \texttt{Officiel}) + \texttt{loi})}{nb(\texttt{Journal} + \texttt{loi}) + nb(\texttt{Officiel} + \texttt{loi})} = 0.159$

Dans la pratique, le premier exemple revient à effectuer les trois requêtes suivantes : `"Jeux Olympiques" AND loi`, `Jeux AND loi`, `Olympiques AND loi`.

La mesure prenant en compte le contexte $C = \{$`loi`$\}$ permet de privilégier la définition `Journal Officiel` à associer à l'acronyme `JO`. Cette mesure revient à calculer le coefficient de Dice à partir des seules pages contenant le mot `loi`.

En utilisant le contexte $C = \{\texttt{sport}\}$, nous obtenons :

$$DeMTC_{Dice}(\texttt{JO}^1) = 0.025 \text{ et } DeMTC_{Dice}(\texttt{JO}^2) = 0.010$$

Ainsi, dans ce cas, la définition `Jeux Olympiques` est privilégiée. Bien entendu, le fait de donner un contexte encore plus riche (composés de plusieurs mots) permet d'accentuer les écarts des scores pour les deux définitions. Par exemple, avec $C = \{\texttt{sport, natation}\}$ nous avons : $DeMTC_{Dice}(\texttt{JO}^1) = 0.190$ et $DeMTC_{Dice}(\texttt{JO}^2) = 0.008$.

Notons enfin que la mesure $DeMTC_{Dice}$ qui est proposée dans nos travaux et les mesures qui sont présentées dans la section suivante sont indépendantes des langues des textes étudiés.

#### *DeMTC* fondée sur l'Information Mutuelle et l'Information Mutuelle au Cube

De manière similaire à la formule (1.16), les formules (1.17) et (1.18) présentent respectivement la mesure *DeMTC* fondée cette fois-ci sur l'Information Mutuelle et l'Information Mutuelle au Cube.

$$DeMTC_{IM}(a^j) = \frac{nb((\bigcap_{i=1}^{n} a_i^j) + C)}{\prod_{i=1}^{n} nb(a_i^j + C; a_i^j \notin M_{outils})} \text{ où } n \geq 2 \tag{1.17}$$

$$DeMTC_{IM3}(a^j) = \frac{nb((\bigcap_{i=1}^{n} a_i^j) + C)^3}{\prod_{i=1}^{n} nb(a_i^j + C; a_i^j \notin M_{outils})} \text{ où } n \geq 2 \tag{1.18}$$

L'ensemble de ces mesures sont utilisées afin de valider certains termes complexes (cf. chapitre 2) ainsi que des termes construits (cf. chapitre 3). Par ailleurs, ces mesures de base ont été enrichies et/ou adaptées aux deux problématiques (désambiguisation d'acronymes et analyse de sentiments) décrites dans les sections suivantes (sections 1.2.2 et 1.2.3).

### 1.2.2 Fonctions de rang et désambiguisation d'acronymes

Dans cette section, nous présentons la manière dont nous avons enrichi les mesures de base présentées précédemment pour la désambiguisation des acronymes/définitions. Dans un premier temps nous décrivons un processus global qui s'appuie sur différents traitements de fouille de textes qui sont résumés dans la section 1.2.2.1.

Le processus global de gestion des acronymes/définitions est détaillé dans [27]. Dans le cadre de ce processus, nous avons étendu les mesures de désambiguisation comme nous le montrerons en section 1.2.2.3.

#### 1.2.2.1 Processus global de désambiguisation d'acronymes

Cette section s'intéresse au problème spécifique des acronymes qui sont particulièrement propices au problème de polysémie. Un acronyme est l'abréviation d'un groupe de mots formé, en général, par les initiales de ces mots. Une distinction existe entre les *sigles* dont chaque lettre est épelée (par exemple, SNCF) contrairement aux *acronymes* qui sont prononcés comme des mots classiques (par exemple, OVNI). Cependant, dans cette section nous utiliserons le même mot "acronyme" pour désigner ces deux situations qui peuvent se révéler difficiles à distinguer de manière automatique. Au même titre que les mots, les acronymes ont souvent plusieurs sens. Par exemple, comme nous l'avons vu, l'acronyme "JO" peut être associé aux définitions "Jeux Olympiques" ou "Journal Officiel". Quelques ressources plus ou moins spécialisées existent et proposent des définitions possibles pour un même acronyme. À titre d'exemple, le site http ://www.sigles.net/ fournit une telle liste.

Le problème concerne les textes pour lesquels aucune définition d'acronyme n'est présente. La difficulté est donc de choisir de manière automatique la définition la plus adaptée.

Dans ce contexte, posons $a$ un acronyme donné (par exemple, $a$ =JO). Pour chaque $a$ dont la définition n'est pas présente dans un document $d$, considérons que nous avons une liste de $n$ définitions possibles pour $a$ : $a^1...a^n$ (par exemple, $a^1$=Jeux Olympiques, $a^2$=Journal Officiel). L'objectif de notre approche est de déterminer $k$ ($k \in [1, n]$) tel que $a^k$ soit la définition pertinente pour le document $d$. Pour effectuer un tel choix, nous proposons des mesures de qualité, $DeMT$ et $DeT$, qui s'appuient notamment sur les ressources du Web. La figure 1.1 résume le processus global appliqué. Dans ce schéma la mesure de rang sera appelée $DefAcro$ afin de se ramener à la problématique de désambiguisation des définitions propres aux acronymes.

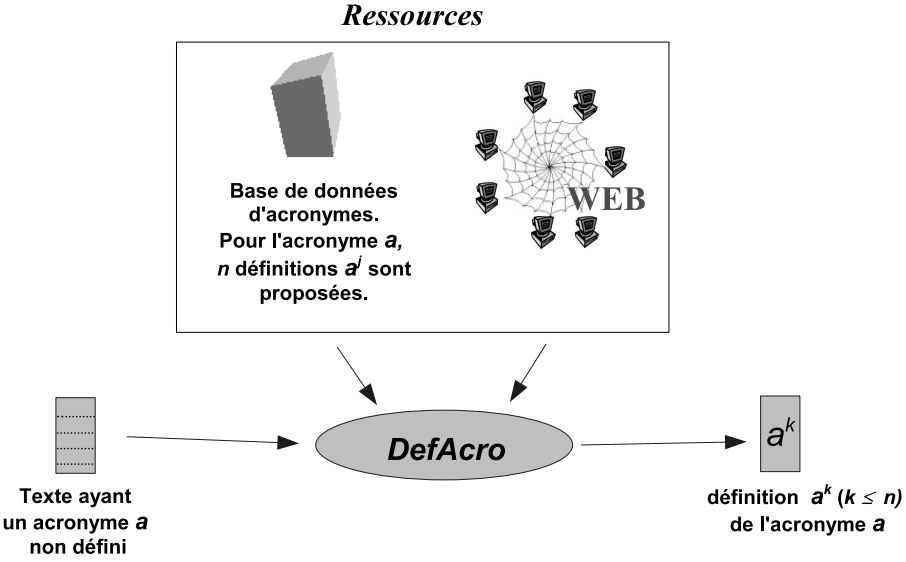


FIGURE 1.1 – Processus global.

Nos travaux issus du projet ProSigles [9] développé par le LIRMM possèdent deux phases distinctes qui sont synthétisées ci-après.

- **Phase 1 :** Les acronymes et leur(s) définition(s) sont tout d'abord extraits à partir de données textuelles quelconques (domaines spécialisés ou non, différentes langues, etc). Une telle phase permet d'acquérir ou d'enrichir des dictionnaires plus ou moins spécialisés. La méthode mise en œuvre possède deux étapes qui sont détaillées dans la section 1.2.2.2 : extraction des acronymes candidats et filtrage de ces derniers.
- **Phase 2 :** Une fois ces dictionnaires constitués, nous utilisons les mesures statistiques qui consistent à déterminer la définition pertinente d'un acronyme présent dans un document. Dans ces documents, la définition n'est cependant pas présente d'où la difficulté du traitement. Dans ce contexte, il est donc essentiel d'avoir à disposition un dictionnaire adapté ce qui justifie la première phase du processus présenté. Notre approche décrite dans la section 1.2.2.3 utilise entre autres la mesure $DeMT$ pour sélectionner la définition adaptée.

#### 1.2.2.2 Extraction des acronymes/définitions (phase 1 du projet ProSigles)

De nombreuses méthodes pour extraire les acronymes et leur définition existent dans la littérature. La plupart des approches de détection d'acronyme dans les textes s'appuient sur l'utilisation de marqueurs spécifiques. La méthode de [Yeates, 1999] consiste dans un premier temps à séparer les phrases par fragments en utilisant des marqueurs spécifiques (parenthèses, points, etc) comme frontières. Par exemple, la phrase :
`Les JO (Jeux Olympiques) ont été organisés en Chine en 2008.`
devient
`Les JO | Jeux Olympiques | ont été organisés en Chine en 2008 |`

L'étape suivante a pour but de comparer chaque mot de chacun des fragments avec les fragments précédents et suivants. Ainsi, dans notre exemple, les comparaisons suivantes sont effectuées : `Les` avec `Jeux Olympiques` ; `Jeux` avec `Les JO` ; `JO` avec `Jeux Olympiques` ; `Olympiques` avec `Les JO`, etc.

Ensuite, les couples acronymes/définitions sont testés. Les candidats acronymes sont retenus si les lettres de l'acronyme sont mises en correspondance avec les premières lettres des définitions potentielles. Dans notre cas, le couple "JO/Jeux Olympiques" est un candidat acronyme. La dernière étape consiste à utiliser des heuristiques spécifiques pour retenir les candidats pertinents. Ces heuristiques s'appuient sur le fait que les acronymes ont une taille plus petite que leur définition, qu'ils sont en majuscule, que les définitions des acronymes ayant une longueur importante ont tendance à posséder davantage de mots outils (par exemple, les articles et les prépositions), etc. Dans notre cas, le couple "JO/Jeux Olympiques" qui vérifie ces heuristiques peut alors être considéré comme un acronyme. De nombreuses approches [Larkey et al., 2000, Chang et al., 2002] utilisent des méthodes similaires fondées sur la présence de marqueurs associés à des heuristiques spécifiques.

9. Projet financé par le Conseil Scientifique de l'Université Montpellier 2, France.

Quelques méthodes recherchent des définitions des acronymes en utilisant le Web. Par exemple, l'utilisation d'un moteur de recherche intervient dans les travaux de [Larkey et al., 2000] afin d'enrichir un corpus initial de pages Web utiles pour la recherche d'acronymes. Pour ce faire, à partir d'une liste d'acronymes, des requêtes sont soumises au moteur de recherche AltaVista [10]. Ceci permet d'acquérir des pages Web dont les URLs sont à leur tour explorées pour enrichir le corpus des pages Web.

D'autres travaux s'appuient sur la présence de marqueurs linguistiques souvent associées à des heuristiques [Chang et al., 2002, Larkey et al., 2000]. Par exemple les travaux de [Okazaki and Ananiadou, 2006] utilisent des mesures statistiques pour l'extraction de la terminologie issue de domaine de spécialité (biomédecine). Ainsi, [Frantzi et al., 2000, Nenadic et al., 2003] appliquent la mesure appelée C-value dans un tel contexte. Celle-ci favorise l'extraction de termes qui n'apparaissent pas souvent dans des termes plus longs. Par exemple, dans un corpus d'ophtalmologie le terme 'soft contact' n'est pas pertinent ; il est souvent présent dans un terme plus long ('soft contact lens'). L'avantage de l'approche de [Okazaki and Ananiadou, 2006] tient au fait qu'elle est indépendante de l'alignement de caractères entre l'acronyme et son expansion.

D'autres approches, fondées sur des méthodes supervisées, utilisent des descripteurs numériques (longueur, présence de caractères spéciaux, contexte) [Xu and Huang, 2007]. Contrairement aux approches d'apprentissage supervisé de la littérature [Stevenson et al., 2009, Joshi et al., 2006] qui ont été développées dans le domaine biomédical, l'approche utilisée dans le contexte du projet ProSigles s'appuie sur une méthode non supervisée. Elle se décline en deux étapes distinctes :

***Étape 1 : Extraction des candidats acronymes/définitions***

De manière similaire aux travaux décrits précédemment, notre méthode, consistant à extraire les candidats acronymes/définitions, utilise des marqueurs (parenthèses, crochets, etc). Ils visent à identifier soit un acronyme, soit une définition, ce qui nécessite la prise en compte de deux traitements différents :

*Premier cas :* l'acronyme se situe avant la définition qui se trouve entre les marqueurs (les parenthèses dans le cas le plus courant). Exemple : `"... S.I.G. (Solde Intermédiaire de Gestion)..."`

*Deuxième cas :* la définition se trouve avant l'acronyme qui se situe entre les marqueurs. Exemple : `"... les Systèmes d'Informations Géographiques (SIG) ..."`. Dans ce cas, la taille de la définition est pour le moment indéterminable. C'est pourquoi, il est nécessaire de la définir arbitrairement en fonction du nombre de lettres composant l'acronyme. Ce choix doit tenir compte des mots qui pourraient venir s'intercaler sans pour autant nous intéresser, comme par exemple les prépositions ou les articles. Nous avons expérimentalement fixé cette taille à trois fois le nombre de lettres composant l'acronyme. Par exemple, la définition potentielle choisie pour l'exemple "SIG" sera composée des neuf mots qui précèdent cet acronyme.

Le but de cette première étape consiste à extraire tous les candidats acronymes/définitions

10. http ://www.altavista.com/

pertinents. Bien entendu, cette phase retourne une quantité importante de bruit car elle s'appuie seulement sur les marqueurs pour identifier un candidat potentiel. Ainsi, la seconde étape de notre approche consiste à filtrer les couples acronymes/définitions pertinents parmi la liste retournée.

***Étape 2 : Filtrage des candidats***

La seconde étape de notre application utilise donc les résultats obtenus lors du premier traitement qui vient d'être décrit. Les résultats sont triés afin de : supprimer les paires acronymes/définitions non pertinentes ; extraire précisément les définitions présentes dans les définitions potentielles (ces dernières pouvant être trop longues puisque coupées arbitrairement lors du second cas de la recherche des candidats).

Pour permettre un tel filtrage, nous effectuons un alignement des lettres contenues dans l'acronyme avec les mots de la définition. Cet alignement consiste à vérifier la correspondance entre les lettres des acronymes avec les premières lettres de chacun des mots des définitions. Dans notre méthode, si le premier caractère des mots de la définition candidate ne peut être aligné, les caractères qui suivent au sein des mots sont considérés. Par exemple, cette méthode permet de reconnaître "Extraction Itérative de la Terminologie" comme la définition de l'acronyme EXIT dans lequel la lettre "X" a pu être alignée. Notons enfin que les mots outils (prépositions, articles, etc) sont considérés comme des mots quelconques sans traitement spécifique contrairement à certains travaux qui utilisent une telle liste dans le processus [Larkey et al., 2000]. Ceci a pour but d'avoir une méthode indépendante des langues et de considérer ces mots qui permettent de constituer un acronyme (par exemple, la préposition "de" pour "GDF / Gaz de France").

Nous présentons dans la table 1.1 une évaluation de notre système d'alignement des acronymes avec les définitions candidates. Pour cette évaluation, nous nous appuyons sur les données issues du site http ://www.sigles.net/ proposant 25463 acronymes et leurs définitions issus de 17 langues. L'évaluation consiste à extraire aléatoirement de ce site des acronymes de 2, 3 et 4 caractères et d'évaluer le taux de réussite de l'alignement (nombre d'acronymes alignés avec les définitions du site en utilisant la version actuelle de notre logiciel). Le tableau 1.1 présente les résultats de 800 cas de mise en correspondance qui se sont révélées globalement très satisfaisants (taux de réussite de 78% à 98%). Par ailleurs, ce tableau montre que les acronymes longs sont plus difficiles à aligner. Ceci est, par exemple, dû à la présence de lettres en majuscule accentuées qui ne sont pas encore considérées par notre logiciel. De telles améliorations techniques sont assez aisées à mettre en œuvre. Cependant, notons que de nombreux cas particuliers plus difficiles à traiter peuvent exister comme l'alignement de caractères numériques / non numériques (par exemple, "3D / Trois Dimensions", "ST2I / Sciences et Techniques de l'Informatique et de l'Ingénierie").

| Nb de lettres | Nb d'acronymes | Nb de définitions | Nb de définitions non reconnues | Pourcentage de réussite |
|---|---|---|---|---|
| 2 | 100 | 616 | 11 | 98.2 % |
| 3 | 50 | 157 | 10 | 93.6 % |
| 4 | 20 | 32 | 7 | 78.1 % |

TABLE 1.1 – Taux de réussite de l'alignement acronymes/définitions.

#### 1.2.2.3 Sélection des expansions pertinentes par *DeMT* et *DeT* (phase 2 du projet ProSigles)

Outre les mesures de base décrites dans la section 1.2.1.3 et dans [26], cette section présente des mesures plus riches qui sont adaptées à la problématique de la désambiguisation des acronymes. Ces dernières sont également détaillées dans [28].

**Mesure *DeT***

La mesure *DeMT* que nous avons présentée consiste à calculer la Dépendance entre les Mots constituant le Terme. Ceci est plus proche des travaux liés à la terminologie. Les travaux présentés dans ce document reposent en grande partie sur une telle mesure car nos différents travaux consistent à extraire et/ou construire des syntagmes pertinents.

Dans cette section, nous proposons également de mesurer la **dépendance entre les termes**. Dans notre contexte, ceci consiste à calculer la dépendance entre l'acronyme et son expansion possible. Les mesures présentées ci-dessous appelées *DeT* (Dépendances entre les Termes) sont plus proches des travaux de [Turney, 2001].

Dans un premier temps, nous proposons une approche générique qui permet de mesurer la dépendance entre les termes $a$ et $b$. Dans ce cas, le dénominateur $nb(a)$ propre à la formule de l'information mutuelle n'est pas utile car il constitue une constante qui n'influence pas l'ordre des $b_i$ déterminé par notre mesure *DeT*.

$$DeT_{IM}^{And}(b_i) = \frac{nb(a\ AND\ b_i)}{nb(b_i)} \tag{1.19}$$

En appliquant cette mesure à notre problématique de désambiguisation d'acronymes, nous obtenons la formule (1.20).

$$DeT_{IM}^{And}(\ a^j\ ) = \frac{nb(\ a\ \text{AND}\ \bigcap_{i=1}^{n} a_i^j\ )}{nb(\ \bigcap_{i=1}^{n} a_i^j\ )} \tag{1.20}$$

Par exemple, $nb(\ a\ \text{AND}\ \bigcap_{i=1}^{n} a_i^j\ )$ avec $a =$ `IR` et $\bigcap_{i=1}^{2} a_i^j =$ `Information` $\cap$ `Retrieval` calcule le nombre de pages retournées avec la requête `IR AND "Information Retrieval"`. Nous obtenons le nombre de fois où les termes IR and "Information Retrieval" sont présents dans la même page.

De manière plus précise, pour calculer la dépendance entre les termes $a$ (par exemple, "IR") et $\bigcap_{i=1}^{n} a_i^j$ (par exemple, "Information Retrieval"), nous pouvons considérer le nombre de pages où les mots sont présents dans une même fenêtre en utilisant l'opérateur *NEAR*.

La formule (1.21) calcule cette dépendance :

$$DeT_{IM}^{Near}(\ a^j\ ) = \frac{nb(\ a \text{ NEAR } \bigcap_{i=1}^{n} a_i^j\ )}{nb(\ \bigcap_{i=1}^{n} a_i^j\ )} \tag{1.21}$$

Dans la suite, nous pouvons étendre ces fonction de rang en utilisant d'autres mesures statistiques (Information Mutuelle au Cube, Dice) associées à différents opérateurs (*AND*, *NEAR*).

$$DeT_{IM3}^{And}(\ a^j\ ) = \frac{nb(\ a \text{ AND } \bigcap_{i=1}^{n} a_i^j\ )^3}{nb(\ \bigcap_{i=1}^{n} a_i^j\ )} \tag{1.22}$$

$$DeT_{IM3}^{Near}(\ a^j\ ) = \frac{nb(\ a \text{ NEAR } \bigcap_{i=1}^{n} a_i^j\ )^3}{nb(\ \bigcap_{i=1}^{n} a_i^j\ )} \tag{1.23}$$

$$DeT_{Dice}^{And}(\ a^j\ ) = \frac{2 \times nb(\ a \text{ AND } \bigcap_{i=1}^{n} a_i^j\ )}{nb(\ a\ ) + nb(\ \bigcap_{i=1}^{n} a_i^j\ )} \tag{1.24}$$

$$DeT_{Dice}^{Near}(\ a^j\ ) = \frac{2 \times nb(\ a \text{ NEAR } \bigcap_{i=1}^{n} a_i^j\ )}{nb(\ a\ ) + nb(\ \bigcap_{i=1}^{n} a_i^j\ )} \tag{1.25}$$

**Mesure** $DeTC$

Pour la mesure $DeMT$, nous avons ajouté un contexte et proposé la mesure $DeMTC$. De la même manière, nous pouvons contextualiser la mesure $DeT$. Ces mesures contextualisées sont appelées $DeTC$. Concrètement, nous ajoutons un contexte $C$ (en appliquant l'opérateur *AND*) aux formules (1.20), (1.21), (1.22), (1.23), (1.24) et (1.25). Un tel contexte constitue également une extension de la mesure de base de Peter Turney.

**Quelques exemples**

Nous appliquons ici les mesures de qualité avec l'acronyme *IR* afin de comparer les différents modes de calcul fondés sur les approches présentées dans ce chapitre. Nous donnons le détail des calculs avec l'expansion possible *Information Retrieval*.

- $DeMT_{Dice}$ : $\frac{2 \times nb(\texttt{Information} \cap \texttt{Retrieval})}{nb(\texttt{Information}) + nb(\texttt{Retrieval})}$ [formule (1.16)]
- $DeMT_{IM}$ : $\frac{nb(\texttt{Information} \cap \texttt{Retrieval})}{nb(\texttt{Information}) \times nb(\texttt{Retrieval})}$ [formule (1.17)]

- $DeMT_{IM3}$ : $\frac{nb(\texttt{Information} \cap \texttt{Retrieval})^3}{nb(\texttt{Information}) \times nb(\texttt{Retrieval})}$ [formule (1.18)]
- $DeT^{And}_{Dice}$ : $\frac{2 \times nb(\texttt{IR AND (Information} \cap \texttt{Retrieval))}}{nb(\texttt{IR}) + nb(\texttt{Information} \cap \texttt{Retrieval})}$ [formule (1.24)]
- $DeT^{Near}_{Dice}$ : $\frac{2 \times nb(\texttt{IR NEAR (Information} \cap \texttt{Retrieval))}}{nb(\texttt{IR}) + nb(\texttt{Information} \cap \texttt{Retrieval})}$ [formule (1.25)]
- $DeT^{And}_{IM}$ : $\frac{nb(\texttt{IR AND (Information} \cap \texttt{Retrieval))}}{nb(\texttt{Information} \cap \texttt{Retrieval})}$ [formule (1.20)]
- $DeT^{Near}_{IM}$ : $\frac{nb(\texttt{IR NEAR (Information} \cap \texttt{Retrieval))}}{nb(\texttt{Information} \cap \texttt{Retrieval})}$ [formule (1.21)]
- $DeT^{And}_{IM3}$ : $\frac{nb(\texttt{IR AND (Information} \cap \texttt{Retrieval))}^3}{nb(\texttt{Information} \cap \texttt{Retrieval})}$ [formule (1.22)]
- $DeT^{Near}_{IM3}$ : $\frac{nb(\texttt{IR NEAR (Information} \cap \texttt{Retrieval))}^3}{nb(\texttt{Information} \cap \texttt{Retrieval})}$ [formule (1.23)]

Bien sûr nous pouvons ajouter un contexte à ces mesures, dans ce cas ces mesures seront nommées $DeMTC$ et $DeTC$.

La section suivante présente les expérimentations menées pour la tâche de désambiguisation des acronymes/définitions avec les différentes mesures présentées dans cette section.

#### 1.2.2.4 Expérimentations

Dans ces expérimentations, nous avons utilisé le moteur de recherche Exalead car il a un bon comportement pour le français et l'anglais. Une autre raison majeure d'utiliser ce moteur de recherche réside dans le fait qu'il propose l'opérateur *NEAR* (avec une fenêtre de 16 mots [11]) qui est utile pour les formules (1.25), (1.21), (1.23). Les travaux présentés dans ce mémoire ont été initiés en 2007 bien avant l'apparition de l'opérateur du même type par google fin 2010 (opérateur $AROUND$).

##### Corpus en français – Domaine général

Dans ces expérimentations, nous nous sommes appuyés sur l'acronyme polysémique "JO". Notons que ce choix a été motivé par le fait que les premières pages retournées par le moteur de recherche Google sont réparties de manière semblable [12].

Dans ces expérimentations, nous avons utilisé un corpus provenant du défi DEFT'06 (DÉfi Fouille de Textes). La deuxième édition de ce défi francophone de fouille de textes consistait à déterminer les segments thématiques de corpus écrits en français issus de domaines différents (politiques, juridiques, scientifiques). Dans nos expérimentations, nous nous sommes appuyés sur le corpus juridique propre à des articles de loi de l'Union

11. Informations sur l'opérateur NEAR du moteur de recherche Exalead : (1) http ://www.searchengineshowdown.com/blog/exalead/, (2) http ://moritzlegalinformation.blogspot.com/2006_06_01_archive.html

12. Expériences effectuées avec les 50 premières pages retournées en février 2007 par le moteur de recherche google avec la requête "JO". Manuellement, nous avons évalué le fait que 10 pages sont propres au "Journal Officiel" et 10 pages sont relatives aux "Jeux Olympiques".

Européenne [13]. Les 1303 articles (11 Mo) possédant l'acronyme JO sont pris en compte.

Cet acronyme est généralement utilisé dans ce corpus pour faire référence à un ou des articles précis du Journal Officiel (par exemple, les références "JO 308 du 18.12.1967" ou "JO no L 249 du 8.9.1988" pour lesquelles l'acronyme JO n'est pas défini). Pour chacun des articles de loi, nous mesurons si l'acronyme JO doit être associé à la définition "Journal Officiel". Le tableau 1.2 présente les taux d'erreur obtenus à partir de ce corpus avec différents contextes (de un à trois mots). Notons que dans ces expérimentations, il est nécessaire d'exécuter 23454 requêtes : 1303 articles de loi et 6 requêtes par article avec 3 jeux de test (contextes de un à trois mots).

| | Nombre d'acronymes correctement associés | Taux d'erreur |
|---|---|---|
| | **Contexte d'1 mot** | |
| $DeMTC_{IM}$ | 190 | 85.4% |
| $DeMTC_{IM3}$ | 1040 | 20.2% |
| $DeMTC_{Dice}$ | 842 | 35.4% |
| | **Contexte de 2 mots** | |
| $DeMTC_{IM}$ | 434 | 66.7% |
| $DeMTC_{IM3}$ | 1234 | 5.3% |
| $DeMTC_{Dice}$ | 1200 | 7.9% |
| | **Contexte de 3 mots** | |
| $DeMTC_{IM}$ | 650 | 50.1% |
| $DeMTC_{IM3}$ | 1281 | 1.7% |
| $DeMTC_{Dice}$ | 1274 | 2.2% |

TABLE 1.2 – Taux d'erreurs sur le corpus juridique de DEFT'06.

Le tableau 1.2 montre que notre méthode donne des résultats de très bonne qualité avec le corpus de DEFT'06, plus particulièrement dans le cas d'un contexte plus riche c'est-à-dire constitué de deux ou trois mots. Nous pouvons confirmer que le coefficient de Dice et l'Information Mutuelle au Cube donnent un taux d'erreur faible respectivement de 2.2% et 1.7% avec un contexte de trois mots.

Le fait que l'Information Mutuelle au Cube et le coefficient de Dice donnent des résultats de bonne qualité à partir des deux corpus étudiés s'explique par le fait ces mesures privilégient les co-occurrences fréquentes. Dans notre cas, le nombre de pages Web partageant la définition d'un acronyme associée à un contexte pertinent est important. Ceci a pour conséquence d'accorder un score élevé à ces mesures qui sont relatives à un grand nombre de pages. Notons que dans le cas où le nombre de pages retournées pour différentes définitions est du même ordre, la fréquence n'est pas toujours un critère pertinent, d'où la nécessité de s'appuyer sur des mesures statistiques. Ceci sera d'ailleurs confirmé dans la section 3.4.3.1 du chapitre 3.

#### Corpus en anglais – Domaine Spécialisé (Biomédical)

Dans cette section, nous nous sommes intéressés à la désambibiguisation de défini-

13. Corpus disponible à l'adresse suivante : http ://www.lri.fr/ia/fdt/DEFT06/corpus/donnees.html

tions du domaine biomédical [20] fournies par l'application Acromine [14]. Pour chaque acronyme donné, Acromine fournit une liste d'expansions possibles. Ainsi, 102 paires acronymes/définitions ont été aléatoirement extraites à partir d'Acromine. Pour chaque acronyme, 4 à 6 définitions possibles sont proposées. Les acronymes étudiés ont une taille de deux, trois ou quatre caractères. Par exemple, le tableau 1.3 présente les définitions possibles pour l'acronyme PKD.

| |
|---|
| polycystic kidney disease |
| protein kinase D |
| proliferative kidney disease |
| paroxysmal kinesigenic dyskinesia |
| pyruvate kinase deficiency |

TABLE 1.3 – Exemple de définitions possibles de l'acronyme PKD en biomédecine.

Pour chaque paire, des résumés d'articles ont été extraits de la base de données bibliographique Medline [15]. Ces résumés contiennent les acronymes et leurs expansions. Nous avons alors manuellement constitué un ensemble de 204 documents (deux documents par couple acronyme/expansion). Le but de ces expérimentations est de déterminer, pour chaque document, si la définition prédite par nos mesures correspond à l'expansion réelle.

La distribution des acronymes par rapport aux 204 documents est donnée dans le tableau 1.4. Ce dernier montre que nous avons besoin de tester 960 expansions ($12 \times 6 + 120 \times 5 + 72 \times 4$).

| **Nb de documents** | **Nb d'expansions possibles par document** |
|---|---|
| 12 | 6 |
| 120 | 5 |
| 72 | 4 |

TABLE 1.4 – Nombre de définitions possibles pour les 204 documents.

Les expérimentations ont nécessité l'exécution de 7340 requêtes [16] :
- Calcul de 6 mesures $DeT$ : $DeT$ nécessite $2 \times 960$ requêtes pour le numérateur (avec les opérateurs $AND$ et $NEAR$) et $2 \times 960$ pour le dénominateur (pour la mesure de Dice) : 3840 requêtes
- Calcul de 3 mesures $DeMT$ : $DeMT$ nécessite 960 requêtes pour le numérateur et 2540 pour le dénominateur (le nombre de requêtes du dénominateur dépend du nombre de mots de chaque expansion) : 3500 requêtes

**Résulats des mesures $DeMTC$ et $DeTC$**

14. http ://www.nactem.ac.uk/software/acromine/
15. http ://www.ncbi.nlm.nih.gov/PubMed/
16. Expérimentation réalisées en août 2009.

Le tableau 1.5 présente les résultats expérimentaux obtenus pour les mesures $DeMTC$ et $DeTC$.

- La première colonne correspond au nombre de fois où la définition correcte est retournée en première position
- La deuxième colonne correspond au nombre de fois où la définition correcte est retournée en première ou deuxième position.
- La troisième colonne correspond au nombre de fois où la définition correcte est retournée à une des trois premières positions.

| **Rang** | **1** | **1 or 2** | **1, 2, or 3** |
|---|---|---|---|
| $DeMTC_{Dice}$ | 73 (35.8%) | 127 (62.3%) | 161 (78.9%) |
| $DeMTC_{IM}$ | 62 (30.4%) | 111 (54.4%) | 149 (73.0%) |
| $DeMTC_{IM3}$ | 72 (35.3%) | 118 (57.8%) | 165 (80.9%) |
| $DeTC_{Dice}^{And}$ | 111 (**54.4%**) | 150 (**73.5%**) | 174 (**85.3%**) |
| $DeTC_{Dice}^{Near}$ | 104 (51.0%) | 142 (69.6%) | 174 (85.3%) |
| $DeTC_{IM}^{And}$ | 94 (46.1%) | 139 (68.1%) | 169 (82.8%) |
| $DeTC_{IM}^{Near}$ | 90 (44.1%) | 137 (67.1%) | 170 (83.3%) |
| $DeTC_{IM3}^{And}$ | 104 (51.0%) | 145 (71.1%) | 174 (85.3%) |
| $DeTC_{IM3}^{Near}$ | 102 (50.0%) | 146 (71.6%) | 170 (83.3%) |

TABLE 1.5 – Nombre de définitions correctement prédites à partir des résumés de la base bibliographique PubMed/Medline

| **Mesures** | **Somme** |
|---|---|
| $DeMTC_{Dice}$ | 470 |
| $DeMTC_{IM}$ | 516 |
| $DeMTC_{IM3}$ | 481 |
| $DeTC_{Dice}^{And}$ | **389** |
| $DeTC_{Dice}^{Near}$ | 403 |
| $DeTC_{IM}^{And}$ | 422 |
| $DeTC_{IM}^{Near}$ | 424 |
| $DeTC_{IM3}^{And}$ | 401 |
| $DeTC_{IM3}^{Near}$ | 405 |

TABLE 1.6 – Somme des rangs des définitions pertinentes.

Les expérimentations ont été menées avec un contexte formé d'un seul mot (mot le plus fréquent de chaque document). Le fait que cette étude ait été menée sur un domaine de spécialité implique que des requêtes avec plus de mots retournent souvent des valeurs nulles lorque nous utilisons un moteur de recherche généraliste tel qu'Exalead.

Le tableau 1.5 nous permet de répondre à un certain nombre de questions et d'établir des remarques résumées ci-dessous :

**Quelle mesure de qualité adopter pour la désambiguisation d'acronymes / définitions ?**

Le tableau 1.5 montre que la mesure $DeTC$ a le meilleur comportement. Ceci montre

que pour la tâche de désambiguisation de définitions, il est préférable de calculer la dépendance entre acronyme et expansion plutôt que la dépendance entre les mots des définitions. Notons que pour valider les termes en eux même (cf. chapitres 2 et 3) seule $DeMTC$ peut être appliquée.

Par ailleurs, nous remarquons que les mesures $IM3$ et Dice ont un bon comportement avec un résultat légèrement meilleur obtenu avec la mesure de Dice. Nos mesures permettent donc d'étendre la mesure de qualité proposée par [Turney, 2001] qui s'appuie sur l'Information Mutuelle.

La table 1.5 montre que les performances obtenues avec les opérateurs $AND$ and $NEAR$ sont en fait très proches. Ce résultat diffère des travaux présentés par [Turney, 2001]. Ceci peut s'expliquer par la spécificité de l'utilisation des acronymes dont l'expansion est souvent assez proche dans les documents retournés par les moteurs de recherche.

Le tableau 1.5 montre que $DeTC^{And}_{Dice}$ fournit la définition pertinente au rang 1 dans 54.4% des cas. Ceci est significatif au regard d'une prédiction aléatoire (score de 22% dans une telle situation [17]).

Dans le but de déterminer de manière plus précise la qualité de ces mesures, nous avons calculé la somme des rangs des définitions pertinentes. La meilleure mesure correspond à celle qui retourne la somme la plus faible. Une telle méthode est équivalente aux approches fondées sur les courbes ROC (Receiver Operating Characteristics) et au calcul de l'aire sous ces dernières ([Ferri et al., 2002], [23]).

La Table 1.6 confirme que $DeTC^{And}_{Dice}$ est la mesure qui a le meilleur comportement dans le cadre du traitement de documents spécialisés en biomédecine. De manière globale les mesures $DeTC$ obtiennent un meilleur rang (plus faible somme) que la meilleure mesure $DeMTC$. Par ailleurs, ce tableau montre que la mesure de Dice améliore aussi bien les mesures $DeTC$ que $DeMTC$. Notons cependant que les mesures de Dice et l'Information Mutuelle au Cube donnent des résultats assez proches.

#### Propriété des données

Dans le domaine biomédical, le fait que plusieurs termes puissent être associés à un même concept ou un concept proche est une problématique assez classique. Par exemple, à l'acronyme ZO nous pouvons associer les définitions suivantes : *zonula occludens*, *zona occludens*, *zonulae occludentes*. Avec un tel exemple, nous pouvons remarquer les phénomènes de flexions pour un même terme (singulier *vs.* pluriel) ou des termes lexicalement très proches (*zonula* signifie "small zone" *vs. zona*). Ces variations s'expliquent par des fonctions linguistiques classiques. Ainsi, quelques erreurs de prédiction peuvent être causées par de telles variations linguistiques qui font en fait référence à un même concept.

Par ailleurs, quelques définitions équivalentes pourraient aussi être identifiées avec l'aide d'un expert du domaine. Par exemple, *terminal* et *termini* peuvent être vus comme

17. Ce score alétaoire est calculé de la manière suivante : 1 chance sur 4 d'avoir la définition pertinente à la première position dans 72 cas, 1 chance sur 5 dans 120 cas et 1 chance sur 6 dans 12 cas, ceci correspondant au nombre de documents avec respectivement 4, 5 et 6 définitions possibles (cf. Table 1.4).

des flexions latines dans les couples suivants : *carboxy terminal* / *carboxy termini*, *COOH terminal* / *COOH termini* et *C02H-terminal* / *CO2H termini*. De plus, il est crucial que l'expert relève que *COOH*, *CO2H* et *carboxy* sont également des formes équivalentes. Ainsi, cette expertise pourra aider à améliorer les résultat de notre système en identifiant que tous ces termes sont totalement équivalents.

Les fonctions de rang proposées se révèlent tout à fait adaptées pour une tâche de désambiguisation. Dans cette section, nous nous sommes intéressés à l'identification des liens de synonymie entre les acronymes et les définitions. Dans de nombreuses autres situations, il est nécessaire d'identifier des liens de synonymie ou, de manière plus globale, des proximités sémantiques entre mots et/ou syntagmes. Ceci est le cas, dans le domaine de la fouille de données d'opinion qui consiste, par exemple, à construire des dictionnaires relatifs à un sentiment. Une telle tâche est détaillée dans la section suivante.

### 1.2.3 Fonctions de rang et construction d'un dictionnaire d'opinion

Dans cette section, nous décrivons un processus global pour construire un dictionnaire d'opinion. Comme nous le montrerons dans la section suivante, dans le cadre de ce processus global, nous avons utilisé et adapté la mesure $DeMTC$.

Ce processus s'appuie sur différents traitements de fouille de textes menés en collaboration avec l'équipe Tatoo du LIRMM et le LGI2P (Nîmes). Ces traitements sont résumés dans la section 1.2.3.1 et détaillés dans [10]. La fouille de données d'opinion et de sentiment est en essor constant ces dernières années comme le démontrent les éditions récentes dans le domaine [Jackiewicz et al., 2010] (Membre du comité éditorial) et [25] (Co-éditeur avec Pascal Poncelet).

#### 1.2.3.1 Processus de fouille de données d'opinion

Avec le développement du Web, et surtout du Web 2.0, le nombre de documents décrivant des opinions devient de plus en plus important. Il devient ainsi possible de donner son avis sur un produit ou sur un film. Récemment, les chercheurs de différentes communautés (fouille de données, fouille de textes, linguistique, etc) se sont intéressés à l'extraction automatique de données d'opinions sur le Web. Traditionnellement, les approches de détection d'opinions cherchent à déterminer les caractéristiques d'opinions positives ou négatives à partir d'ensembles d'apprentissage. Des algorithmes de classification sont alors utilisés pour classer automatiquement les documents extraits du Web. Dans cette section, nous nous intéressons plus particulièrement à l'étape d'acquisition du vocabulaire caractérisant une opinion positive ou négative d'un document. De manière à caractériser ces dernières, les principaux travaux de recherche considèrent que l'orientation sémantique d'une opinion est exprimée par l'intermédiaire des adjectifs [Turney, 2002, Taboada et al., 2006, Kamps et al., 2004] bien que les verbes puissent également véhiculer un sentiment [Sokolova and Lapalme, 2008]. Des approches ont enrichi l'apprentissage des adjectifs à l'aide de ressources existantes, par exemple WordNet [Miller, 1995]. Dans ce cadre, il s'agit d'intégrer automatiquement les synonymes et les antonymes [Andreevskaia and Bergler, 2006] ou d'acquérir des mots porteurs d'opinions [Voll

and Taboada, 2007, Hu and Liu, 2004]. Ces dictionnaires représentent la base essentielle pour déterminer la polarité générale véhiculée par un document [Taboada et al., 2011]. Notons que des traitements complémentaires, comme la prise en compte de la négation pour le changement de polarité, sont souvent déterminants [Wiegand et al., 2010, Taboada et al., 2011].

Cependant, la plupart des approches qui s'appuient sur des dictionnaires existants ou sur des listes prédéfinies d'adjectifs se trouvent confrontées au problème suivant. Considérons, les deux phrases "*The picture quality of this camera is high*" et "*The ceilings of the building are high*". Dans le cas de la première phrase (par exemple, une opinion exprimée sur une caractéristique d'un produit), l'adjectif *high* est positif. Par contre dans la seconde phrase (par exemple, un document sur l'architecture), l'adjectif est neutre. Notre objectif est de proposer une méthode de détection automatique des adjectifs correspondant à une opinion exprimée sur un domaine spécifique. De la même manière, [Sokolova and Lapalme, 2011] focalisent leur étude sur les mots dits "non émotionnels" (comme le mot *high* de notre exemple) qui sont malgré tout porteurs d'une opinion. Ces travaux sont menés dans un cadre d'apprentissage supervisé contrairement à notre approche décrite dans [10] et synthétisée ci-après.

- **Phase 1 : Acquisition de données sources**
  Pour construire un dictionnaire d'opinion, la premiere étape consiste à acquérir de manière automatique un corpus adapté. Pour cela, nous considérons deux ensembles $P$ et $N$ de mots "germes" dont les orientations sémantiques sont respectivement positif et négatif [Turney, 2002].
  - P = $\{good, nice, excellent, positive, fortunate, correct, superior\}$
  - N = $\{bad, nasty, poor, negative, unfortunate, wrong, inferior\}$

  Pour chaque mot germe, nous utilisons un moteur de recherche avec une requête spécifiant un domaine d'application $d$, le mot germe recherché et les mots à éviter. Par exemple, si nous considérons le moteur de recherche google, pour obtenir des opinions sur des films avec le mot germe "good", la requête suivante est effectuée : "+opinion +review +cinema +good -bad -nasty -poor -negative -unfortunate -wrong -inferior". Cette requête donnera comme résultat des documents d'opinions sur le cinéma contenant le mot good mais ne contenant pas les mots bad, nasty, poor, ... inferior. De manière générale, les requêtes propres aux opinions positives (resp. négatives) d'un domaine $d$ sont de la forme : "+opinion +mots$_d$ +germes$_{pos}$ -germes$_{neg}$. Dans le cadre de nos expérimentations, nous avons utilisé le moteur de recherche BlogGooglesearch.com spécialisé dans la recherche sur les blogs pour obtenir des données sources. Ainsi, pour chaque mot germe de l'ensemble P (resp. N) et pour un domaine donné, nous collectons automatiquement K documents où il n'apparaît aucun mot de l'ensemble N (resp. P). Nous obtenons ainsi, 14 corpus : 7 positifs et 7 négatifs.

- **Phase 2 : Extraction des adjectifs porteurs d'opinion**
  Les corpus obtenus lors de l'étape précédente ne contiennent que des documents correspondant à un domaine spécifique. L'objectif de la seconde phase est de rechercher dans ces corpus les adjectifs porteurs d'opinion. Pour cela, à partir des corpus

collectés, nous cherchons les corrélations entre les mots germes et les adjectifs des documents collectés pour enrichir les ensembles de mots germes avec des adjectifs pertinents. Pour cela, les deux étapes successives ci-dessous sont appliquées.

1. ***Prétraitement et règles d'association***
   Comme dans [Taboada et al., 2006, Kamps et al., 2004], nous considérons les adjectifs comme des mots représentatifs pour déterminer l'opinion. Ainsi, après l'application d'un étiqueteur grammatical [Schmid, 1994], nous ne conservons que les adjectifs des documents traités. Nous déterminons l'association sémantique entre les termes des documents et les mots germes des ensembles positifs et négatifs à l'aide d'un algorithme de recherche de règles d'association de type "Apriori" [Agrawal and Srikant, 1994]. Soit $I = \{i_1, ....i_n\}$ un ensemble d'items, et $D$ un ensemble de transactions, où chaque transaction correspond à un sous-ensemble d'éléments de $I$. Une règle d'association est une implication de la forme X$\rightarrow$Y, où X$\subset I$, Y$\subset I$, et X $\cap$ Y $= \oslash$. Une règle a un support $s$ si $s$% des transactions de $D$ contiennent X$\cup$Y. Dans notre contexte, les items correspondent aux adjectifs et les transactions aux phrases.

2. ***Filtrage***
   De manière à minimiser le nombre de faux positifs et de faux négatifs, les adjectifs trouvés dans les documents qui sont en corrélation avec un seul mot germe sont supprimées. En effet, comme nous ne voulons pas utiliser de dictionnaire extérieur, en ne retenant que des adjectifs corrélés à plus d'un mot germe, nous souhaitons véritablement rechercher ceux qui sont fortement corrélés à des opinions positives ou négatives. Ensuite, pour les adjectifs qui apparaissent à la fois dans les listes positives et négatives, ceux qui sont corrélés avec plusieurs mots germes d'une même orientation ayant un support élevé et une moyenne d'apparition plus grande que 1, sont retenus comme mots appris de la part de ces mots germes, autrement ils sont éliminés.

   Enfin, de manière à améliorer la qualité des résultats obtenus, nous appliquons la mesure $DeMTC_{IM3}$. Cette approche consiste à mesurer la dépendance entre un adjectif de la liste des adjectifs appris par rapport à la liste des mots germes. Ce point sera détaillé en section 1.2.3.2 de ce manuscrit.

- **Phase 3 : Classification**
  Pour chaque document à classer, nous calculons son orientation positive ou négative en fonction du nombre d'adjectifs, appris dans la phase précédente, contenus dans le document. Nous comptons le nombre d'adjectifs positifs, puis le nombre d'adjectifs négatifs, et nous faisons la différence. Si le résultat est positif (resp. négatif), le document sera classé dans la classe positive (resp. négative).

#### 1.2.3.2 Sélection du vocabulaire d'opinion par $DeMTC$

De manière à améliorer la qualité des résultats obtenus de l'étape de filtrage de la phase 2, nous appliquons la mesure $DeMTC_{IM3}$. Dans nos travaux, nous souhaitons

calculer la dépendance entre deux adjectifs $x$, $y$. Dans ce cas, $nb(x, y)$ représente le nombre de pages Web où $x$ et $y$ sont présents ensemble et de manière consécutive. Le numérateur de la mesure de base $DeMTC_{IM3}$ calcule le nombre de pages où le mot $x$ précède le mot $y$ dans un contexte $C$. Dans ces travaux de fouille de données d'opinion, nous allons aussi considérer la situation symétrique, c'est-à-dire lorsque le mot $y$ précède le mot $x$. En effet, dans le cadre des termes étudiés précédemment (cas des définitions des acronymes), l'ordre des mots d'un terme a une importance majeure en faisant référence à des concepts différents. Par exemple le terme "ciel bleu" renvoie à un concept qui diffère de la couleur "bleu ciel", ce dernier terme correspondant à une locution adjectivale.

Dans les travaux présentés ici, nous étudions si deux adjectifs $x$ et $y$ peuvent être voisins en prenant en considération les situations symétriques (par exemple, lors des énumérations d'adjectifs). Ainsi, dans la pratique, $nb(x, y)$ correspond au nombre total de pages retournées par un moteur de recherche avec les deux requêtes "$x$ $y$" et "$y$ $x$" (somme totale du nombre de pages retournées par ces deux requêtes).

**Exemple :** Considérons l'adjectif *funny* qui doit être associé à une orientation sémantique positive. Nous allons chercher la dépendance entre l'adjectif *funny* et les mots germes positifs cités dans la section 1.2.3.1 pour le contexte *movie*. En appliquant $DeMTC$ avec l'adjectif germe *good* relativement à l'adjectif *funny*, nous obtenons :

$$DeMTC_{IM3}(\texttt{funny}) = \frac{(nb((\texttt{good} \ \cap \ \texttt{funny}) + \ \texttt{movie}) + nb((\texttt{funny} \ \cap \ \texttt{good}) + \ \texttt{movie}))^3}{nb(\texttt{good} + \ \texttt{movie}) \times nb(\texttt{funny} + \ \texttt{movie})}$$

Cette mesure est appliquée en utilisant le moteur de recherche Google. La même formule est appliquée entre l'adjectif *funny* et tous les mots germes (voir figure 1.2) (un logarithme peut également être appliqué afin de "lisser" les valeurs obtenues). Ensuite, la moyenne de toutes les valeurs obtenues (*17.37*) est calculée (cf. figure 1.2). Puisque, cette valeur est supérieure à un seuil déterminé expérimentalement (0.05), l'adjectif *funny* sera retenu comme adjectif positif appris [10].

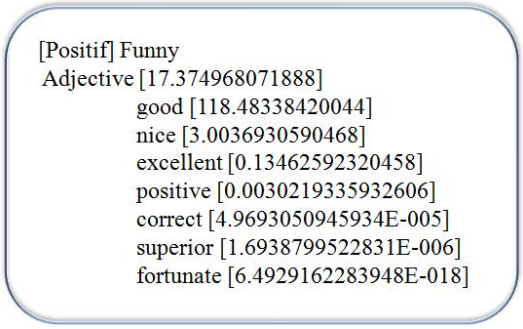


FIGURE 1.2 – Exemple du mot *funny* en appliquant $DeMT_{IM3}$

Par ailleurs d'autres filtres sont également appliqués afin d'éliminer les adjectifs qui sont souvent communs aux deux corpus.

L'utilisation de la mesure *DeMTC* dans notre approche permet d'améliorer significativement la détection d'un texte véhiculant une opinion positive (66.9% *vs.* 75.9%) et négative (30.4% *vs.* 57.1%). Les détails du processus appliqué et les expérimentations complètes sont décrites dans [10].

## 1.3 Discussion générale

Le travail mené ici, montre que les informations endogènes permettent de déterminer des connaissances sémantiques cruciales (section 1.1). Par ailleurs, les connaissances exogènes sont essentielles pour améliorer les approches de fouille de textes (section 1.2).

Les connaissances exogènes traitées dans ce mémoire sont focalisées sur les informations issues du Web. Pour mener à bien des processus complets de fouille de textes (dans notre cas, la désambiguisation des acronymes/définitions et la fouille de données d'opinion), des informations complémentaires liées aux corpus sont souvent nécessaires. Le chapitre suivant place quant à lui le corpus au cœur du processus afin d'identifier les descripteurs pertinents.

De manière plus globale la mise en relief des descripteurs linguistiques en contexte permet une meilleure interprétation des résultats. Par exemple, dans les travaux menés en collaboration avec l'équipe Tatoo et l'entreprise Pikko [18], nous avons développé une méthode permettant d'extraire des connaissances dans des bases de données (sous forme de motifs séquentiels) et visualiser les résultats en contexte (via les publications scientifiques). Cette visualisation offre une assistance à la découverte de nouveautés en facilitant l'accès aux documents ayant un lien plus ou moins fort avec les gènes examinés [30]. Ceci met en exergue l'importance du contexte et de sa prise en compte globale en fouille de textes.

Afin de prendre en considération les informations contextuelles de manière plus significative, dans [28] nous avons enrichi la méthode *DeMT*, en prenant en compte la proximité des contextes textuels comme dans les travaux de désambiguisation lexicale [Navigli, 2009]. Dans le chapitre suivant, nous adoptons ce même point du vue en traitant les descripteurs linguistiques (mots, syntagmes) en corpus.

18. http ://www.pikko-software.com/

# Chapitre 2

# Extraction et traitement des descripteurs linguistiques en corpus

| *Thèmes de Recherche* | *Types de travaux* | *Années* |
|---|---|---|
| Extraction (80%) | | Exploitation (20%) |
| Extraction de connaissances en Français Médiéval | **Projet STICS/UM2 – *TSAL***<br>porteur, co-responsable scientifique<br>*Stage Ingénieur CNAM* | 2006-2007 |
| Extraction (70%) | | Exploitation (30%) |
| Classification de Blogs | **Collaboration industrielle** – PAPERBLOG (Paris)<br>responsable scientifique<br>*Stage Ingénieur* | 2007-2008 |
| Classification de Tweets | **Collaboration industrielle** – WEBREPORT (Bordeaux)<br>co-responsable scientifique<br>*Stage M2 Recherche* | 2010-2011 |
| Recherche de gloses | **Projet CNRS PEPS – *RESENS***<br>porteur, co-responsable scientifique | 2010-2011 |
| Extraction (50%) | | Exploitation (50%) |
| Extraction d'information dans les fichiers logs | **THÈSE** CIFRE – H. Saneifar (en collaboration avec *Satin Technologies*) | 2008-2011 |
| Titrage automatique | **Collaboration industrielle** – Open-S/EvalAccess (Montpellier)<br>co-responsable scientifique<br>*Stages M2 Recherche*<br>**THÈSE** (Région/UM2) – C. Lopez | 2008-2009<br><br>2009-2012 |

Après avoir décrit les méthodes de traitement des mots/termes en eux mêmes, nous allons étudier ces descripteurs linguistiques en contexte, c'est-à-dire en corpus. Dans un premier temps, nous nous sommes intéressés à l'extraction puis à l'utilisation du mot comme descripteur pour des tâches de classification de textes (cf. section 2.1). Dans un second temps, les descripteurs présents sous forme syntagmatique sont présentés (cf. section 2.2). Les syntagmes (cf. section 2.2.1) sont utiles pour le traitement de données textuelles "classiques" pour des tâches de titrage par exemple (cf. section 2.2.2). Nous monterons également que leur extraction se révèle cruciale à partir de données textuelles atypiques tels que les textes en français médiéval et les fichiers logs (cf. section 2.2.3).

## 2.1 Les mots

Dans cette section, nous nous intéresserons aux mots utilisés comme descripteurs pour des tâches de classification de textes (section 2.1.1) et nous discuterons du comportement de ces mots dans le cas d'une application de détection de catastrophes naturelles à partir de dépêches (section 2.1.2).

### 2.1.1 Le mot en corpus, un descripteur linguistique de base en fouille de textes

Dans cette section, nous présentons un système de Recherche d'Information qui s'appuie sur une représentation vectorielle des textes. Ceci permet, en particulier, d'effectuer des tâches de classification [Scott and Matwin, 1999, Sebastiani, 2002] à partir des données textuelles issues du Web 2.0. Ce point sera détaillé en section 2.1.2.

Une telle approche "sac de mots" a également été utilisée dans le cadre de collaborations menées avec le LIA (Avignon) [11]. Ce travail a consisté à classer les candidatures les plus adaptées (CV et lettres de motivation) au regard d'une offre d'emploi décrite en langage naturel. Dans ce mémoire, nous allons nous focaliser sur une telle approche appliquée dans le cadre des réseaux sociaux.

L'approche classique de Recherche d'Information est fondée sur une représentation vectorielle [Salton et al., 1975]. Cette approche dite *sac de mots* est relativement efficace pour les problèmes de classification à grande échelle de documents. Les limites de ces méthodes résident dans le fait que l'ordre des mots, et donc les informations syntaxiques inhérentes, n'est pas considéré. Ce point sera discuté en section 2.2. Par ailleurs, d'autres approches ne s'intéressent pas nécessairement à la prise en compte de l'ensemble des informations textuelles d'un document mais préfèrent se focaliser sur les phrases véhiculant les informations les plus significatives afin de catégoriser un document [Dulac-Arnold et al., 2011]. Bien que les améliorations de ce type soient tout à fait pertinentes, les approches de classification de base s'appuient sur le modèle suivant les trois étapes décrites ci-après :

- **Étape 1 : Acquisition d'un corpus**

  La première étape est une phase d'acquisition de données textuelles afin d'obtenir un corpus homogène sur la forme. L'acquisition du corpus est en général dépendante des tâches à mener et peut être très spécifique, comme par exemple la constitution de corpus multilingues propres à une même actualité [Pattabhi et al., 2010] ou l'acquisition d'un corpus de SMS via une collecte sous forme de dons [Fairon and Paumier, 2006]. Dans les travaux menés au LIRMM depuis 2005, selon les tâches de classification, des corpus plus ou moins spécifiques seront utilisés. Bien qu'étant souvent différents en termes de langue ou de thème, le point commun de ces corpus tient au fait qu'ils nécessitent une phase de normalisation.

- **Étape 2 : Représentation du corpus**

  Une fois le corpus acquis et normalisé, ce dernier peut alors être représenté sous forme vectorielle. Chaque texte sera considéré comme un *sac de mots*. Dans cette représentation dite "Saltonienne" [Salton et al., 1975], un traitement préalable peut consister à éliminer les mots outils ou fonctionnels (préposition, articles, etc).

Nous pouvons également appliquer des traitements statistiques et/ou linguistiques décrits ci-après.

***Traitement statistique***

Deux types de représentations statistiques simples peuvent alors être effectuées : une représentation booléenne (présence/absence des mots) et/ou fréquentielle (nombre d'occurrences des mots dans chaque document).

Notons que d'autres types de représentations numériques peuvent être appliquées. Par exemple, la mesure TF-IDF [Salton and Buckley, 1988] consiste à calculer l'importance d'un mot dans un document relativement à une collection. Ainsi, un mot présent dans tous les documents d'une collection aura un poids moindre. De manière plus concrète, le TF-IDF est calculé de la manière suivante.
La fréquence d'un mot (term frequency ou TF) est d'abord calculée. Elle correspond au nombre d'occurrences (normalisé) de ce terme dans le document considéré (formule (2.1)). Cette fréquence peut déterminer l'"importance" et/ou la "représentativité" d'un mot dans un texte.

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2.1}$$

– $n_{i,j}$ est le nombre d'occurrences du terme $t_i$ dans le document $d_j$, et le dénominateur est la somme du nombre d'occurrences pour chaque terme dans le document $d_j$.

La fréquence inverse de document (inverse document frequency ou IDF) permet de donner un poids plus important aux mots les plus discriminants, c'est-à-dire n'apparaissant pas dans de nombreux documents du corpus (formule (2.2)).

$$IDF_i = log\frac{|D|}{|d_j : t_i \in d_j|} \tag{2.2}$$

– $|D|$ : nombre total de documents dans le corpus.
– $|d_j : t_i \in d_j|$ : nombre de documents où le terme $t_i$ apparaît.

Enfin, le produit (formule (2.3)) de ces deux critères est calculé afin d'obtenir une valeur globale :

$$TF - IDF_{i,j} = TF_{i,j} \times IDF_i \tag{2.3}$$

D'autres types de pondérations existent comme la mesure Okapi très répandue dans le domaine [Robertson et al., 1994]. Un large panel de pondérations est donné dans [Turney and Pantel, 2010].

Notons que pour éviter d'obtenir des vecteurs trop creux, une phase d'élagage consistant à ne prendre en compte que les mots présents un minimum de fois peut être appliquée [23].

***Traitement linguistique***

Par ailleurs, différents traitements linguistiques peuvent également être appliqués. Dans la suite, nous allons illustrer ces traitements à partir de l'exemple ci-après : *Il vend son bien de Montpellier.*

Un premier traitement consiste à déterminer les informations morphosyntaxiques des mots. Ceci a deux avantages. Le premier est de conserver les mots ayant une étiquette grammaticale donnée qui sont les plus porteurs d'information pour des tâches de classification. Il est courant de conserver les noms, verbes, adjectifs et éventuellement leur donner des poids différents [1]. Dans notre exemple, nous conserverons les mots *vend*, *bien* et *Montpellier*. Le deuxième avantage est de distinguer certains mots ayant différentes catégories grammaticales possibles avec une sémantique qui leur est propre. Ce traitement permet alors d'associer le mot *bien* à un nom et non pas à un adverbe ou un adjectif. Pour ce traitement, nous pouvons utiliser différents étiqueteurs grammaticaux qui sont détaillés dans la section 2.2.3 de ce chapitre.

Par la suite, les mots d'une même famille présents sous forme fléchie peuvent être rassemblés (singuliers/pluriels, féminins/masculins, verbes conjugués/infinitif, etc). La représentation canonique des mots permet un regroupement de ces derniers. Ainsi, les formes lemmatisées des mots sélectionnés à l'étape précédente sont *vendre*, *bien* et *Montpellier*. D'autres types de préparation des données textuelles consistent à ne considérer que la forme radicale de chaque mot [Porter, 1980].

Outre les informations grammaticales et lexicales, des connaissances sémantiques [Miller, 1995] peuvent aussi être utilisées (par exemple, en remplaçant lorsque cela se révèle possible, chaque mot par son concept plus général). Dans l'exemple précédent, nous pourrions alors identifier que le verbe *vendre* est en général lié à un concept de Commerce. Nous pouvons aussi repérer que *Montpellier* est une Entité Nommée d'un lieu (cf. section 3.2 du chapitre 3 pour plus de détails sur les Entités Nommées). Notons cependant que l'utilisation des information sémantiques n'a pas toujours un impact positif en classification [Scott and Matwin, 1999].

D'autres traitements linguistiques plus complexes peuvent aussi être menés à l'échelle du corpus. En effet, nous pouvons supposer que le pronom (*il*) de l'exemple est en relation anaphorique [Orasan et al., 2008], il renvoie à une entité qui précède. Le remplacement du pronom par son antécédent nécessitant la résolution des anaphores [Weissenbacher and Nazarenko, 2007] permet d'avoir des informations plus précises dans la représentation des données textuelles.

Notons enfin, qu'après ces traitements, les mots, lemmes ou radicaux seront appelés les *descripteurs linguistiques* des documents.

- **Étape 3 : Classification d'un nouveau document**

  Dans la dernière phase utile pour les tâches de classification, nous calculons la similarité entre un nouveau document et les vecteurs existants. Ces derniers peuvent être associés à une étiquette par exemple relative à un thème. Ainsi, lorsqu'un nouveau document partage souvent les mêmes descripteurs (par exemple, les mots), les textes peuvent être déclarés comme proches. Pour calculer cette similarité, des mesures telles que le cosinus (décrit dans la section 1.1 du chapitre 1) peuvent être appliquées. Les différents algorithmes de classification de la littérature [Sebastiani, 2002] s'appuient sur le principe général énoncé ci-dessus.

  Depuis 2005, j'ai mené de nombreux travaux de classification de documents qui s'appuient sur des algorithmes classiques (K Plus Proches Voisins, Naives Bayes, Support Vector Machines) en me focalisant sur la représentation de différents types de données, en particulier des dépêches d'actualité [3], textes d'opinion [19], documents administratifs [12], blogs [1], tweets [29].

  Notons qu'au préalable des "compressions" ou "approximations" globales des données peuvent être appliquées. En effet, avant de regrouper les documents (vecteurs) une représentation du corpus de type LSA (Latent Semantic Analysis) peut être appliquée ([Landauer and Dumais, 1997], [3]) afin d'approximer la matrice d'origine dont chaque document représente un vecteur. 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 $USV^T$. $U$ et $V$ sont des matrices orthogonales et $S$ une matrice diagonale. 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 $U$ et $V$. La matrice $U_k S_k V_k^T$ peut alors être considérée comme une version compressée de la matrice originale $A$.

  [Bestgen, 2004] précise que la taille des contextes (documents) est primordiale pour obtenir une qualité des résultats satisfaisante. Cette affirmation confirme les travaux de [Rehder et al., 1998] qui ont effectué des expérimentations pour estimer la taille minimale d'un contexte afin d'obtenir des résultats intéressants avec LSA. Ces expérimentations ont consisté à découper les documents d'un corpus correspondant à des essais d'étudiants en documents de 10 mots, 20 mots, et ceci jusqu'à 200 mots. Les expérimentations ont montré que si les contextes (documents) possèdent moins de 60 mots alors la méthode LSA se révèle décevante.

  Ces résultats représentent une limite pour les données textuelles que nous traitons (textes courts issus du Web2.0) et qui sont présentées dans la section suivante.

### 2.1.2 Le mot est-il un descripteur pertinent dans le contexte Web 2.0 ?

Nous venons de montrer de quelle manière utiliser le mot comme descripteur pour des tâches de classification de textes. Nous allons maintenant appliquer ce principe et discuter ses limites dans le contexte du Web 2.0.

Les premiers travaux sur les données dites du Web 2.0 ont été proposées dans le cadre d'une collaboration avec la société Paperblog [1] en 2008. Ces travaux ont consisté à classer automatiquement des blogs en utilisant le processus décrit dans la section 2.1.1. À partir de 2010, une collaboration avec la société WebReport a consisté à classer les données issues des tweets. Ce travail publié dans [29] est synthétisé ci-dessous.

Ces dernières années, le développement du web social et collaboratif 2.0 a rendu les internautes plus actifs au sein des réseaux participatifs. Les blogs pour diffuser son journal intime de manière massive, les tweets pour publier ses faits et gestes en 140 caractères maximum et autres dépêches RSS sont extrêmement répandus. Simples à créer et gérer, ces outils sont utilisés par les internautes, les entreprises ou autres organisations pour communiquer. Ces nouvelles formes de publication s'inscrivent désormais dans une logique d'intelligence collective et de gestion des connaissances, et ont un potentiel inattendu en termes de veille stratégique. En effet, les professionnels de l'information peuvent les utiliser comme nouvelles ressources documentaires pour y rechercher de l'information. Or, ces derniers se confrontent à l'abondance d'informations. Comment effectuer un tri efficace à partir de cette masse de ressources, pour ne conserver que les informations pertinentes en fonction d'une problématique ?

Dans le cadre d'une collaboration avec la société Web Report [1] une étude sur la classification de tweet a été menée. La société est spécialisée dans l'animation de communautés, la valorisation d'avis et commentaires d'internautes et consommateurs. Des webmasters éditoriaux rédigent des articles, notes de blog, guides d'achat, brèves et autres dossiers multimédia. La société souhaitait développer un outil de veille stratégique pour détecter les informations avant même leur apparition dans les nouvelles des agences de presse. Nous adoptons ici une approche similaire à celle de Google [Ginsberg et al., 2009] qui a montré un lien entre les requêtes des internautes qui utilisent des termes liés à la grippe et le nombre de personnes présentant les symptômes de cette maladie.

Le système Langma développé par la société "Web Report" en collaboration avec le LIRMM vise à fournir un support pour produire puis vérifier des informations sur les catastrophes naturelles qui, si elles sont publiées par un site public, seront qualifiées de "scoop". Cet outil se rapproche de la méthode proposée par [Sakaki et al., 2010] qui détecte les tremblements de terre au Japon via les tweets et dont est issu le site Toretter. [Sakaki et al., 2010] considèrent les tweets comme autant de capteurs produisant des informations sensorielles. Ainsi l'application des méthodes de fouille de textes décrites

1. http ://www.webreport.fr/

dans la section 2.1.1 se sont révélées tout à fait efficaces sur des données de taille réduite et spécifiques comme les tweets [29].

Cette collaboration nous a permis de mettre en relief certaines spécificité des tweets. Contrairement aux documents textuels traditionnels, les descripteurs pertinents des tweets ne sont pas nécessairement les mots-clés d'un "dictionnaire classique". Par exemple la présence d'URL peut se révéler particulièrement pertinente pour classer/discriminer des tweets [Duan et al., 2010]. De plus, des émoticônes (suites de quelques caractères représentant une émotion) présents dans les tweets peuvent aussi révéler un certain sentiment qui peut être détecté automatiquement [Davidiv et al., 2010]. Enfin, les phénomènes d'allongement de lettres (par exemple, le mot *suuuuuuuuper*) permettent également de détecter un sentiment [Brody and Diakopoulos, 2011]. Ceci ouvre des perspectives tout à fait intéressantes qui sont discutées dans le chapitre 4.

De manière globale, les approches "sacs de mots" ne sont pas toujours suffisantes. Dans un processus de fouille de textes, la prise en compte des groupes de mots (syntagmes) ou des associations de mots (par exemple, motifs séquentiels) peut se révéler crucial comme nous l'avons montré dans les travaux liés à l'identification des opinions [19] et des éléments inattendus dans les données textuelles [13].

Ainsi, dans la suite nous allons décrire des méthodes d'extraction de syntagmes en fouille de textes qui sont des groupes de mots qui apportent des informations plus complètes et sémantiquement plus justes.

## 2.2 Les syntagmes

L'étude des syntagmes parfois appelés des *collocations* est décrite dans la section 2.2.1. Nous présenterons des méthodes d'extraction de la terminologie à partir de textes classiques dans la section 2.2.2 et de textes plus atypiques dans la section 2.2.3.

### 2.2.1 Le syntagme en corpus, un descripteur plus riche en fouille de textes

[Clas, 1994] donne deux propriétés définissant une collocation. Premièrement, une collocation est définie comme un groupe de mots ayant un sens global qui est déductible des unités (mots) composant le groupe. Par exemple, "lumière vive" est considéré comme une collocation car le sens global de ce groupe de mots peut être déduit des deux mots "lumière" et "vive". En nous appuyant sur cette définition, l'expression "tirer son chapeau" n'est pas une collocation car son sens ne peut pas être déduit de chacun des mots. De telles formes sont appelées des combinaisons figées. Une deuxième propriété est ajoutée par [Clas, 1994] pour définir une collocation. Le sens des mots qui composent la collocation doit être limité. Par exemple "acheter un chapeau" n'est pas une collocation car le sens de "acheter" et de "chapeau" n'est pas limité. En effet, de multiples objets,

voire des personnes, peuvent être achetés. De tels groupes de mots sont appelés des combinaisons libres. Notons cependant qu'il reste très difficile de différencier par des méthodes automatiques issues du TAL les locutions figées, libres et les collocations. La définition générale des collocations étant donnée, elle peut être enrichie avec deux caractéristiques supplémentaires : les aspects sémantiques et syntaxiques [Heid, 1998, Laurens, 1999]. Le premier point s'appuie sur des caractéristiques sémantiques communes de certaines collocations. Par exemple, "lait tourné" et "beurre rance" ont des sens très proches liés à un phénomène de dégradation. Les aspects sémantiques définissant formellement les collocations sont pris en considération dans de nombreux travaux [Mel'čuk et al., 1999, Heid, 1998, Laurens, 1999]. Ainsi, [Mel'čuk et al., 1999] ont introduit les fonctions lexicales qui s'appuient sur des caractéristiques sémantiques pour définir les relations entre les unités des collocations.

La deuxième caractéristique est liée à la structure syntaxique des collocations. A titre d'exemple, "lumière vive" et "marque distinctive" ont une même structure syntaxique de type Nom-Adjectif. Une classification de la structure syntaxique des collocations que nous donnons ci-dessous est proposée dans de nombreux travaux [Clas, 1994, Laurens, 1999] : Nom-Verbe (par exemple, "interpréter un film"), Nom-Adjectif (par exemple, "cinéma muet"), Nom-Nom/Nom-Préposition-Nom (par exemple, "plateau de cinéma"), Verbe-Adverbe (par exemple, "boire goulûment"), Adverbe-Adjectif (par exemple, "gravement malade").

Les méthodes de TAL ne permettent pas toujours d'identifier les collocations proprement dites qui s'appuient sur les définitions linguistiques énoncées, c'est la raison pour laquelle nous allons nous intéresser à l'extraction globale des syntagmes ou termes.

De multiples approches de recherche terminologique ont été développées afin d'extraire les termes pertinents à partir d'un corpus. Nous ne traiterons pas ici les approches d'aide à la structuration et au regroupement conceptuel des termes qui sont détaillées dans les travaux de [Aussenac-Gilles and Bourigault, 2003].

Les méthodes d'extraction de la terminologie sont fondées sur des approches syntaxiques et/ou statistiques. Le système Termino de [David and Plante, 1990] est un outil précurseur qui s'appuie sur une analyse syntaxique afin d'extraire les termes nominaux. Cet outil effectue une analyse morphologique à base de règles, suivie de l'analyse des syntagmes à l'aide d'une grammaire. Lexter de [Bourigault, 1993] et Syntex de [Bourigault and Fabre, 2000] s'appuient essentiellement sur une analyse syntaxique afin d'extraire la terminologie du domaine. La méthode consiste à extraire les syntagmes nominaux maximaux. Ces syntagmes sont alors décomposés en termes de "têtes" et d'"expansions" à l'aide de règles grammaticales. Les termes sont alors proposés sous forme de réseau organisé en fonction de critères syntaxiques.

Les travaux de [Smadja, 1993] (Xtract) s'appuient sur une méthode statistique. Xtract extrait, dans un premier temps, les termes binaires situés dans une fenêtre de dix mots. Les termes binaires sélectionnés sont ceux qui dépassent d'une manière statistiquement significative la fréquence due au hasard. L'étape suivante consiste à extraire les

syntagmes plus généraux (syntagmes de plus de deux mots) contenant les termes binaires trouvés à la précédente étape.

La grande majorité des systèmes d'extraction de la terminologie est finalement mixte. Ainsi, ACABIT de [Daille, 1994] effectue une analyse linguistique afin de transformer les termes nominaux en termes binaires. Ces derniers sont ensuite triés selon des mesures statistiques. Le système EXIT [21] permet quant à lui de sélectionner des termes binaires et/ou ternaires sur la base de critères linguistiques et/ou statistiques puis de construire itérativement des termes complexes. TERMEXTRACTOR [Sclano and Velardi, 2007] s'appuie également sur des mesures statistiques (entropie) afin de sélectionner les termes pertinents. Un panorama de différentes approches linguistiques, statistiques et mixtes est proposé dans [21].

L'extraction de la terminologie représente les travaux menés au cours de ma thèse soutenue à l'Université Paris-Sud [21]. Les travaux effectués depuis 2005 à Montpellier s'intéressent à l'extraction des termes dans un contexte plus complexe.

Ces dernières années, nous nous sommes intéressés à l'extraction des syntagmes nominaux (SN) complexes et spécifiques tels que des titres (section 2.2.2.1). Par ailleurs, nous avons étudié des SN spécifiques appelés gloses (section 2.2.2.2). Les corpus utilisés sont assez classiques (par exemple, articles journalistiques).

Nous nous sommes également intéressés à l'extraction des SN "classiques" dans des corpus réputés complexes (en particulier, les corpus en ancien français et des fichiers logs). Ces études seront présentées en section 2.2.3.

### 2.2.2 Extraction d'une terminologie complexe à partir de corpus classiques

#### 2.2.2.1 Terminologie et titrage

L'étude présentée dans cette section est issue de la thèse de Cédric Lopez (2009-2012) que je co-encadre avec Violaine Prince. Le travail sur l'extraction de syntagmes propres au titrage a été publié dans [14, 16]. Cette section s'appuie sur l'article [16] publié à la conférence INFORSID'2011.

Les pages Web contiennent une multitude d'informations concernant des domaines divers et variés. L'utilisateur doit souvent fournir de grands efforts cognitifs pour localiser l'information recherchée. Pour les handicapés, alors que l'accès à Internet est un formidable vecteur d'intégration dans la société, la localisation des informations recherchées demeure complexe. Un des domaines clés de l'accessibilité des pages Web tel que défini par la norme proposée par les associations sur le handicap (norme W3C) concerne le titrage (et par expansion le sous-titrage) de pages Web. Côté lecteur, le principal objectif est d'augmenter la lisbilité des pages tout venant obtenues à partir d'une recherche sur mot-clé et dont la pertinence est souvent faible, décourageant les lecteurs devant four-

nir de grands efforts cognitifs. Côté producteur de site Web, l'objectif est d'améliorer l'indexation des pages pour une recherche plus pertinente.

Par ailleurs, le procédé de titrage automatique peut s'intégrer dans diverses applications hors Web. Par exemple, un système d'aide à la rédaction est envisageable, proposant à l'auteur un contenu textuel segmenté thématiquement [Prince and Labadié, 2007] puis titré automatiquement. Ainsi, une nouvelle application industrielle fondée sur le titrage automatique inclurait la génération automatique de sommaires, permettant un gain de temps considérable pour l'auteur.

Le titre (et par extension le sous-titre) est une entité à part entière, possédant ses propres fonctions et se distinguant nettement des tâches de résumé et d'index. L'objectif du titrage automatique est de proposer un/des titre(s) respectant les contraintes mentionnées ci-avant. Les méthodes de TAL seront exploitées dans le but de respecter les contraintes morphosyntaxiques et sémantiques auxquelles doivent se confronter les titres et sous-titres.

Dans cette section, nous proposons un système ayant un double intérêt. Tout d'abord, faciliter l'assimilation du contenu sémantique d'un ensemble de documents textuels à l'utilisateur, ensuite, lui permettre de récupérer l'information pertinente rapidement. S'appliquant sur les ressources textuelles, le traitement proposé consiste à mettre en avant de manière pertinente les sujets abordés en utilisant les titres générés automatiquement, et ainsi faciliter la communication et la localisation des informations.

La détermination de titres pour un document nécessite tout d'abord de savoir quelle est la construction morphosyntaxique des titres et sous-titres habituellement utilisés. À partir de nos études statistiques portées sur les caractéristiques morphosyntaxiques, nous avons mis en place une approche, nommée POSTIT (Utilisation d'information de POsition et d'informations Statistiques pour le TITrage automatique), composée de deux étapes principales consistant à extraire le SN le plus pertinent pour le proposer en tant que titre. La première étape consiste à extraire tous les SN du texte tandis que la seconde étape détermine le syntagme le plus pertinent.

#### Travaux antérieurs en titrage automatique

Le titre est un élément primordial du document. Il désigne le sujet traité par un groupe de mots bien formé, expression, phrase ou simple mot, permettant à la fois de structurer le texte et d'informer le lecteur du contenu. Plusieurs groupes de mots bien formés peuvent donc convenir à un titre. Autrement dit, un texte peut avoir plusieurs titres possibles. Il peut varier en fonction de sa taille (en nombre de mots), de sa forme ou bien du sujet mis en avant. Ainsi, le jugement humain sur la qualité d'un titre sera toujours subjectif.

Les titres ont fait l'objet de nombreuses études linguistiques et sont vus de différentes manières [Peñalver Vicea, 2003] : « porte qui s'ouvre au lecteur », « ensemble de petites unités textuelles », ou encore « élément le plus important de la plupart des textes ». Ces différences d'appréciation induisent que plusieurs titres sont possibles pour un même texte. Le titrage a pour objectif de représenter pertinemment le contenu des documents

en quelques mots. Il peut utiliser des métaphores, l'humour, des jeux de mots [2] ou encore des reformulations.

Les titres peuvent avoir plusieurs fonctions. D'une part, le titre peut être vu comme objet textuel [Ho-Dac et al., 2004] : polices de caractères, tailles, couleurs, etc. Ceci n'est pas la partie que nous étudierons pour l'instant.

D'autre part, le titre permet a priori d'avoir un aperçu de l'article associé. Ainsi, il est doté d'un contenu sémantique qui a trois fonctions : intéresser/captiver le lecteur, informer le lecteur et introduire le sujet de l'article. Il est admis que les éléments apparaissant dans le titre sont souvent présents dans le corps du texte. [Baxendale, 1958] a montré que les premières et dernières phrases des paragraphes sont importantes.

Les récents travaux de [Jacques and Rebeyrolle, 2004] et [Kastner and Monz, 2009] viennent appuyer cette idée et montrent que le taux de recouvrement des mots de titres est très important dans les deux premières et deux dernières phrases du texte. Ainsi, une grande partie de l'information permettant la détermination d'un titre se trouve aux extrémités du document. [Vinet, 1993] remarque que très souvent, une définition est donnée dès les premières phrases suivant le titre. En d'autres termes, des mots pertinents apparaîtront dans les premières phrases du texte.

Dans nos travaux, nous commencerons par analyser statistiquement (taux de recouvrement, nombre de mots, présence de noms communs, verbes, etc) les titres de notre corpus, pour chaque catégorie. Nous mettrons en évidence l'importance de la sélection des syntagmes nominaux pour le titrage. Les résultats portés par les statistiques constitueront une base permettant de déterminer un processus global de titrage automatique, nommé Postit, s'appuyant sur des méthodes de sélection statistique et lexicale.

#### Taux de recouvrement des mots des titres et sous-titres

Afin d'analyser le comportement des titres et sous-titres réels d'articles journalistiques, nous avons constitué un corpus en utilisant la base de données Factiva [3] qui répertorie, entre autres, les articles des grands journaux. Ce corpus contient des articles issus de trois grands journaux français : Le Monde, Le Figaro, Les Echos. Le choix des journaux a été dépendant de la présence des sous-titres dans les articles, afin de faciliter la constitution du corpus. Celui-ci est composé de 300 articles, soit 300 titres, relevant de domaines variés (politique, sport, sciences, etc). Les sous-titres sont au nombre de 354. Le corpus admet un total de 169.796 mots.

L'analyse statistique est primordiale pour envisager une construction automatique de titre. Nous nous sommes intéressés au taux de recouvrement des mots du titre et sous-titre dans le texte (c'est-à-dire à quelle fréquence retrouve-t-on les termes du titre dans le texte). Dans ce calcul, nous n'avons pas tenu compte de la présence de mots fonctionnels (i.e. déterminants, prépositions, etc), ni de la présence de ponctuation. Notons que ces statistiques ont été obtenues après étiquetage, via le TreeTagger [Schmid, 1994], où les entités nommées (EN) correspondent aux noms propres (étiquette NAM du TreeTagger). Les résultats indiquent que dans notre corpus, entre 65% et 68% des mots contenus dans

2. Exemple : « A Montpellier, Ségolène fait un retour royal »
3. http ://factiva.com/

les titres se retrouvent dans le texte (cf. Table 2.1).

En ce qui concerne les sous-titres, le taux de recouvrement est calculé en tenant seulement compte du "sous-texte", c'est-à-dire la partie du texte dépendante du sous-titre. 70% des mots contenus dans les sous-titres se retrouvent dans le sous-texte (cf. Table 2.2). Notons que Le Monde obtient un taux plus faible que les autres journaux (55%), celui-ci préférant utiliser des tournures différentes ou expressions françaises dont les mots ne se retrouvent que rarement dans le texte référé par le sous-titre.

Finalement, une grande partie de l'information nécessaire à la construction d'un titre est présente dans le contenu de l'article. Nous supposons que cette information est suffisante pour la détermination des titres et sous-titres d'un article.

| Journaux | Le Monde | Le Figaro | Les Echos | **Moyenne** |
|---|---|---|---|---|
| Nb. de mots moyen par titre | 6.3 | 4.5 | 5.5 | 5.3 |
| Verbes (en %) | 55 | 52 | 68 | 58 |
| Noms Communs (en %) | 99 | 98 | 99 | 99 |
| Entités Nommées (en %) | 75 | 70 | 72 | 72 |
| **Taux de recouvrement (en %)** | **66** | **65** | **68** | **66** |

TABLE 2.1 – Caractéristiques des titres d'articles journalistiques

| Journaux | Le Monde | Le Figaro | Les Echos | **Moyenne** |
|---|---|---|---|---|
| Nb. de mots moyen par sous-titre | 2.7 | 2.5 | 2.4 | 2.5 |
| Verbes (en %) | 5 | 7 | 10 | 8 |
| Noms Communs (en %) | 99 | 98 | 100 | 99 |
| Entités Nommées (en %) | 7 | 16 | 12 | 12 |
| **Taux de recouvrement (en %)** | **55** | **82** | **74** | **70** |

TABLE 2.2 – Caractéristiques des sous-titres d'articles journalistiques

Afin de savoir comment sont réparties ces informations dans l'article et dans quelles proportions, nous avons découpé le texte en huitièmes (ce qui constitue des segments de taille adéquate pour notre étude). Pour chacune de ces parties, nous comptons les mots du titre et sous-titre s'y trouvant (hors mots fonctionnels). Les résultats sont présentés sous forme de graphe (cf. Figure 2.1). L'abscisse représente les huit parties du texte et les ordonnées donnent le nombre de mots du titre retrouvé dans le texte. Par exemple, un peu plus de 500 (sur 1630 au total) mots présents dans les titres ont été retrouvés dans la deuxième partie des textes de notre corpus (cf. Figure 2.1). Une étude similaire a été réalisée sur les sous-titres (cf. Figure 2.2).

En ce qui concerne les titres, la courbe Total (cf. Figure 2.1) représente la somme des résultats obtenus pour les trois courbes (Le Figaro, Le Monde, Les Echos). Elle est strictement décroissante avec toutefois une exception concernant le dernier huitième du texte qui croit légèrement. Notons que les trois journaux ont globalement le même comportement. Concernant les sous-titres, la courbe Total adopte globalement le même comportement que pour les titres, même si le point culminant de la courbe apparaît peu

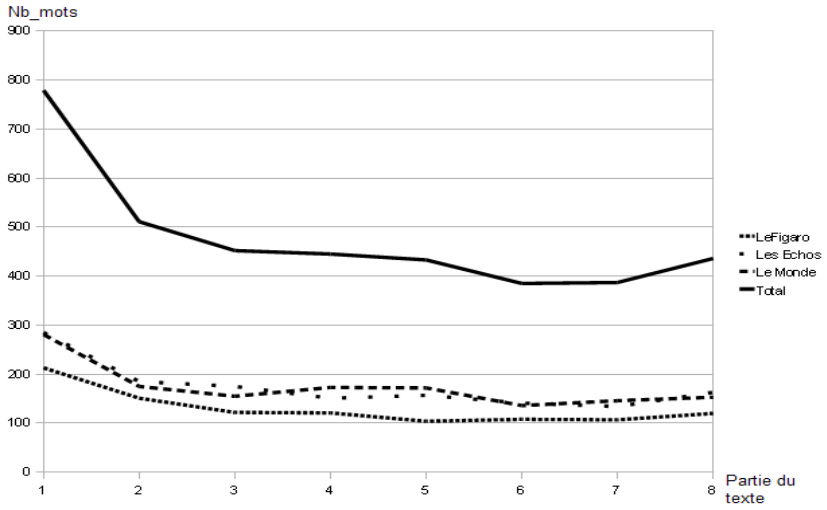


FIGURE 2.1 – Courbes présentant la répartition des mots du titre dans le texte.

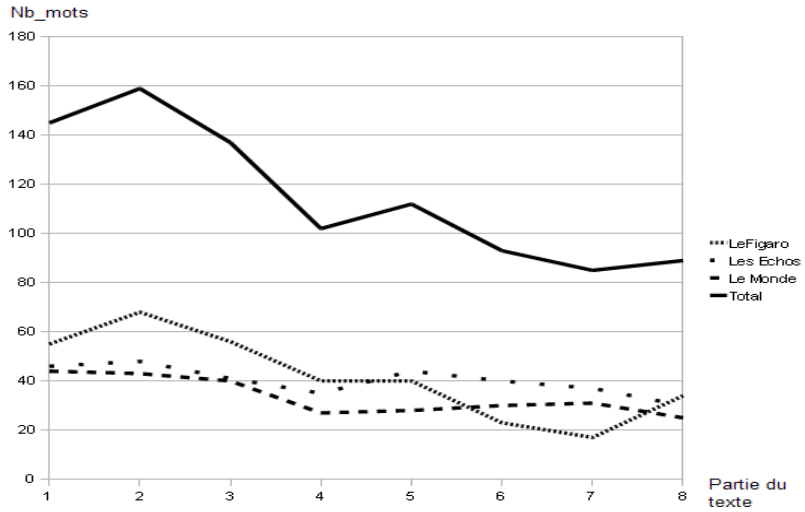


FIGURE 2.2 – Courbes présentant la répartition des mots du sous-titre dans le sous-texte.

après le début du texte (deuxième partie).

Nous pouvons donc considérer que pour les articles journalistiques de notre corpus, les termes pertinents pour le titrage et sous-titrage sont présents en début de texte.

Dans les tableaux 2.1 et 2.2, la ligne "Verbes (en %)" indique le pourcentage de titres contenant au moins un verbe. Ces résultats indiquent une forte prédominance des noms communs et entités nommées par rapport aux verbes. De ce fait, nous proposons une approche consistant à déterminer le syntagme nominal (i.e. un syntagme dont la tête est un nom) le plus pertinent du texte, qui se verra attribuer la fonction de titre. La première étape consiste donc à extraire les syntagmes nominaux candidats au titrage.

POSTIT **: Une approche de titrage automatique par extraction de syntagmes nominaux**

D'après les études statistiques précédemment menées, nous proposons un processus global de titrage automatique (POSTIT) composé de deux étapes :

- **Étape 1 :** *Extraction des syntagmes nominaux candidats au titrage.* L'extraction utilise des filtres syntaxiques tout en s'appuyant sur les études statistiques précédemment menées.

- **Étape 2 :** *Détermination du titre.* Nous mettons en œuvre une méthode statistique permettant le calcul d'un score et mettant en avant les meilleurs syntagmes pour le titrage. Ce score est fondé sur la position du SN dans le texte et sur la pertinence des termes qui le compose.

Ce processus permet d'attribuer un titre à chaque document et à chacune de ses sections. Appliqué à un ensemble de documents, la génération du sommaire permet à l'utilisateur de visualiser facilement les sujets abordés dans cet ensemble. Chaque titre est associé à la section de texte qu'il représente.

Dans la suite, nous présentons les étapes de notre processus global de titrage automatique, illustrées par des exemples issus de notre système.

**Étape 1 : Extraction des syntagmes nominaux (SN)**

Cette première étape consiste à extraire tous les syntagmes nominaux du texte (il est supposé que tous représentent potentiellement un titre). Pour cela, nous utilisons le TreeTagger [Schmid, 1994] qui permet un étiquetage morpho-syntaxique du texte. Nous n'exploitons pas la partie de lemmatisation que cet outil propose. Nous nous sommes appuyés sur les travaux de [Daille, 1994] qui a déterminé des patrons syntaxiques permettant l'extraction de syntagmes nominaux (SN). Par exemple, $Nom-Adjectif$, $Nom-Det-Nom$, $Nom-Nom$ etc. Ainsi, nous avons mis en place un ensemble de patrons syntaxiques per-

mettant l'extraction de syntagmes nominaux pour le français. Nos patrons syntaxiques sont composés des étiquettes suivantes : Nom Commun, Adjectif, Nom Propre, Déterminant, Ponctuation, Préposition, etc.

Par exemple, ces syntagmes nominaux sont extraits automatiquement d'un article issu de Le Monde : "affaire des présumés emplois fictifs à la mairie de la capitale", "au nom de l'éthique", "une quinzaine de militants du collectif", "la réparation", "le maire socialiste de Paris", etc.

Notons que cet ensemble de patrons syntaxiques est dédié au français puisque leur construction est fondée sur des titres réels d'articles français. Un travail similaire pourrait être mené sur les articles journalistiques d'autres langues. Notre processus de titrage POSTIT peut donc s'appliquer sur des articles de langues diverses, à condition de mettre en place un ensemble de patrons syntaxiques adapté à la langue choisie.

Parmi les SN extraits, nous devons déterminer quel est le plus pertinent afin de lui attribuer la fonction de titre. Idéalement, il doit contenir l'information majeure du texte.

#### Étape 2 : Détermination du titre

Notre approche consiste à déterminer le SN le plus pertinent au titrage (et sous-titrage). Le SN retenu sera celui de meilleur score. Le score de chaque SN dépend à la fois de la pertinence des termes qui le composent (utilisation du TF-IDF) et de la position du SN dans le texte.

##### *Le score statistique $SN_{TF-IDF}$*

Nous utilisons le TF-IDF pour calculer le score de chaque syntagme nominal extrait du texte (cf. section 2.1.1 de ce chapitre). Notons que si un nouvel article est inséré dans le corpus, le TF-IDF est recalculé.

Un premier score, $SN_{TF-IDF}$ est calculé pour chaque syntagme nominal. Il s'agit de la somme du TF-IDF de chaque terme composant le SN (hors mots fonctionnels) (formule (2.4)). Le maximum des TF-IDF des termes que nous avons également expérimenté donne des résultats du même ordre [14].

$$SN_{TF-IDF} = \sum_{terme=1}^{n} (TF \times IDF)_{terme} \tag{2.4}$$

À partir de l'exemple précédent associé à l'étape d'extraction, nous obtenons les deux SN suivants : "affaire des présumés emplois fictifs à la mairie de la capitale" ($SN_{TF-IDF} = 0.12$), "la réparation" ($SN_{TF-IDF} = 0.13$).

Dans notre contexte, le principal inconvénient de ce score est qu'il ne tient pas compte de l'emplacement des SN dans le texte. Ainsi, si deux syntagmes nominaux $SN1$ (placé en début de texte) et $SN2$ (placé en milieu de texte) obtiennent un score identique, ils seront considérés comme étant de même degré de pertinence. Or, notre objectif est de corriger ce score en prenant en considération l'information de position absolue des SN dans le texte ($SN_{POS}$).

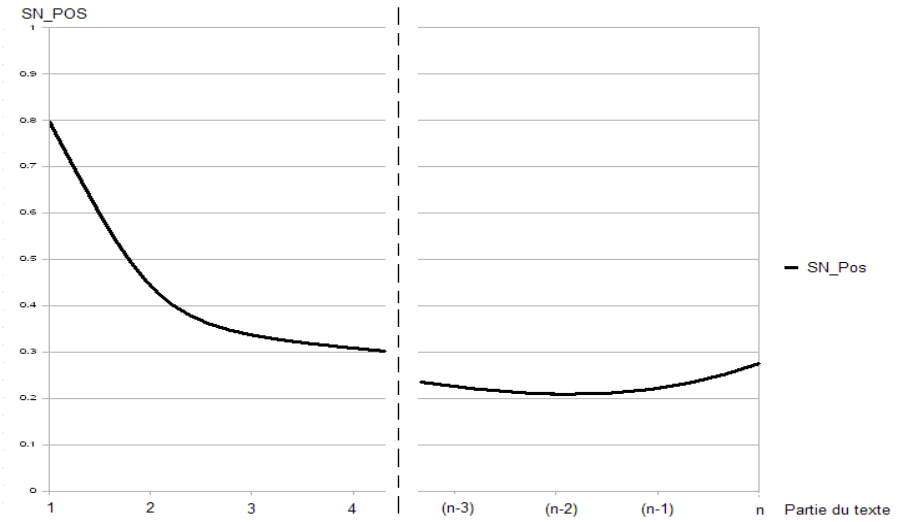


FIGURE 2.3 – Allure de la courbe représentant le score de position $SN_{POS}$.

*Le score de position $SN_{POS}$*

D'après les résultats de notre étude statistique (cf. Figure 2.1), la présence des mots du titre s'atténue au fur et à mesure de l'avancement dans le texte (voir aussi [Zajic et al., 2002]), sauf pour la fin du texte où elle semble à nouveau prendre de l'importance. Afin de tenir compte de cette analyse, notre méthode considère un score $SN_{POS}$. L'intérêt de $SN_{POS}$ est de tenir compte de l'emplacement absolu des SN dans le texte. Le texte est divisé en plusieurs segments de tailles égales (en terme de mots dans notre étude). Soient $n$ le nombre de segments du texte et $P$ la partie du texte où se trouve le syntagme nominal traité ($P \in [1, n]$).

Notre étude statistique a montré que le taux de recouvrement (TR) maximal est obtenu au début du texte. De plus, TR diminue fortement dans les deux premières parties du texte, puis modérément jusqu'à la pénultième partie. Nous traduisons ce phénomène par l'utilisation de la fonction exponentielle (formule (2.5)).

$$SN_{POS}(P) = \begin{cases} e^{1-P} & si\ P \in [1, n-2] \\ e^{2-n} & si\ P = n-1 \\ e^{3-n} & si\ P = n \end{cases} \tag{2.5}$$

Finalement, le calcul de $SN_{POS}$ (formule (2.5)) que nous avons proposé dans [16] traduit fidèlement l'allure globale de la courbe présentée à la Figure 2.1 : décroissante jusqu'à $n-2$ (d'où l'exponentielle) et modestement croissante à partir de $n-2$. Localement, ceci permet d'obtenir une forme hyperbolique centrée autour de $n-2$, pour laquelle nous avons $SN_{POS}(n-3) = SN_{POS}(n-1)$ et $SN_{POS}(n-4) = SN_{POS}(n)$.

Dans ce cadre, le SN "affaire des présumés emplois fictifs à la mairie de la capitale" qui est situé dans la deuxième partie du texte de notre exemple ($P = 2$) obtient

$SN_{POS} = 0.36$. Le SN "la réparation" est situé dans la quatrième partie du texte ($P = 4$) et obtient $SN_{POS} = 0.05$. Le premier SN est donc privilégié grâce à sa position dans le texte. Dans la section suivante, nous proposons une méthode qui combine $SN_{TF-IDF}$ et $SN_{Score}$.

*Le score $SN_{POSTIT}$*

L'objectif de notre étude est de tenir compte de la position du SN dans le texte. Cette dernière information est traduite par le score dit "de position" ($SN_{POS}$) qui vient corriger le score calculé à partir du TF-IDF ($SN_{TF-IDF}$). La variation du coefficient $\lambda$ permet de pondérer le score de position et le score fondé sur le TF-IDF (formule (2.6)). La valeur optimale de $\lambda \in [0, 1]$ pour notre corpus est discutée plus loin.

$$SN_{POSTIT}(P) = \lambda \times SN_{POS} \ + \ (1 - \lambda) \times SN_{TF-IDF} \tag{2.6}$$

Si nous reprenons l'exemple de la section précédente, nous obtenons les résultats suivants : "affaire des présumés emplois fictifs à la mairie de la capitale" ($SN_{POSTIT} = 0.49$), "la réparation" ($SN_{POSTIT} = 0.19$). Alors que ces deux syntagmes nominaux ont un $SN_{TF-IDF}$ quasi identique (resp. 0.12 et 0.13), grâce au score de position, le premier syntagme est privilégié. Le titre attribué à ce texte sera donc : "Affaire des présumés emplois fictifs à la mairie de la capitale".

**Évaluation**

L'objectif de l'évaluation présentée dans cette section est double. Tout d'abord, l'évaluation dite *en surface* consiste à examiner les titres obtenus par notre méthode sur un ensemble de différents textes. Ensuite, elle peut-être associée à une évaluation dite *en profondeur*, concernant le choix du SN parmis tous les SN extraits. À l'issue de ces évaluations, nous proposerons une valeur optimale pour $\lambda$. Dans cette étude, nous posons $n = 8$, c'est-à-dire que chaque texte est découpé en huit parties de taille identique.

*Évaluation en surface*

La première évaluation est réalisée à partir de 90 articles journalistiques issus de notre corpus (30 articles de chaque journal) [4]. Les articles retenus pour cette évaluation sont les trente premiers publiés (du 11 au 15 septembre 2010) pour chaque journal avec la condition qu'ils présentent au moins un sous-titre.

Dans la formule 2.6, nous avons posé $\lambda \in [0, 1]$. Lorsque $\lambda = 0$, alors $SN_{POSTIT}$ ne dépend que de $SN_{TF-IDF}$. Lorsque $\lambda = 1$, alors $SN_{POSTIT}$ ne dépend que de $SN_{POS}$. La variation de $\lambda$ entre 0 et 1 permet de déterminer la valeur adaptée à notre corpus. Au total, ce sont 270 titres qui ont été évalués manuellement (30 articles, soient 30 titres selon 9 valeurs de $\lambda$).

Pour chaque titre, un utilisateur a attribué une des deux étiquettes, "titre pertinent"

4. Rappelons que les trois journaux constituant notre corpus sont : Le Monde, Le Figaro et Les Echos.

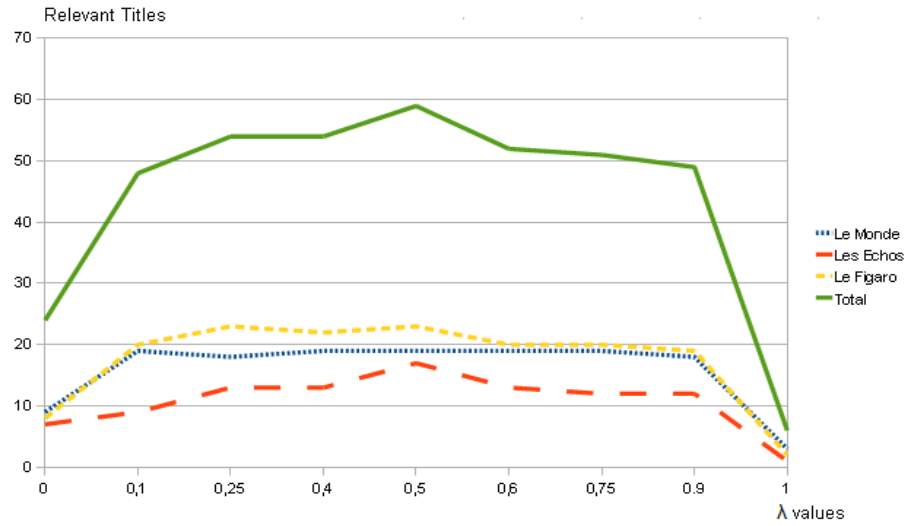


FIGURE 2.4 – Evaluation en surface des titres d'articles journalistiques.

ou "titre non pertinent". Il a été considéré qu'un titre pertinent est un groupe de mots syntaxiquement bien formé donnant un aperçu pertinent du contenu du texte.

La figure 2.4 permet de visualiser la pertinence des titres en fonction de la valeur attribuée à $\lambda$. Les résultats indiquent que pour $\lambda = 0$, 25 articles sont titrés de manière pertinente, contre seulement 8 pour $\lambda = 1$. Les meilleurs résultats de titrage automatique sont obtenus pour $0.4 \leq \lambda \leq 0.6$. Il semble donc que, pour notre corpus, l'information de pertinence (c.-à-d. $SN_{TF-IDF}$) et l'information de position (c.-à-d. $SN_{POS}$) montrent autant d'importance l'une que l'autre. Ainsi, en posant $\lambda = 0.5$, notre méthode permet d'attribuer un titre pertinent à deux titres sur trois (58 titres pertinents pour 90 articles).

Rappelons que plusieurs titres (donc plusieurs SN) peuvent être pertinents pour un même article. Il est donc nécessaire d'étudier la pertinence du choix du SN parmis tous les SN extraits.

*Évaluation en profondeur*

Cette évaluation est réalisée à partir de trois articles journalistiques (un de chaque journal), soit 1681 mots. Tous les syntagmes nominaux extraits de ces articles ont été évalués manuellement, soit 696. Les critères de mesures de performance de notre méthode sont le rappel, la précision et la F-mesure, mesures classiques en fouille de textes.

La précision est définie par le nombre de syntagmes nominaux pertinents retrouvés au regard du nombre total de syntagmes nominaux extraits par notre méthode (formule (2.7)).

$$Précision = \frac{Nombre\ de\ SN\ pertinents\ extraits}{Nombre\ total\ de\ SN\ extraits} \tag{2.7}$$

Le rappel est défini par le nombre de syntagmes nominaux pertinents retrouvés au regard du nombre de syntagmes nominaux pertinents (formule (2.8)). Celui-ci peut-être calculé car tous les SN ont été évalués, permettant ainsi d'obtenir l'ensemble des SN pertinents.

$$Rappel = \frac{Nombre\ de\ SN\ pertinents\ extraits}{Nombre\ total\ de\ SN\ pertinents} \tag{2.8}$$

Enfin, la $F_{Mesure}$ est une mesure populaire qui combine la précision et le rappel (formule (2.9)). Afin de ne privilégier ni la précision, ni le rappel, nous posons $\beta = 1$ pour la suite de notre étude.

$$F_{Mesure} = \frac{(1+\beta^2)(Précision \times Rappel)}{\beta^2 \times Précision + Rappel} \tag{2.9}$$

*Résultats*

Le tableau 2.3 présente la précision, le rappel et la F-mesure pour $\lambda \in [0, 1]$. Le seuil, compris entre 5% et 40% (au-delà de 40%, les résultats sont similaires), correspond au nombre de SN retrouvés par POSTIT, par rapport au nombre total de SN extraits par nos filtres syntaxiques. L'intérêt est d'étudier la présence de titres pertinents retrouvés par POSTIT en fonction du seuil, sachant que plusieurs titres pertinents peuvent apparaître dans la liste de SN. Par exemple, si 260 SN sont extraits du texte, un seuil de 10% indique que 26 SN de plus haut score $SN_{Score}$ extraits par POSTIT sont proposés à l'utilisateur. Un système de bonne qualité devra proposer des titres pertinents en tête de classement.

| Seuil | $\lambda$ | 0 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5% | Précision | 3 | 22 | 28 | 44 | 44 | 44 | 44 | 44 | 44 | 44 | 15 |
| | Rappel | 0 | 56 | 76 | 93 | 93 | 93 | 93 | 93 | 93 | 93 | 17 |
| | **F-mesure** | **0** | **31.59** | **40.92** | **59.74** | **59.74** | **59.74** | **59.74** | **59.74** | **59.74** | **59.74** | **15.94** |
| 10% | Précision | 4 | 17 | 21 | 24 | 24 | 24 | 24 | 24 | 24 | 24 | 21 |
| | Rappel | 20 | 93 | 93 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 83 |
| | **F-mesure** | **6.67** | **28.75** | **34.26** | **38.71** | **38.71** | **38.71** | **38.71** | **38.71** | **38.71** | **38.71** | **33.52** |
| 20% | Précision | 12 | 13 | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 11 |
| | Rappel | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | **F-mesure** | **21.43** | **23.01** | **21.43** | **21.43** | **21.43** | **21.43** | **21.43** | **21.43** | **21.43** | **21.43** | **19.82** |
| 40% | Précision | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| | Rappel | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | **F-mesure** | **11.32** | **11.32** | **11.32** | **11.32** | **11.32** | **11.32** | **11.32** | **11.32** | **11.32** | **11.32** | **11.32** |

TABLE 2.3 – Évaluation des titres d'articles journalistiques (en %).

Les résultats du tableau 2.3 indiquent que les titres les plus pertinents sont obtenus pour $0.30 \leq \lambda \leq 0.90$ (F-mesure = 59,74%) et un seuil de 5%. Finalement, les titres les plus pertinents sont localisés parmi les premiers SN de plus haut score ($SN_{POSTIT}$).

Remarquons que quelque soit $\lambda$ compris entre 0.30 et 0.90, le rappel atteint 100% dès le seuil de 10%. Autrement dit, de manière plus générale, notre méthode permet de concentrer tous les SN pertinents en tant que titres, en tête de classement.

Notons que des expérimentations menées sur d'autres types de corpus (forums, emails) sont présentées dans [14] .

**Discussion**

Notre approche POSTIT résulte de l'analyse ayant mis en évidence l'importance du TF-IDF et de la position du syntagme nominal dans le texte.

À partir d'analyses statistiques s'intéressant à la construction morphosyntaxique des titres réels d'articles journalistiques, un ensemble de patrons syntaxiques a été mis en place permettant l'extraction de syntagmes nominaux dans le texte. La prise en compte de la position des SN et du TF-IDF permet d'extraire un SN pertinent parmi tous les SN extraits par nos patrons syntaxiques.

D'après les résultats des évaluations précédemment menées, notre approche POSTIT propose des titres pertinents pour les articles journalistiques français. En posant $\lambda = 0.5$, celle-ci permet de placer tous les SN pertinents du texte dans les dix premiers pourcents de la liste de SN extraits (souvent supérieure à 200 SN). Par ailleurs, l'évaluation en surface montre que deux titres sur trois proposés par notre approche sont pertinents.

L'approche POSTIT proposée, considère que les titres et sous-titres sont un groupe de mots bien formé ne contenant pas de verbe. Cependant, les statistiques ont montré que, dans notre corpus, un pourcentage non négligeable (58%) des titres d'articles journalistiques contiennent des verbes. Une amélioration possible consisterait à prendre en compte les verbes dans les filtres syntaxiques permettant l'extraction de nouveaux syntagmes, ce qui rapprocherait morphosyntaxiquement les titres déterminés automatiquement des titres réels. Toutefois, rappelons que la priorité de cette approche est d'extraire l'information pertinente du texte et de la présenter en tant que titre. En particulier, l'utilisation de tournures humoristiques ou reformulation ne sont pas l'objet de cette étude.

Un sous-titre peut être considéré comme un titre, qui doit cependant se différencier sur quelques points. Précisément, la taille moyenne d'un sous-titre d'article journalistique est de trois mots. Sa construction morphosyntaxique est donc plus simple que celle d'un titre. Par ailleurs, même si la courbe "Total" de la Figure 2.2 présente une courbe décroissante, celle-ci est largement influencée par le comportement de la courbe pour Le Figaro. Il semble donc que l'on puisse distinguer deux types de sous-titres. Un premier type où les sous-titres suivent le même comportement que les titres (Le Figaro) et un second type où les sous-titres sont construits sans privilégier les premiers termes apparaissant dans le texte (Les Echos, Le Monde).

Nous avons montré que les titres possédaient des spécificités qui demandent la mise en place de méthodes originales. Outre les difficultés liées à la découverte de ces syntagmes, la problématiques des liens entre ces derniers mérite d'être étudiée. Pour attaquer ce problème, nous proposons d'étudier les spécificités langagières voire stylistiques qui nécessitent la prise en compte de connaissances linguistiques plus pointues. Dans ce cadre, nous proposons une étude sur l'identification de *gloses* qui sont des syntagmes tout à fait spécifiques.

#### 2.2.2.2 La terminologie pour la recherche de gloses

Le travail sur les gloses [18] est issu du projet PEPS RESENS (projet pluridisciplinaire financé par le CNRS) que je co-dirige scientifiquement avec Augusta Mela (Université Montpellier 3). Dans le cadre ce projet, linguistes et informaticiens se sont associés pour vérifier dans quelle mesure "le mot et sa glose" permettent un accès efficace au sens des mots.

L'acquisition automatique de connaissances lexicales à partir de textes vise à identifier divers types d'unités lexicales (termes, entités nommées, syntagmes, mots nouveaux, mots à sens nouveau) ainsi que leurs propriétés syntaxiques et sémantiques. Elle constitue une aide précieuse pour la construction de dictionnaires, thesaurus et terminologies, qu'ils soient de langue générale ou spécialisée. Elle intéresse également la recherche documentaire grâce à l'expansion de requêtes.

Les gloses [Mela, 2004, Mela, 2005] sont des commentaires en situation parenthétique, souvent introduits par des marqueurs tels que *appelé*, *c'est-à-dire*, *ou* qui signent la relation de sémantique lexicale mise en jeu : équivalence avec *c'est-à-dire*, *ou* ; spécification du sens avec *au sens* ; nomination avec *dit*, *appelé* ; hyponymie avec *en particulier*, *comme* ; hyperonymie avec *et/ou*, etc. Elles sont en apposition au mot glosé, le plus souvent de catégorie nominale.

La glose est spontanée. Elle partage cette caractéristique avec la définition, dite naturelle parce qu'elle n'est pas le fruit du travail réfléchi d'un lexicographe. Cependant la glose est parenthétique alors que la définition naturelle est l'objet principal du propos. Ainsi, dans *L'accouchement, également appelé travail, naissance ou parturition, est l'aboutissement de la grossesse, la sortie d'un enfant de l'utérus de sa mère*, la glose en *appelé* révèle que *travail*, *naissance* ou *parturition* sont d'autres façons de nommer l'accouchement alors que *l'aboutissement de la grossesse, la sortie d'un enfant de l'utérus de sa mère* est la définition naturelle du mot accouchement. La définition naturelle a été décrite en vue de son traitement automatique par [Rebeyrolles, 2000]. Sa configuration et la copule *être* ne sont pas des marques suffisamment discriminantes mais il n'est pas rare qu'elle soit concomitante d'une glose.

Nous avons réalisé le repérage de gloses en adoptant une première approche dite *p*ar patrons comme dans les travaux de [Jacques and Aussenac-Gilles, 2006, Aussenac-Gilles and Jacques, 2008]. Nous synthétisons notre méthode qui est décrite dans [18] en reprenant le cas des gloses de nomination en *appelé*. Nous partons du principe que $X$ (mot glosé) et $Y$ (mot glosant) sont des syntagmes nominaux (SN [5]), ce qui est le cas général. De plus, nous prenons en compte le fait que $Y$ peut être une coordination de SN, comme dans *L'accouchement, également appelé travail, naissance ou parturition*. La variante du schéma abstrait de glose peut s'écrire : "$X$ marqueur $Y_i$" ($Y_i$ représentent les différents SN glosant).

5. Le corpus est étiqueté au préalable, ce qui nous permet d'utiliser les patrons de termes de [Daille, 1994] pour reconnaître les SN. Nous avons privilégié l'extraction des syntagmes nominaux maximaux.

Un premier patron détecte $Y_1$, le premier SN à droite du marqueur. Par exemple, à partir de la phrase *"Un disque microsillon, appelé disque vinyle"*, nous pouvons extraire $X$ = *Un disque microsillon* et $Y_1$ = *disque vinyle*. Un deuxième patron prend en compte l'éventualité d'une coordination en position $Y$, pour extraire d'une séquence telle que de *Un disque microsillon, appelé disque vinyle ou Maxi*, deux SN : *disque vinyle* et *Maxi*.

Enfin, d'autres heuristiques ont été ajoutées, par exemple, nous privilégions les unités entre guillemets. Ces signes paralinguistiques sont des marques assez discriminantes d'unités "glosantes".

Par la suite, nous avons mis en place une extraction plus large afin d'avoir une couverture plus importante des syntagmes. Dans ce cadre, une fois le marqueur *appelé* détecté dans un corpus, nous recherchons les SN situés entre le marqueur et une frontière droite. Cette dernière est soit un verbe conjugué, soit une ponctuation forte. Pour illustrer notre méthode, posons l'exemple ci-dessous :

*Le Maxi 45 Tours (aussi appelé Maxi 45 ou même tout simplement Maxi ou encore Super 45 tours) est un format de disque microsillon (ou disque vinyle) très apprécié des disc-jockeys et des collectionneurs.*

À partir de cet exemple, quatre SN seront extraits. Le premier instancie $X$, le SN glosé : *Le Maxi 45 Tours*. Les trois autres sont les SN glosant ($Y_i$) qui sont respectivement : *Maxi 45*, *Maxi* et *Super 45 tours*. Des méthodes statistiques issues de combinaison d'approches de fouille du web décrites dans le chapitre 1 permettent de classer les SN extraits en adaptant la mesure $DeMT$ [18]. Le but est alors de présenter à l'utilisateur les syntagmes glosant les plus pertinents.

L'ensemble des résultats de notre méthode est donné dans [18]. Sur une étude menée à partir de 219 SN candidats $Y_i$, l'approche par patron fournie une excellente précision (96 %) mais un rappel plus modeste (63%). Ceci signifie que même si le nombre de SN extrait est assez faible, ces derniers se révèlent d'excellente qualité. A contrario, la méthode étendue retourne un excellent rappel (100%) mais une précision plus faible (75%). Ainsi, tous les SN pertinents ont été extraits mais un quart des SN retournés sont erronés. Les méthodes étendues associées aux fonctions de rang (mesure $DeMT$) permettent de proposer aux experts les SN les plus pertinents de manière ordonnée [18].

Comme pour l'application sur le titrage automatique, nous avons montré, dans cette section, l'importance du syntagme dans les textes classiques pour une tâche spécifique comme l'étude des gloses. Nous nous intéressons dans la section suivante à l'extraction de syntagmes simples à partir de corpus dits complexes.

### 2.2.3 Extraction d'une terminologie classique à partir de corpus complexes

Dans le domaine du TAL, deux types de difficultés sont en général rencontrées. La première est liée aux tâches à accomplir (par exemple, la traduction, la segmentation thématique, l'identification de variations diachronique, etc). Pour confronter les méthodes

selon les tâches, différents challenges (par exemple, TREC [6] ou DEFT [7]) sont organisés. Outre la difficulté inhérente aux tâches, certains challenges s'attelant au traitement de corpus complexes comme par exemple les textes de biomédecine. Ceci constitue une seconde difficulté majeure liée au TAL.

Dans cette section, nous allons plus spécifiquement nous intéresser au traitement (extraction de la terminologie) à partir de textes extrêmement complexes car atypiques. Ce travail nécessite la mise en place d'approches nouvelles et de combinaisons originales de méthodes.

#### 2.2.3.1 La terminologie dans les textes en français médiéval

L'étude présentée dans cette section est issue du stage d'Ingénieur CNAM d'Emmanuel Cazal (2006-2007) que j'ai co-encadré avec Anne Laurent et Maguelonne Teisseire. Ce travail [36, 37] complète la thèse en littérature de Claire Serp menée à l'Université Montpellier 3.

##### Contexte

Les travaux présentés ici reposent sur l'extraction de la terminologie à partir de corpus en ancien français. Le corpus étudié comprend plus de deux milles pages réparties en deux grands ensembles, le cycle Lancelot-Graal (5 ouvrages) et le Perlesvaus. L'ancien français pose deux problèmes majeurs lorsque l'on souhaite en avoir une vision d'ensemble. Tout d'abord, comme le latin dont elle est issue, c'est une langue à déclinaisons. C'est-à-dire que bien que le système soit plus simple que le latin, les mots en ancien français portent des marques particulières en fonction de leur place dans la phrase (par exemple, le mot *chevalier* au singulier en position de sujet s'écrit avec un S, tandis que le même mot en complément d'objet direct s'écrit sans S). La deuxième particularité de cette langue est qu'elle n'a pas de normes orthographiques "fixes", ce qui veut dire que les écrivains utilisent différentes formes pour un même mot, et cela au sein d'un même texte. Nous pouvons par exemple citer le mot *soeur*, que l'on peut trouver dans le tome VII du Lancelot sous les formes suivantes : *soeur*, *serours*, *seur*, *seror*, *seurs*, *serours*, *suer*. Dès lors, il paraît évident qu'un lexique, aussi complet soit-il, ne peut intégrer toutes les variantes orthographiques d'un même mot, et doit se limiter à répertorier les formes les plus usitées.

Les recherches menées s'appuient sur le relevé d'occurrences qui a été réalisé dans le cadre d'une thèse en littérature médiévale sur le thème "Filiation, identité et problèmes de parenté dans les romans du Graal en prose". Celle-ci a permis de mettre en évidence des différences importantes dans le traitement de l'imaginaire de la parenté d'un texte à l'autre, notamment grâce à l'étude du contexte dans lequel apparaissait le terme de *frère*. Il est alors apparu important de mettre en œuvre un processus d'extraction de la terminologie adapté à l'ancien français afin d'aider à cette analyse.

6. Text REtrieval Conference : http ://trec.nist.gov/
7. DÉfi Fouille de Textes : http ://deft.limsi.fr/

Parmi les travaux sur le traitement de textes en ancien français, nous pouvons citer ceux de la BFM (Base de Français Médiéval) fondés sur le système Weblex développé au laboratoire ICAR de l'Ecole Normale Supérieure de Lettres et Sciences Humaines [Heiden and Guillot, 2003]. Les textes sont tout d'abord normalisés selon le format XML permettant d'obtenir le document selon une structure arborescente avec les méta-données associées.

Nos travaux proposent une étude sur la terminologie issue des textes en ancien français. Avant d'extraire la terminologie, il est nécessaire d'utiliser un étiqueteur grammatical. Deux étiqueteurs ont été appliqués à notre corpus médiéval : l'étiqueteur de Brill et le TreeTagger. Le système TERVOTIQ que nous avons proposé permet de classer les termes extraits à partir des textes traités par plusieurs étiqueteurs. À partir d'un corpus étiqueté grammaticalement, le système EXIT [28] utilisé permet, entre autres, d'extraire les candidats termes respectant des patrons syntaxiques simples (Nom-Nom, Nom-Préposition-Nom, etc). La section suivante présente les systèmes d'étiquetage grammatical les plus usités.

#### Étiquetage morphosyntaxique

Cette section présente deux étiqueteurs qui seront utilisés pour traiter des textes en français médiéval : le TreeTagger et l'étiqueteur de Brill.

Le TreeTagger de [Schmid, 1994] estime la probabilité qu'un mot ait une étiquette grammaticale (Nom, Adjectif, Déterminant, etc) en s'appuyant sur des arbres de décision binaires. Ces derniers sont construits récursivement à partir d'un ensemble de trigrammes connus (ensembles de trois étiquettes grammaticales consécutives constituant l'ensemble d'apprentissage). Le processus complet de construction des arbres de décision est décrit dans les travaux de [Schmid, 1994].

L'étiqueteur de Brill appose une étiquette grammaticale à chacun des mots d'un texte en utilisant un lexique, des règles lexicales et des règles contextuelles. Dans ses travaux, E. Brill [Brill, 1994] s'appuie sur un corpus d'apprentissage du Wall Street Journal. Le but est alors d'apprendre des règles d'étiquetage à partir d'un corpus annoté manuellement. À chaque étape d'apprentissage, des règles sont modifiées et le résultat de l'étiquetage avec les nouvelles règles est comparé avec le corpus représentant l'ensemble des annotations justes. Tant qu'un nombre d'erreurs supérieur à un seuil subsiste, le processus d'apprentissage continue. Les transformations des étiquettes s'effectuent (1) en changeant une étiquette par une autre suivant les mots ou les étiquettes des mots proches, (2) en utilisant certaines caractéristiques pour les mots inconnus (lettres en majuscules pour les noms propres, suffixe des mots, etc).

N'ayant pas de corpus étiquetés manuellement en relation directe avec le corpus spécialisé étudié, nous ne pouvons mettre en œuvre une phase d'apprentissage supervisé comme dans les travaux de [Stein, 2003]. Dans un premier temps, notre approche consiste à construire un lexique adapté à l'ancien français.

**Étiquetage morphosyntaxique des textes en français médiéval**

Afin de réaliser un étiquetage de bonne qualité des textes en ancien français, nous utilisons deux lexiques. Le premier est un lexique en français moderne contenant plus de 440 000 formes fléchies obtenu auprès de l'INaLF (Institut National de la Langue Française). Le second est un lexique en ancien français qui contient un peu plus de 45 000 formes fléchies. Celui-ci est issu des travaux de Douglas C. Walker de l'Université de Calgary [8]. Dans ces lexiques, chaque mot est associé à une ou plusieurs étiquettes. Par exemple, le lexique en français moderne possède, pour chaque ligne, la structure suivante : *mot étiquette*$_1$ *étiquette*$_2$ *... étiquette*$_n$. Ces deux lexiques seront par la suite désignés par AF pour le lexique en ancien français et par FM pour le lexique en français moderne. Afin d'améliorer la qualité de l'étiquetage, nous avons fusionné les deux lexiques. Lors de cette fusion, les mots et étiquettes présents dans AF sont privilégiés par rapport aux mots et étiquettes de FM. Notons que dans certains cas, nous pouvons avoir un mot de AF égal à un mot de FM mais avec des étiquettes différentes. Dans le lexique en français moderne, les étiquettes sont associées au mot selon un ordre de probabilité (les étiquettes les plus probables pour un mot précèdent les moins probables). Malheureusement, cette information ne nous est pas donnée par le lexique en ancien français. Par conséquent, dans cette situation, nous positionnons les étiquettes communes au mot de AF et au mot de FM dans l'ordre donné par le mot de FM (ce qui semble plus adapté qu'un ordonnancement aléatoire des étiquettes). Par exemple, l'entrée de AF *de Adverbe Préposition* et celle de FM *de Préposition Déterminant* co-existent. L'occurrence de sortie dans le lexique mixte sera donc *de Préposition Adverbe Déterminant*, les étiquettes de AF étant toujours privilégiées lors de la construction du lexique mixte.

**L'algorithme appliqué pour réaliser la fusion se décline en trois étapes :**

1. La première consiste à vérifier l'existence de chacune des occurrences de AF dans le lexique FM. Si les mots sont égaux, nous vérifions d'abord la présence d'étiquettes communes entre ces mots. Si la similitude d'étiquettes est avérée, nous les positionnons dans l'ordre du mot de FM, sinon nous écrivons dans le lexique mixte l'occurrence du lexique AF.
2. La deuxième étape consiste à ajouter dans le lexique mixte toutes les entrées de AF qui ne sont pas dans FM.
3. Puis, la troisième étape complète le lexique mixte par les mots de FM qui ne sont pas dans AF.

Le lexique AF contenait 2138 mots qui existaient aussi dans FM ce qui représente 4,7%. 43083 entrées de AF étaient quant à elles inconnues du lexique FM, ce qui représente 95.3%. Ainsi, ces entrées qui n'apparaissaient pas dans FM ont été intégrées dans le lexique mixte. Même si cette fusion des lexiques était utile, nous pouvons regretter le fait que le lexique AF ne représente qu'un neuvième du lexique FM en terme de nombres d'entrées.

8. Données disponibles à l'URL : http ://www.acs.ucalgary.ca/ dcwalker/Dictionary/dict.html

**Combinaison des étiqueteurs pour l'extraction de la terminologie : le système** TERVOTIQ

L'approche présentée dans cette section, appelée TERVOTIQ (TERminologie par VOte selon l'éTIQuetage) s'appuie sur un système de vote [Màrquez et al., 1999, Illouz, 1999, Sjobergh, 2003]. En effet, la combinaison d'étiqueteurs [Illouz, 1999] ou d'analyseurs syntaxiques [Monceaux, 2002] donne en général des résultats particulièrement intéressants d'où notre objectif de mettre en place un système de vote fondé sur des étiqueteurs indépendants.

*Attribution des indices de fiabilités initiaux*

L'objectif de TERVOTIQ est de déterminer la pertinence de l'ensemble des termes extraits qu'ils soient fréquents ou non. La méthode de validation retenue repose sur un protocole spécifique. Ce protocole attribue un indice de fiabilité à chaque candidat terme extrait en utilisant un texte traité avec un étiqueteur donné. L'attribution d'un indice à un étiqueteur a été réalisée suite à l'évaluation par l'experte médiéviste du fragment étiqueté par Brill puis par le TreeTagger. Cette évaluation manuelle a montré que le taux d'erreur de Brill était plus élevé que celui du TreeTagger, c'est la raison pour laquelle l'indice de Brill a été attribué à 1 et celui du TreeTagger à 2. Plus l'indice est élevé, plus nous pouvons attribuer une confiance importante aux candidats termes extraits selon l'étiquetage préalable. Grâce au système TERVOTIQ, nous pouvons attribuer une note à chacun des candidats termes extraits.

*Combinaison des indices de fiabilité*

Notre système s'appuie sur l'intersection des candidats termes extraits en leur attribuant la somme des indices de fiabilité initiaux. Les termes obtenus sont alors classés par ordre décroissant selon les indices de fiabilité puis sur la base du nombre d'occurrences pour les termes ayant le même indice global (somme). La principale différence entre les travaux de [Illouz, 1999] et [Sjobergh, 2003] et notre système réside dans le fait que nous ne cherchons pas à mettre en place un système de vote retournant des résultats d'étiquetage de meilleure qualité mais à déterminer les termes les plus pertinents en appliquant différents étiqueteurs. Les résultats détaillés dans [36] montrent que certains termes communs qui sont rares (nombre d'occurrences faible) sont mis en valeur avec notre approche car ils sont présents parmi les candidats termes à indice élevé (indice 3 dans notre cas).

Pour extraire la terminologie propre à ce corpus spécifique, nous avons adapté et combiné les méthodes et outils existants. Cette méthodologie s'est révélée efficace sur des corpus atypiques (ayant des informations lexicales et syntaxiques spécifiques). Dans la suite, nous nous interrogeons sur la pertinence de l'utilisation de méthodes de TAL et de Recherche d'Information à partir de données textuelles encore davantage éloignées des

corpus classiquement traités dans la littérature tels que les fichiers logs.

#### 2.2.3.2 La terminologie dans les fichiers logs

L'étude présentée dans cette section est issue de la thèse de Hassan Saneifar (2008-2011) effectuée en collaboration avec la société Satin Technologies que je co-encadre avec Pascal Poncelet. Le travail sur l'extraction de la terminologie à partir des fichiers logs ainsi que le système global de Recherche d'Information ont notamment été présentés à [32, 33, 34]. Cette section s'appuie sur les articles [31, 34] publiés à JFO'09 (Journées Francophones sur les Ontologies) et CORIA'10 (COnférence en Recherche d'Informations et Applications).

##### Introduction

Dans de nombreux domaines d'application, les systèmes numériques produisent automatiquement des rapports. Ces derniers, appelés *logs* de système, représentent une principale source d'information sur la situation des systèmes, des produits ou même des causes des problèmes ayant pu se produire [Yamanishi and Maruyama, 2005, Facca and Lanzi, 2005].

L'exploitation des fichiers logs liés aux circuits intégrés a pour but d'extraire de l'information sur l'état et les conditions des produits finaux. Dans ce contexte, la création d'une ontologie de domaine [Maedche and Staab, 2001, Bourigault et al., 2004] peut se révéler essentielle afin de généraliser des patrons d'extraction ou étendre des requêtes dans le cadre de la Recherche d'Information. Les ontologies, qui sont particulièrement utilisées dans le domaine de l'Ingénierie des Connaissances, rassemblent sous forme de *concepts* (aussi appelées *classes conceptuelles*) les termes d'un domaine (*instances des concepts*). Les concepts peuvent être liés sémantiquement les uns avec les autres.

Définir le vocabulaire du domaine est l'une des premières étapes de la construction d'une ontologie. Or, selon les spécificités des logs, les méthodes classiques de TAL ne sont pas nécessairement adaptées. Nous avons alors proposé l'approche Exterlog (EXtraction de la TERminologie à partir de LOGs) qui consiste, après avoir prétraité et normalisé les données, à extraire des co-occurrences (avec et sans utilisation de patrons syntaxiques). Le processus mis en œuvre est détaillé ci-dessous.

##### Extraction d'information dans les logs

Dans la conception des circuits intégrés, il existe plusieurs niveaux. À chaque niveau, plusieurs outils de conception peuvent être utilisés. Malgré le fait que les logs issus du même niveau de conception contiennent les mêmes informations, les structures peuvent significativement différer selon l'outil de conception utilisé. Plus précisément, pour la même information, chaque outil de conception possède souvent son propre vocabulaire. Par exemple, dans l'étape de vérification, nous faisons produire deux fichiers logs (par exemple, log "A" et log "B") par deux outils de vérification différents. Une information

comme "`Statement coverage`" sera exprimée sous la forme suivante dans le log "A" :

```
    TOTAL     COVERED  PERCENT
    Lines        10       11     12
statements       20       21     22
```

Mais cette même information dans le log "B", sera exprimée par cette simple ligne :

"`EC: 2.1%`"

Tel que montré ci-dessus, la même information, dans deux fichiers logs produits par deux outils différents, est représentée par des structures et un vocabulaire totalement différents. En outre, les outils de conception évoluent au cours du temps et cette évolution se produit souvent de manière imprévisible. Par conséquent, nous avons besoin de méthodes généralisées pouvant être appliquées sur différents logs hétérogènes (*données textuelles multi-format*). Pour généraliser au mieux ces schémas d'extraction, nous avons besoin d'ontologie du domaine qui fait la correspondance entre des termes utilisés dans les logs issus des outils différents. Nous utiliserons cette ontologie pour généraliser certains patrons d'extraction appliqués aux logs.

Par exemple, pour vérifier "*l'absence des attributs*" sur les logs, nous devons chercher des phrases différentes dans les logs en fonction de la version et du type d'outil de conception utilisé :

- "`Do not use map_to_module attribute`"
- "`Do not use one_cold or one_hot attributes`"
- "`Do not use enum_encoding attribute`"

Au lieu d'utiliser plusieurs patrons, chacun adapté à un fragment, en associant les termes "`map_to_module attribute`", "`one_hot attributes`" et "`enum_encoding attribute`" au concept "*Absence d'attributs*", nous utiliserons un patron générique. Plusieurs approches s'appuient également sur les ontologies de domaine pour mieux guider des démarches d'extraction d'information [Even and Enguehard, 2002].

La création d'ontologie nécessite tout d'abord une analyse lexicale de corpus afin d'identifier les termes du domaine [Sclano and Velardi, 2007]. Nous cherchons donc à identifier la terminologie propre aux logs de chaque outil de conception. Nous insistons ici sur le fait que dans notre contexte, l'ontologie du domaine doit être créée en s'appuyant principalement sur le corpus de logs car l'extraction d'information sera effectuée sur ce même corpus très spécialisé.

### EXTERLOG : EXtraction de la TERminologie à partir de LOGs

Dans la suite, nous allons détailler les trois étapes de notre approche EXTERLOG qui permet d'extraire la terminologie à partir de fichiers logs.

***Étape 1 : Normalisation et étiquetage***

Afin d'avoir des patrons généralisés dans la phase d'extraction d'information du projet, nous effectuons une série de pré-traitements et de normalisation. Les ponctuations, les lignes de séparation et les en-têtes des tableaux sont alors remplacées par des caractères spéciaux pour réduire l'ambiguïté. Nous ne traiterons pas ici la phase de segmentation qui est synthétisée en section 4.2 du chapitre 4 et détaillée dans [35].

Dans notre contexte, des difficultés et des limites subsistent pour appliquer un étiqueteur grammatical classique sur de telles données textuelles. En effet, les techniques classiques sont développées selon une grammaire standard du langage naturel. Par exemple, une phrase se termine par un point, ce qui n'est pas le cas dans les fichiers logs que nous traitons. Malgré ce contexte difficile, nous avons appliqué l'étiqueteur grammatical de Brill [Brill, 1994] en utilisant des règles *contextuelles* et *lexicales* spécifiques à nos données. À titre d'exemple, un mot commençant par un nombre est considéré comme un "cardinal" par l'étiqueteur de Brill. Or, dans les fichiers logs, il existe de nombreux mots comme `12.1vSo10` ; qui ne devraient pas être étiquetés comme "cardinal".

***Étape 2 : Extraction des bigrammes***

Afin d'identifier les co-occurrences dans les logs, nous considérons deux solutions :

1. Extraction des bigrammes de mots en utilisant des patrons syntaxiques (ci-après "*Bigrammes-AP*" [9]),
2. Extraction des bigrammes de mots sans utilisation de patrons syntaxiques (ci-après "*Bigrammes-SP*" [10]).

Dans la première, nous utilisons le filtrage de mots par des patrons syntaxiques. Les patrons syntaxiques déterminent les mots adjacents ayant les rôles grammaticaux définis :
- "\AJ - \NN" (Adjectif-Nom)
- "\NN - \NN" (Nom-Nom)

Notons que les patrons syntaxiques complexes prenant en compte plus de deux mots ne se sont pas révélés pertinents pour les experts du domaine.

Avec la deuxième solution, l'extraction des bigrammes-SP (sans utilisation des patrons syntaxiques) ne dépend pas du rôle grammatical des mots. Les bigrammes extraits représentent deux mots adjacents quelconques.

Ces termes extraits à ce stade doivent être filtrés afin de favoriser les termes les plus pertinents du domaine. Pour cela, nous avons utilisé la mesure $DeMTC$ présentée dans la section 1.2.1.4 du chapitre 1. Nous l'avons appliquée en utilisant le moteur de recherche Google dans le contexte lié aux Circuits Intégrés (CI).

9. AP : Avec Patron
10. SP : Sans Patron

***Étape 3 : Filtrage des termes avec la mesure*** *DeMTC*

Avec la formule *DeMTC*, le contexte est caractérisé par les mots significatifs du domaine (par exemple, *chiffrage*, *information* et *code* pour représenter le contexte Cryptographie). Dans le but de déterminer les mots qui représentent le contexte des logs, nous avons construit un corpus de documents comprenant les documents de référence des outils de conception de circuits intégrés ainsi que trois documents de domaine proche. Nous classons les mots du corpus à l'aide de la mesure *TF-IDF* (cf. section 2.1.1 de ce chapitre). Ensuite, nous avons sélectionné les cinq premiers mots retournés par cette mesure comme contexte.

Notons que comparativement à la mesure *DeMTC* décrite dans le chapitre 1, nous n'utilisons pas une conjonction entre ces mots formant le contexte mais une disjonction. En effet, le fait de travailler à partir d'un contexte très spécialisé nécessite de relâcher certaines contraintes, d'où le choix d'utiliser une disjonction (c.-à-d. un opérateur 'OR' au niveau des requêtes).

La section suivante discute des résultats expérimentaux obtenus en appliquant un tel processus.

**Expérimentations d'**EXTERLOG

Nous expérimentons d'abord les deux méthodes utilisées pour extraire la terminologie à partir des logs : (1) avec utilisation de patrons syntaxiques (*bigrammes-AP*) et (2) sans utilisation de patrons syntaxiques (*bigrammes-SP*). Ensuite, nous validons les termes extraits via un protocole automatique puis manuel. Enfin, la prise en compte de notre méthode EXTERLOG dans le système global de Recherche d'Information [33, 34] est évaluée.

Dans toutes les expérimentations, le corpus d'une taille de 950 Ko est constitué de logs de tous les niveaux de conception.

***Évaluation automatique des bigrammes.***

Pour analyser la performance des deux approches choisies pour l'extraction des bigrammes, dans un premier temps, nous comparons les *bigrammes-AP* et *bigrammes-SP* aux termes extraits à partir des documents de référence du domaine. Pour chaque niveau de conception des circuits intégrés, nous utilisons certains documents, qui expliquent les principes de la conception et particulièrement les détails des outils de conception. Nous employons ces documents comme "références expertes" dans le cadre d'une validation automatique. En effet, si un terme extrait des logs est utilisé dans les références du domaine, nous pouvons le considérer comme un terme valide du domaine. Pourtant, il existe plusieurs termes propres aux logs, en particulier les termes techniques qui ne sont pas utilisés dans le corpus de référence. C'est la raison pour laquelle, une validation par un expert sera indispensable pour compléter la validation automatique.

Le corpus de documents de référence est composé d'environ trois documents par niveau de conception. Ces documents sont de taille considérable. Chaque document est constitué d'environ 600 pages. Étant donné que le corpus de référence est constitué des textes écrits en langue standard contrairement aux logs, nous appliquons une méthode classique d'extraction de la terminologie à partir du corpus de référence (application d'un étiquetage "classique" suivi d'une phase d'extraction des bigrammes [21]).

| Niveau 1 | | Niveau 2 | | Niveau 3 | | Niveau 4 | | Niveau 5 | |
|---|---|---|---|---|---|---|---|---|---|
| AP | SP | AP | SP | AP | SP | AP | SP | AP | SP |
| 67.7 | 11.3 | 20.7 | 6.5 | 37.8 | 9.9 | 40.1 | 6.5 | 19.6 | 5.1 |

TABLE 2.4 – Proportion de bigrammes-AP et de bigrammes-SP retrouvés dans les documents de références

Nous calculons la proportion $P$ (formule (2.10)) des bigrammes-AP et SP retrouvés dans les documents de référence.

$$P = \frac{\mid Bigrammes \cap Termes\,de\,références \mid}{\mid Bigrammes \mid} \tag{2.10}$$

Le tableau 2.4 montre que l'extraction de la terminologie fondée sur les patrons syntaxiques est tout à fait pertinente sur les données logs. Malgré le fait que la normalisation et l'étiquetage des données logs ne soient pas une tâche facile, nos expérimentations montrent qu'un effort dans ce sens est tout à fait utile dans le but d'extraire une terminologie de qualité.

***Évaluation manuelle des bigrammes.***

La validation des termes par un expert est une tâche difficile en raison du nombre des termes extraits par EXTERLOG. Nous avons alors effectué nos expérimentations sur un échantillon composé des 700 termes (bigrammes-AP) les plus fréquents dans le corpus de logs.

Une fois $DeMTC$ calculé, nous classons les termes en fonction de leur score. Plus la valeur de $DeMTC$ est *élevée*, plus le terme est *représentatif* dans notre contexte. Ainsi, nous favorisons les termes les mieux classés dans le but d'augmenter la précision des résultats.

Pour évaluer la performance de $DeMTC$ en tant que mesure de qualité, nous avons fait valider les termes par deux experts du domaine. Un des experts a annoté les termes comme "pertinents" ou "non pertinents" et le second a confirmé ces annotations. Ensuite, nous évaluons la fonction de rang en utilisant les courbes ROC (Receiver Operating Curve). La méthode des courbes ROC 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. Une courbe ROC permet de mesurer la capacité d'une fonction de rang (dans notre cas $DeMTC$) à placer les positifs (termes pertinents) devant les négatifs (termes

| $m$ | $AUC_{IM}$ | $AUC_{IM3}$ | $AUC_{Dice}$ |
|---|---|---|---|
| 200 | 0.53 | **0.60** | 0.59 |
| 300 | 0.61 | **0.70** | **0.70** |
| 400 | 0.62 | **0.67** | 0.64 |
| 500 | 0.66 | **0.72** | 0.71 |
| 600 | 0.72 | **0.75** | **0.75** |
| 700 | 0.74 | **0.77** | 0.76 |

TABLE 2.5 – AUC obtenue à chaque niveau de filtrage en fonction des courbes ROC de la fonction de rang $DeMTC$

non pertinents). On trouve en abscisse des axes représentant une courbe ROC le taux de faux positifs et en ordonnée le taux de vrais positifs.

Il existe également un indicateur synthétique dérivé des courbes ROC : l'AUC (Area Under Curve). Ce critère d'évaluation correspond à la surface entre la courbe et l'axe des abscisses. Ceci indique la probabilité d'un individu (dans notre cas, un terme) positif soit classé devant un individu négatif. Notons qu'en classant les individus aléatoirement, l'AUC sera égale à 0.5.

Le tableau 2.5 présente l'AUC calculée en considérant les $m$ meilleurs termes classés avec nos trois fonctions de rang $DeMTC_{IM}$, $DeMTC_{IM3}$, $DeMTC_{Dice}$. Les résultats montrent que la mesure $DeMTC$ fondée sur IM3 et Dice a tendance à mieux classer les termes extraits en fonction de leur pertinence. Par ailleurs nous avons mené d'autres expérimentations montant qu'une sélection du contexte à partir de l'ensemble des étiquettes *nom*, *adjectif* et *verbe* associé au calcul du TF-IDF (c.-à-d. sélection des mots ayant les valeurs les plus élevées) améliore les résultats.

***Évaluation de la terminologie dans le système de Recherche d'Information global***

Dans cette section, nous discutons de la pertinence des informations terminologiques dans le cadre d'un approche de recherche d'information et plus précisément du système de questions/réponses global mis en place. Ce système précisément décrit dans [33, 34] s'appuie sur une extension des mesures de type TF-IDF tout en prenant en considération le principe du Retour de Pertinence (Relevance Feedback) [Lv and Zhai, 2009].

Dans ce contexte, une approche d'expansion de requêtes a été mise en œuvre. Nous avons alors expérimenté l'utilisation ou non d'informations terminologiques. Dans ce cadre, nous avons calculé la Moyenne des Réciproques du Rang (MRR) qui est un critère utilisé dans le challenge TREC [11] [Voorhees, 1999]. Cette mesure tient compte du rang de la réponse correcte au regard d'un ensemble de questions données. De plus, le rappel a également été évalué.

11. Text REtrieval Conference (Question Answering Track) : http ://trec.nist.gov/data/qa.html

Selon les meilleures configurations de notre système de questions/réponses, l'utilisation des informations terminologiques obtenues avec EXTERLOG permet d'augmenter le rappel (de 88% à 91%) et le score de MRR (de 80% à 85%).

Dans cette section, nous avons décrit un type particulier de données textuelles : les fichiers logs générés par des outils de conception de circuits intégrés. Pour extraire la terminologie des logs, nous avons extrait des co-occurrences. Pour cela, nous avons adapté les méthodes de prétraitement, de normalisation et d'étiquetage grammatical. Le système EXTERLOG proposé fournit des résultats tout à fait intéressants et surtout utiles pour le processus global de Recherche d'Information. Il serait maintenant intéressant d'étudier la pertinence de termes plus complexes (formé de plusieurs mots) qui sont souvent plus pertinents dans d'autres domaines spécialisés [Okazaki and Ananiadou, 2006].

## 2.3 Discussion générale

Les méthodes présentées dans cette section placent le corpus au centre du processus. Le mot en corpus constitue le matériau de base tout à fait pertinent pour des méthodes robustes de classification de textes (c.f. section 2.1.1). L'utilisation de descripteurs plus précis tels que les syntagmes a tendance à améliorer les résultats de ces méthodes [19] comme discuté dans les travaux à partir de textes classiques (pour le titrage et l'identification de gloses) et les textes plus complexes (comme les textes en français médiéval ou les logs). L'extraction des syntagmes peut se révéler extrêmement complexe à partir du moment où les corpus traités sont atypiques. Ainsi, il est nécessaire de mettre en place des solutions originales fondées sur les méthodes de TAL classiques.

Outre les difficultés inhérentes à la complexité des corpus à traiter, une autre difficulté est liée aux tâches à effectuer. En effet, les méthodes à mettre en place dépendent de celles-ci. Par exemple, l'extraction d'un syntagme dans un processus de classification de textes [19] est tout à fait différent comparativement à un syntagme issu d'une application de titrage automatique [14].

Dans ce chapitre, nous nous sommes focalisés sur l'extraction des descripteurs et surtout leur mise en relief par différentes approches (statistiques, linguistiques et le plus souvent mixtes). Ceci peut représenter une limite flagrante pour certaines tâches. Par exemple, les syntagmes proposés comme titres doivent être informatifs et accrocheurs. Ce dernier critère, qui peut se révéler tout à fait capital selon les types de documents à traiter, est beaucoup plus difficile à obtenir par les méthodes d'extraction. C'est ainsi que, de manière globale, il semble essentiel d'étudier des méthodes d'*induction textuelle* qui seront décrites dans le chapitre suivant.

# Chapitre 3

# Induction de syntagmes à partir de corpus

| *Thèmes de Recherche* | *Types de travaux* | *Années* |
|---|---|---|
| Extraction (50%) | Exploitation (50%) | |
| Acquisition de classes conceptuelles | **THÈSE** (Région/CNRS) – N. Béchet | 2006-2009 |
| Titrage automatique | **Collaboration industrielle** – Open-S/EvalAccess (Montpellier)<br>co-responsable scientifique<br>*Stages M2 Recherche*<br>**THÈSE** (Région/UM2) – C. Lopez | 2008-2009<br><br>2009-2012 |

Dans les chapitres précédents, nous avons décrit le traitement des descripteurs linguistiques "plongés" ou non en corpus. Dans ce chapitre, nous nous focalisons sur l'étude des descripteurs linguistiques de type syntagmes. Mais, contrairement aux chapitres précédents, ceux-ci ne sont pas extraits directement des corpus mais ils sont construits à partir des mots présents dans les textes. Ils sont appelés des syntagmes *induits*. La plupart de ces travaux ont été effectués dans le cadre de deux thèses co-encadrées respectivement avec Jacques Chauché et Violaine Prince. Ces dernières traitaient de *classification conceptuelle* (cf. section 3.3) et de *titrage automatique* (cf. section 3.4).

## 3.1 Pourquoi induire des relations syntagmatiques ?

En étudiant le domaine du Traitement Automatique du Langage, il est crucial de nous appuyer sur certains concepts issus de la logique. Une partie de notre discipline est d'ailleurs directement liée à cette thématique. À titre d'exemple, la logique des prédicats du premier ordre permet une représentation des connaissances dans le but d'exprimer des règles. Cette représentation permet de *raisonner* à partir des données textuelles. Trois types de raisonnement peuvent être effectuées : la déduction, l'induction et l'abduction

La *déduction* logique s'appuie sur des axiomes produisant des résultats dits tautologiques. Ces derniers sont toujours vrais selon les règles initialement établies. L'*induction* a quant à elle pour but de produire des propositions générales hypothétiques à partir d'observations. Ces hypothèses doivent être vérifiées. Enfin, l'*abduction* est un raisonnement consistant à déterminer la ou les causes les plus probables d'une observation. Ces deux derniers raisonnements sont dits "hypothétiques".

Dans le domaine de la fouille de textes, nous proposons, via des observations sur l'agencement des mots dans les relations syntagmatiques ou leur positionnement dans les textes, un certain nombre d'*hypothèses*. Ces dernières doivent être validées si elles sont reproduites de manière significative. Cette validation nécessite la prise en compte de connaissances exogènes (dans notre cas les statistiques issues du Web), les corpus étudiés n'étant, en général, pas exhaustifs.

Dans le travail présenté dans ce chapitre, nous proposons des méthodes d'induction de relations syntagmatiques utiles pour bon nombre d'applications : recherche d'Entités Nommées (section 3.2), construction de classes conceptuelles (section 3.3) et titrage automatique (section 3.4). Dans ce cadre, les systèmes proposés reposent sur deux grandes étapes : (1) l'extraction des candidats, (2) la validation de ces derniers. La principale différence tient aux algorithmes de génération et de validation mis en œuvre qui dépendent des tâches à effectuer. Notons cependant que les approches de validation ont un point commun : elles s'appuient, en partie, sur des approches de fouille du web décrites dans le chapitre 1. En effet, les approches de fouille du web permettent (1) de valider la cohérence des syntagmes construits (par la présence des associations sur le web), (2) de mesurer l'intensité des associations (utilisation, entre autres, de mesures statistiques).

Dans la suite de ce chapitre, chacune des sections (sections 3.2, 3.3 et 3.4) se décline en cinq parties :

1. Description du contexte des travaux ;
2. Présentation des approches de construction des candidats (syntagmes induits) ;
3. Présentation des approches de validation de ces candidats ;
4. Expérimentations à partir de données réelles ;
5. Discussion.

## 3.2 Induction des termes pour l'identification des Entités Nommées

Le travail d'identification des Entités Nommées (EN) présenté dans cette section a notamment été publié dans [22].

### 3.2.1 Introduction et contexte

#### 3.2.1.1 Définition des Entités Nommées

Les EN sont classiquement définies comme les noms de Personnes, Lieux et Organisations. Initialement, une telle définition est issue des campagnes d'évaluation américaines MUC (Message Understanding Conferences) qui furent organisées dans les années 90. Cette série de campagnes consistait à extraire des informations telles que les EN dans différents documents (messages de la marine américaine, récits d'attentats terroristes, etc). Aujourd'hui, de telles campagnes d'évaluation couvrent des tâches très variées sur

la base de textes de différents domaines (textes spécialisés en biologie, dépêches d'actualités, blogs, etc). Nous pouvons, entre autres, citer les challenges TREC - Text REtrieval Conference (international) et DEFT - DEfi Fouille de Textes (francophone) qui sont aujourd'hui très actifs dans la communauté "fouille de textes".

Comme le précisent [Daille et al., 2000], les classes de base d'EN définies dans le cadre de MUC doivent être enrichies. Par exemple, outre les classes relatives aux Personnes, Lieux et Organisations, [Paik et al., 1994] définissent de nouvelles classes telles que Document (logiciels, matériels, machines) et Scientifique (maladie, médicaments, etc).

De nombreuses méthodes permettent d'identifier les EN [Nadeau and Sekine, 2007]. Par exemple, des méthodes de fouille de données fondées sur l'extraction de motifs permettent de déterminer des règles (appelées *règles de transduction*) afin de repérer les EN [Nouvel and Soulet, 2011]. Ce type de règles utilise des informations syntaxiques propres aux phrases [Nouvel and Soulet, 2011, Brun and Hagège, 2004]. Par ailleurs, pour identifier les EN, de nombreux systèmes s'appuient sur la présence de majuscules [Daille et al., 2000]. Cependant ceci peut se révéler peu efficace dans le cas d'EN non capitalisées et pour le traitement de textes non normalisés (mails, blogs, textes ou fragments de textes inégalement en majuscule ou minuscule, etc). À titre d'exemple, certaines données du défi DEFT'06 étaient constituées de discours politiques entièrement capitalisés [1]. Ainsi, nous avons choisi dans nos travaux de ne pas exploiter ce type d'informations pour identifier les EN. Notons cependant que de telles caractéristiques pourraient être intéressantes à associer à l'approche essentiellement statistique présentée dans cette section.

Plus formellement, pour caractériser les EN, les critères d'unicité référentielle (c'est-à-dire, un nom propre renvoie à une entité référentielle unique) et une stabilité dénominative (c'est-à-dire, peu de variations possibles) sont notamment précisées par [Fort et al., 2009]. Nous allons nous appuyer sur ce dernier critère pour identifier les EN parmi des syntagmes extraits.

#### 3.2.1.2 Principe général du processus

Il est fréquent que les syntagmes de type Nom-Nom aient des formes variées comme nous le montrerons dans la section 3.2.4. Par exemple, le terme *fichier clients* peut se décliner sous les formes Nom-Préposition-Nom : *fichier de clients*, *fichiers pour clients*, etc. A contrario, les EN sont peu sujettes aux variations [Fort et al., 2009] telles que les "variations prépositionnelles". Nous allons nous appuyer sur cette constatation pour identifier les EN nominales à partir d'une liste de syntagmes de type Nom-Nom. Pour cela, pour chaque candidat Nom-Nom, l'approche que nous décrivons ci-dessous va consister à :

1. Construire artificiellement un syntagme de type Nom-Préposition-Nom à partir du candidat Nom-Nom.

1. Corpus disponible à l'adresse suivante : http ://deft.limsi.fr/

2. Mesurer la "pertinence" du syntagme construit en mesurant la dépendance entre chaque mot par des méthodes statistiques.
3. Sélectionner les syntagmes prépositionnels ayant des scores faibles (c.-à-d. syntagmes construits peu pertinents). En effet, si les possibilités de variations du candidat Nom-Nom sont faibles, nous pouvons supposer que celui-ci peut potentiellement être une EN.

Nous allons maintenant décrire de manière précise chacune de ces étapes en nous appuyant sur les exemples *fichier clients* et *logiciel ciel*. Rappelons que le but de nos travaux est de déterminer automatiquement que le second candidat est en fait une EN.

### 3.2.2 Extraction des candidats

L'extraction des candidats s'appuie sur la méthode menée au cours de mes travaux de thèse [21]. Les travaux ont consisté à extraire des termes respectant des patrons syntaxiques simples (Adjectif-Nom, Nom-Adjectif, Nom-Préposition-Nom, Nom-Nom). Ces termes constituent la base des instances des concepts constituant une "ontologie" spécialisée comme nous l'avons montré dans nos travaux récents [23, 24].

Ainsi, nous n'avons pas effectué de traitement particulier de ces termes extraits. Le but est alors de déterminer, à partir de cette liste, les entités nommées. Le processus mis en place est décrit dans la section suivante.

### 3.2.3 Filtrage des Entités Nommées

#### Étape 1 – Construction

Nous allons dans une première étape construire des candidats prépositionnels en nous appuyant, dans ces travaux, sur la préposition "de" qui demeure la plus courante en français. En appliquant ce principe avec nos deux exemples, nous obtenons les résultats suivants :

- *fichier clients* $_{NN}$ $\rightarrow$ *fichier de clients* $_{N-Prep-N}$
- *logiciel ciel* $_{NN}$ $\rightarrow$ *logiciel de ciel* $_{N-Prep-N}$

Notons que lorsque le second Nom du terme de base commence par une voyelle, la préposition qui sera appliquée sera « d' » :

- *mission intérim* $_{NN}$ $\rightarrow$ *mission d'intérim* $_{N-Prep-N}$

#### Étape 2 – Mesure

Le but de la deuxième étape est de mesurer la dépendance entre chaque mot composant les syntagmes construits. Pour cela nous allons nous appuyer sur la mesure $DeMT_{Dice}$ décrite dans le chapitre 1. Sur la base de l'exemple utilisé précédemment, les valeurs ci-dessous sont alors retournées par notre mesure :

$$DeMT_{Dice}(fichier, de, clients) = \frac{3 \times 999.000}{37.200.000 + 6.350.000.000 + 208.000.000} = 0.000454$$

$$DeMT_{Dice}(logiciel, de, ciel) = \frac{3 \times 89.800}{35.000.000 + 6.350.000.000 + 35.400.000} = 0.0000419$$

Ce résultat montre que le score le plus faible dans des proportions importantes (facteur dix) est donné par *logiciel de ciel*. Ainsi, notre mesure peut prédire que le candidat *logiciel ciel* de type Nom-Nom a statistiquement plus de chance d'être une EN comparativement à *fichier clients*. Ceci est tout à fait pertinent car cette EN fait référence à un logiciel de gestion et de comptabilité, ce qui correspond au type d'EN appelé Document.

Les mesures Web donnent une indication de popularité des termes tout à fait intéressante lorsque des données issues d'un domaine plus ou moins général sont traitées. Par ailleurs, l'avantage de ces connaissances exogènes (c.-à-d.Web) tient au fait que nous sommes moins sensibles à la taille des données traitées (c.-à-d. corpus). En effet, cette taille et donc la fréquence d'apparition des termes doit être assez significative lorsque des méthodes statistiques sont appliquées. Avec nos approches de type "Fouille du Web", nous n'avons pas de telles contraintes liées à la fréquence d'apparition des éléments dans les corpus eux-mêmes.

**Étape 3 – Sélection**

Les candidats syntagmatiques de type Nom-Nom qui obtiennent de faibles scores représentent des éléments peu enclins à la variation. Dans notre approche, de tels candidats seront considérés comme des EN. La section suivante évaluera la proportion de candidats sélectionnés qui représentent réellement des EN. Dans notre approche, nous allons introduire un paramètre $S$ qui représente un seuil de sélection. Par exemple, avec un seuil $S = 10$, les dix candidats ayant les scores les plus faibles seront sélectionnés comme EN potentielles. Les résultats selon différentes valeurs de $S$ seront discutés dans la section 3.2.4.

*Quid de notre approche en anglais ?*

La terminologie nominale de type Nom-Nom possède des formes variantes différentes en anglais. Ainsi, les variantes fréquentes d'un terme de type Nom-Nom (par exemple, *knowledge discovery*) sont constituées d'une préposition associée à une permutation entre les noms (par exemple, *discovery of knowledge*). L'ensemble de ces règles pour caractériser les termes variants sont détaillées dans [Jacquemin, 1997]. Après avoir décrit, notre approche d'identification des EN à partir de syntagmes extraits, la section suivante présente les résultats expérimentaux obtenus sur des données réelles.

### 3.2.4 Expérimentations

#### 3.2.4.1 Protocole expérimental

Le premier corpus traité est composé de 1144 Curriculum Vitae (noté CV) fournis par la société VediorBis (120.000 mots). Une des particularités de ce corpus tient au fait qu'il est composé de phrases très courtes avec de nombreuses énumérations. Les

travaux à partir de ce corpus consistent à déterminer les concepts les plus significatifs pour le domaine [21]. D'autres travaux sur ce même corpus avaient pour but de classer les Curriculum Vitae en deux catégories : CVs de cadres et de non cadres [Clech and Zighed, 2003].

Le second corpus de spécialité étudié (noté RH) est composé d'un ensemble de textes également écrits en français qui sont issus du domaine des Ressources Humaines (société PerformanSe : http ://www.performanse.fr/). Les textes correspondent à des commentaires de tests de psychologie de 378 individus (600.000 mots). Les textes sont écrits par un seul auteur qui emploie un vocabulaire spécifique.

Après l'extraction des termes avec le système Exit [21], un élagage des candidats obtenus est appliqué. Le principe d'élagage consiste à considérer seulement les candidats présents un nombre de fois minimum dans le corpus. L'élagage permet, dans la majeure partie des cas, d'exclure les candidats trop rares qui sont souvent peu représentatifs du domaine [23]. Ainsi, classiquement un élagage à 3 est effectué [Thanopoulos et al., 2002, Jacquemin, 1997].

Le tableau 3.1 présente le nombre de candidats obtenu avant et après élagage à 3.

| | Avant élagage | | Après élagage | |
|---|---|---|---|---|
| | *RH* | *CV* | *RH* | *CV* |
| *Nom-Nom* | 98 | 1781 | 11 | 162 |
| *Nom-Préposition-Nom* | 4703 | 3634 | 1268 | 307 |
| *Adjectif-Nom* | 1260 | 1291 | 478 | 103 |
| *Nom-Adjectif* | 5768 | 3455 | 1628 | 448 |

Table 3.1 – Nombre de syntagmes après et avant élagage.

Suivant les domaines de spécialité écrits dans une même langue, les résultats peuvent différer de manière importante. Par exemple, sur le corpus de CVs après élagage, le nombre de termes de type Nom-Nom (162) est beaucoup plus important que celui du corpus des Ressources Humaines (11) également écrit en français. Le corpus des Ressources Humaines a pourtant une taille cinq fois plus importante que le corpus de CVs. Ceci est dû au fait que les CVs sont écrits de manière condensée en employant un vocabulaire très spécifique : *emploi solidarité*, *action communication*, *fichier client*, *service achat*, etc. De tels candidats pourraient être assimilés à des termes de type Nom-Préposition-Nom : *emploi de solidarité*, *action de communication*, *fichier des clients*, *service des achats*, etc. Dans la section suivante, nous allons nous appuyer sur les syntagmes Nom-Nom du corpus de CVs afin d'en identifier les EN.

#### 3.2.4.2 Résultats

Le but de cette section est d'estimer si les syntagmes sélectionnés par notre approche représentent des EN réellement pertinentes. Dans ce cadre, nous nous sommes appuyés sur 70 candidats de type Nom-Nom les plus fréquents qui sont estimés pertinents (en tant que terme ou EN). Ces candidats ont été évalués manuellement (18 EN ont été identifiées

sur les 70 candidats).

Ces candidats ont alors été classés par la mesure de $DeMT$ décrite en section 3.2.3. Nous avons donc évalué la qualité des candidats sélectionnés en utilisant différentes valeurs du seuil $S$ (sélection des $S$ candidats ayant les valeurs de $DeMT$ les plus faibles). L'objectif est alors d'évaluer si les candidats sélectionnés par notre système correspondent à des EN pertinentes. Notons que les mesures appliquées avec les candidats ont nécessité l'exécution automatique de 210 requêtes avec le moteur de recherche Exalead (http ://www.exalead.com/) qui utilise des ressources en français assez riches. Nous avons alors effectué 70 requêtes pour les numérateurs et 140 requêtes pour les deux noms propres aux dénominateurs (les requêtes pour les prépositions ont été appliquées une seule fois pour l'ensemble des calculs).

Nous allons mesurer les résultats selon différents seuils ($S$) sur la base de ces trois critères. Les résultats sont présentés dans le tableau 3.2. Les résultats montrent que la meilleure valeur de F-mesure est obtenue lorsque nous considérons les 50 premiers candidats ($S = 50$). Ceci s'explique par le rappel maximum (de valeur 1) car toutes les EN se situent parmi les 50 premiers candidats. Les résultats du 3.2 montrent également que les premiers candidats sélectionnés sont assez souvent des EN avec notamment une précision de 60% pour les dix premiers candidats retournés ($S = 10$). Ces derniers sont : *lotus note*, *ciel paie*, *agent recenseur*, *chauffeur livreur*, *go sport*, *rayon fruit*, *accueil client*, *france télécom*, *paris nord*, *front page*. Six candidats sont effectivement des EN (société, lieu, logiciel).

| | | Corpus Français | | |
|---|---|---|---|---|
| | *S* | *10* | *40* | *70* |
| $DeMT_{Dice}$ | Précision | **0.60** | 0.35 | 0.26 |
| | Rappel | 0.33 | 0.78 | **1** |
| | *F-mesure* | *0.43* | **0.48** | *0.41* |
| Classement aléatoire | *F-mesure* | *0.18* | *0.35* | *0.41* |

TABLE 3.2 – Évaluation de différentes valeurs de $S$ – Corpus français

Afin d'évaluer la généralisation de notre approche, nous proposons de travailler avec un corpus anglais composé de termes de type Nom-Nom. Cette liste contient 105 EN (issus des domaines de politique, militaire, religieux, etc) et de 200 termes spécifiques (du domaine de la fouille de données).

Nous avons alors classé tous les candidats en appliquant notre fonction de rang. Ces expérimentations ont nécessité l'exécution de 915 requêtes.

Les résultats présentés dans le tableau 3.3 confirment les résultats présentés à partir du corpus français, à savoir, les premiers candidats retournés sont très souvent des entités nommées (précision de 100% avec $S = 10$). Par ailleurs, de manière similaire au corpus français, la meilleure valeur de F-mesure est obtenue lorsque nous considérons la moitié de la liste de candidats.

| | | Corpus Anglais | | | | | |
|---|---|---|---|---|---|---|---|
| | *S* | *10* | *40* | *70* | *140* | *210* | *305* |
| $DeMT_{Dice}$ | Précision | **1** | 0.83 | 0.73 | 0.58 | 0.46 | 0.34 |
| | Rappel | 0.10 | 0.31 | 0.49 | 0.77 | 0.92 | **1** |
| | *F-mesure* | *0.17* | *0.46* | *0.58* | **0.66** | *0.62* | *0.51* |
| Classement aléatoire | *F-mesure* | *0.06* | *0.19* | *0.28* | *0.39* | *0.46* | *0.51* |

TABLE 3.3 – Évaluation de différentes valeurs de $S$ – Corpus anglais

### 3.2.5 Discussion

Cette section présente une méthode de fouille de textes permettant (1) d'extraire des candidats syntagmatiques, (2) de déterminer des Entités Nommées (EN) à partir de cette liste de candidats. Les EN étant a priori "stables", nous construisons des candidats variants et vérifions leur popularité via la mesure $DeMT_{Dice}$ (cf. chapitre 1). Si les candidats variants construits sont peu pertinents (c.-à-d. valeur faible de $DeMT_{Dice}$), ils sont potentiellement considérés comme des EN.

Précisons que ces méthodes ne prétendent pas filtrer de manière exhaustive les EN mais permettent d'obtenir des résultats intéressants à présenter aux experts pour une phase de validation. Dans ce cas, nous devons, en général, privilégier une valeur élevée de précision. Pour le traitement automatique de quantité importante de données (par exemple, pour des tâches de Recherche d'Information), il est souvent nécessaire de privilégier une valeur élevée de rappel même si les méthodes retournent du bruit. Dans nos futurs travaux, nous envisageons d'enrichir les règles de recherche de variantes. Ce point reste donc crucial à améliorer afin de couvrir la grande majorité des variations linguistiquement pertinentes qui seront validées par les approches statistiques décrites dans ce document.

Enfin, nous combinerons ces approches uniquement statistiques pour prendre en compte certaines spécificités lexicales des EN candidates et leur contexte syntaxique.

Dans les chapitres 1 et 2, nous avons distingué l'utilisation de techniques de fouille de textes *en corpus* et *hors corpus* bien que la frontière ne soit pas toujours si "étanche". Dans cette section liée à l'identification des EN, la méthode présentée s'appuie sur l'application de techniques sur les termes en eux-mêmes (sans prendre en compte les informations liées au corpus). À cet égard, ces travaux relatifs à l'induction sont plus proches de ceux présentés dans le chapitre 1. Dans les sections suivantes, nous allons nous situer dans un cadre de fouille en corpus de manière similaire aux travaux du chapitre 2.

## 3.3 Induction des relations syntaxiques pour la classification conceptuelle

L'étude présentée dans cette section est issue de la thèse de Nicolas Béchet (2006-2009) que j'ai co-encadrée avec Jacques Chauché. La première partie a consisté à étudier l'expansion de contexte dans le but d'effectuer des tâches de classification de textes [3]. Le travail lié à l'induction de relations syntaxiques afin de construire des classes conceptuelles qui est présenté dans cette section a été publié dans [4, 5, 2]. Cette section s'appuie sur l'article [4] publié à la conférence EGC'2009 (Extraction et Gestion des Connaissances).

### 3.3.1 Introduction et contexte

L'acquisition de connaissances sémantiques est une importante problématique en TAL. Ces connaissances peuvent par exemple être utilisées pour extraire des informations dans les textes ou pour la classification de documents. De nombreuses autres applications issues du TAL utilisent des connaissances sémantiques, comme la recherche d'information, la traduction automatique ou l'indexation.

La base de ces connaissances sémantiques peut être le regroupement de mots ou termes sous forme de concepts. Par exemple, les mots *hangar*, *maison* et *mas* peuvent être regroupés dans un concept *bâtiment*. De plus, ces concepts peuvent être organisés sous forme hiérarchique formant ainsi une classification conceptuelle. La majorité des travaux de construction de classification conceptuelle prennent en considération les connaissances d'un expert, celui-ci intervenant rarement dans la phase d'apprentissage. Les travaux effectués avec le système Asium [Faure and Nédellec, 1999, Faure, 2000] proposent de faire intervenir l'expert au cours de cette phase. Nos travaux valident la pertinence de relations syntaxiques, plus précisément de relations induites au sens d'Asium. Ces dernières sont essentielles lors de l'étape de construction de concepts.

### 3.3.2 Extraction de relations syntaxiques induites

La méthode d'Asium consiste à regrouper les objets des verbes déterminés comme proches par une mesure de qualité [Faure, 2000]. D'autres approches utilisent également ce principe de regroupement de termes par des mesures de proximité distributionnelle [Hamon and Nazarenko, 1998, Bourigault, 2002]. Par exemple, dans la figure 3.1, si les verbes *consommer* et *manger* sont jugés proches, des objets pouvant être obtenus par le biais d'informations syntaxiques sont regroupés (dans notre cas, les objets *essence*, *légume*, *nourriture* et *fruit*). Cependant, en considérant ce groupe d'objets, nous pouvons intuitivement exclure le mot *essence*. Notons que les objets *essence*, *légume* et *nourriture* appartiennent à un même contexte en tant qu'objets du verbe *consommer*. De plus, les objets *légume* et *nourriture* sont également des objets du verbe *manger* sur la figure 3.1. Nous appelons dans ce cas l'objet *essence* du verbe *consommer*, un objet **complémentaire**

du verbe *manger*. Par opposition, les objets *légume* et *nourriture*[2] sont appelés des objets **communs** au verbe *manger*, car ils sont également objets du verbe *consommer*. La relation syntaxique formée par un verbe et son objet complémentaire est ainsi appelée relation syntaxique **induite** comme les relations *manger essence* et *consommer fruit* de la figure 3.1.

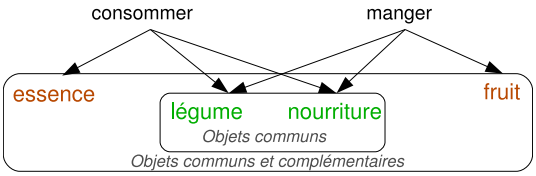


FIGURE 3.1 – Objets communs et complémentaires des verbes "consommer" et "manger".

Notons que ces relations syntaxiques induites sont des connaissances nouvelles "apprisent" à partir des corpus car elles ne sont pas explicitement présentes dans les données textuelles. Notre objectif est de déterminer quels objets complémentaires sont pertinents. Ainsi, nous limitons la tâche de l'expert en proposant de valider uniquement des termes jugés pertinents par nos approches, au lieu de faire valider toutes les relations induites possibles (et donc les éléments des concepts acquis) par un expert. Par exemple, nous souhaitons connaître la pertinence des objets induits de la figure 3.1 *fruit* et *essence* dans un concept automatiquement constitué des mots *essence*, *légume*, *nourriture* et *fruit*. L'évaluation de la qualité des relations syntaxiques induites *consommer fruit* et *manger essence* nous donnera un indice de qualité des objets complémentaires.

Pour déterminer les relations induites à partir de relations syntaxiques issues d'un corpus. Nous utilisons dans un premier temps l'analyseur morphosyntaxique SYGFRAN [Chauché, 1984] pour extraire des relations syntaxiques d'un corpus en français. Nous calculons ensuite la proximité sémantique des verbes issus des relations extraites, en nous appuyant sur la mesure d'ASIUM décrite dans [Faure, 2000]. Ainsi nous considérons que les couples de verbes jugés proches comme par exemple les verbes *consommer* et *manger* de la figure 3.1. Notons que la mesure d'ASIUM considère comme sémantiquement proches deux verbes partageant un certain nombre d'objets en communs. Nous ne conservons finalement que les objets complémentaires des couples de verbes les plus proches sémantiquement.

### 3.3.3 Validation des relations syntaxiques induites

Cette section résume les différentes approches utilisées pour valider les relations syntaxiques induites. La première approche représente une relation syntaxique induite comme une combinaison de concepts issus d'un thésaurus dans un espace vectoriel sémantique. La seconde approche consiste à utiliser la mesure *DeMT* (cf. chapitre 1) pour valider les relations. Nous présentons dans la section suivante ces deux approches ainsi

2. Ce mot est plus général que les autres termes qui tendent à être regroupés. À ce stade seul l'expert peut déterminer une telle situation. Ce point sera abordé en conclusion de la section 3.3 de ce chapitre.

que les combinaisons effectuées entre elles. La section 3.3.4 présente les expérimentations que nous avons menées en évaluant ces approches.

#### 3.3.3.1 Validation par les vecteurs sémantiques

Nous présentons dans cette section notre approche utilisant les vecteurs sémantiques afin de mesurer la proximité sémantique entre le verbe et l'objet des relations syntaxiques. Nous indiquons dans un premier temps, comment sont utilisées des approches similaires dans la littérature. [Wilks, 1998] discute du fait que les différents descripteurs d'un thésaurus comme le Roget peuvent être bénéfiques pour des tâche relatives au TAL. Des approches dites *à la Roget* sont employées dans divers domaines du TAL, comme dans la désambiguïsation sémantique, la recherche d'information, la cohésion textuelle, ou comme mesure de similarité entre termes. Ainsi, [Jarmasz and Szpakowicz, 2003] utilisent la taxonomie de la structure du thésaurus Roget pour déterminer la proximité sémantique entres deux termes. Ils obtiennent de meilleurs résultats que certaines techniques usuelles comme LSA [Landauer and Dumais, 1997] ou PMI-IR [Turney, 2001] 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.

Nous proposons de mesurer la pertinence de l'association d'un verbe avec son objet complémentaire. 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 3.1, il s'agit de mesurer la proximité sémantique des relations *manger fruit* (relation originale) et *consommer fruit* (relation induite). L'objectif est d'attribuer un score à chaque relation induite, afin de les classer par pertinence. Nous avons choisi de représenter une relation induite par un vecteur sémantique. Il est construit en représentant un ou plusieurs termes en le(s) projetant sur un espace de dimension finie de 873 concepts. Ces concepts sont organisés comme une ontologie de concepts définis dans [Larousse, 1992]. Chaque mot est indexé par un ou plusieurs éléments. Par exemple, "*consommer*" est associé à "*fin, nutrition, accomplissement, usage, dépense et repas*" et "*nourriture*" à "*nutrition, éducation, repas et pain*", chaque élément se voyant attribuer un poids de 1.

La représentation vectorielle de la relation syntaxique *consommer nourriture* se traduit par un vecteur de dimension 873, illustrée par la figure 3.2. Ce vecteur résulte d'une combinaison linéaire de la représentation de *consommer* et de *nourriture*, dont les coefficients prennent en compte la structure syntaxique (dans notre cas, un verbe et son objet).

Ainsi, les composantes du vecteur de la relation *consommer nourriture* sont toutes nulles sauf celles associées aux concepts 58, 337, 415, 538, 567, 835, 855 et 857. Afin de mesurer la pertinence d'une relation induite, nous évaluons si cette relation partage les mêmes concepts que la relation syntaxique dont elle est issue. Pour mesurer une telle proximité, nous calculons le cosinus de l'angle formé par les deux vecteurs sémantiques (cf. section 1.1.2 du chapitre 1.1). Notre objectif est d'obtenir un classement des différentes relations syntaxiques par cette mesure de cosinus, afin de valider celles apparaissant en tête du classement. Un exemple de classement de relations induites avec la méthode des

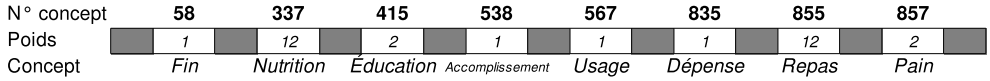


FIGURE 3.2 – Vecteur sémantique de "consommer nourriture".

vecteurs sémantiques est présenté en section 3.3.3.3.
Une autre manière d'attribuer un score aux relations induites est décrite dans la section suivante.

#### 3.3.3.2 Validation Web par application de la méthode $DeMT$

Dans cette section, nous proposons 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, dans notre cas une relation syntaxique, sous forme de chaîne de caractères (par exemple, *"consommer un fruit"*). La requête fournie au moteur de recherche est donnée par la fonction $nb(x)$ qui représente le nombre de pages de résultats retournés par la soumission d'une chaîne de caractères $x$ au moteur de recherche de Yahoo en utilisant une API [3]. En français, langue sur laquelle nous nous appuyons dans nos travaux, un verbe est couramment séparé d'un objet par un article. Ainsi, nous considérons cinq articles fréquents : *un*, *une*, *le*, *la*, *l'* pour composer notre chaîne de caractères représentant notre relation induite. $nb(v, o)$ est alors défini comme le nombre de pages retournées pour la relation syntaxique Verbe-Objet ($v$,$o$) avec respectivement $v$ et $o$, le verbe et l'objet de cette relation. La formule suivante décrit le calcul effectué afin d'obtenir la fréquence d'apparition d'une relation syntaxique :
$nb(v, o) = nb(v$ **un** $o) + nb(v$ **une** o$) + nb(v$ **le** o$) + nb(v$ **la** o$) + nb(v$ **l'** o$)$
$nb(v$ **un** $o)$ est la valeur retournée par le moteur de recherche avec la chaîne de caractères $v$ **un** $o$.
Afin d'évaluer la dépendance entre le verbe $v$ et l'objet $o$ d'une relation induite, nous nous appuyons sur la mesure $DeMT_{IM3}$ décrite dans les sections 1.2.1.3 et 1.2.1.4 du chapitre 1. Les relations induites les plus pertinentes selon les deux approches (vecteurs sémantiques et validation Web) doivent alors se retrouver en tête de liste. Un exemple de classement de relations induites avec la méthode de la validation Web est présentée en section 3.3.3.3.

#### 3.3.3.3 Validation par hybridation

Pour optimiser nos deux approches précédemment présentées, les vecteurs sémantiques (VS) et la validation Web (VW), nous proposons de combiner ces deux techniques.

**Combinaison 1 : Une combinaison pondérée par un scalaire**

La première combinaison entre ces approches consiste à introduire un paramètre

3. http ://api.search.yahoo.com

$k \in [0, 1]$ 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 :

$$combine\_score_r = k \times VS + (1 - k) \times VW \tag{3.1}$$

**Combinaison 2 : Un système hybride adaptatif**

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".

**Exemple de classement avec cinq relations induites**

| Relations Verbe-Objet | | Cosinus |
|---|---|---|
| Induites | Originales | |
| poursuivre réforme | demander réforme | 0,60 |
| dépasser recherche | faire recherche | 0,52 |
| réussir évaluation | faire évaluation | 0,41 |
| dire croisade | poursuivre croisade | 0,37 |
| lancer recherche | mener recherche | 0,27 |

TABLE 3.4 – Résultats avec les vecteurs sémantiques.

Cette section présente un exemple de classement de relations syntaxiques induites avec les différentes approches présentées : les vecteurs sémantiques, la validation Web et les deux approches de combinaisons.
Nous calculons tout d'abord les scores résultant de l'approche des vecteurs sémantiques. Les cinq relations syntaxiques induites et celles existantes sont présentées dans le tableau 3.4. Nous les représentons dans un premier temps sous forme de vecteurs sémantiques avec SYGFRAN. Nous pouvons alors calculer le cosinus entre les relations syntaxiques induites et celles existantes (cf. section 3.3.3.1). Les résultats obtenus pour les cinq relations testées sont présentés dans le tableau 3.4.

| Relations 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 | 0,41 |
| réussir évaluation | 27 600 000 | 57 900 000 | 1 366 | 0,35 |
| dépasser recherche | 15 900 000 | 863 000 | 363 | 0,28 |

TABLE 3.5 – Résultats avec la validation Web.

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 pages retournées par le moteur de recherche (fonction $nb$). Alors nous pouvons calculer $DeMTBase_{IM3}$ pour les cinq relations induites. Les scores obtenus sont présentés dans le tableau 3.5.

Nous appliquons alors la première méthode de combinaison. Nous fixons à titre d'exemple le scalaire $k$ à $0,5$ pour donner le même poids à chacune des approches (VS et VW). De plus, dans cet exemple, nous fixons la constante $n$ de la mesure hybride adaptative à 3. Rappelons que la mesure adaptative consiste à classer dans un premier temps les relations syntaxiques induites avec les vecteurs sémantiques, pour ensuite classer dans notre cas les 3 meilleures avec la validation Web. Les résultats obtenus pour les deux combinaisons [4] sont présentés dans le tableau 3.6. Une fois l'ensemble des scores obtenus pour chacune des approches, nous pouvons ordonner les relations syntaxiques. Le classement ainsi obtenu est donné dans le tableau 3.7.

| Relations Verbe-Objet | VS | VW | Combinaison 1 | Combinaison 2 |
|---|---|---|---|---|
| lancer recherche | 0,60 | 0,49 | 0,55 | 1,49 |
| poursuivre réforme | 0,41 | 0,35 | 0,38 | 1,35 |
| réussir évaluation | 0,52 | 0,33 | 0,43 | 1,33 |
| dépasser recherche | 0,37 | 0,13 | 0,25 | 0,37 |
| dire croisade | 0,27 | 0,71 | 0,49 | 0,27 |

TABLE 3.6 – Relations syntaxiques triées avec l'ensemble des approches.

| 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 |

TABLE 3.7 – Classement obtenu des relations syntaxiques.

### 3.3.4 Expérimentations

#### 3.3.4.1 Protocole expérimental

Nous extrayons d'un premier corpus les relations induites, que nous ordonnons qualitativement par nos différentes approches. Ce premier corpus écrit en français est extrait du site Web d'informations de Yahoo (*http ://fr.news.yahoo.com/*). Il contient 8 948 articles (16,5 Mo). Ce corpus est utilisé comme corpus de test, il sera nommé *corpus T*. Nous avons obtenu à partir de ce corpus, 60 000 relations syntaxiques induites. Afin de mesurer la qualité de nos relations induites, nous utilisons un second corpus (*corpus V*) écrit également en français, de taille plus conséquente (125 Mo). Celui-ci contient plus de 60 000 articles issus du corpus du quotidien Le Monde. Par ailleurs, les deux corpus

4. 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 ajoutons 1, ce qui place ces 3 relations automatiquement en tête de liste (car les scores sont normalisés).

| *Seuil* | *VW* | *VS* | *Seuil* | *VW* | *VS* |
|---|---|---|---|---|---|
| *5000* | 0,61 | 0,51 | *35000* | 0,75 | 0,55 |
| *10000* | 0,65 | 0,52 | *40000* | 0,76 | 0,56 |
| *15000* | 0,68 | 0,54 | *45000* | 0,78 | 0,55 |
| *20000* | 0,70 | 0,54 | *50000* | 0,79 | 0,54 |
| *25000* | 0,72 | 0,55 | *55000* | 0,80 | 0,54 |
| *30000* | 0,74 | 0,55 | *60000* | 0,81 | 0,54 |

TABLE 3.8 – AUC obtenues avec les approches Validation Web et Vecteurs Sémantiques.

sont du même domaine, actualités avec un style journalistique. Nous jugeons comme pertinentes des relations induites créées à partir du *corpus T* si celles-ci sont présentes dans le *corpus V*. Concrètement, si une relation induite est retrouvée dans le *corpus V*, on la qualifiera de **positive**. Dans le cas contraire, elle sera jugée non pertinente et sera donc qualifiée de **négative**. Nous avons opté pour cette validation afin de pouvoir mesurer la qualité de nos approches, sur un grand nombre de relations, de manière automatique (qu'un ou des expert(s) humain(s) ne pourrai(en)t évaluer, faute de temps). Notons que les relations jugées négatives peuvent être de faux négatifs. En effet, une relation qui n'a pas été retrouvée dans le *corpus V* n'est pas pour autant non pertinente. Nous proposons d'évaluer nos différentes approches en utilisant les courbes ROC (cf. section 2.2.3.2 du chapitre 2).

Dans notre cas, on trouve en abscisse des axes représentant une courbe ROC le taux de relations syntaxiques induites non pertinentes (c-à-d. relations non retrouvées dans le *corpus V*) et en ordonnée le taux de relations pertinentes (c-à-d. relations existant dans le *corpus V*). Rappelons que si toutes les relations sont pertinentes, l'AUC (*Area Under the Curve*) a une valeur de 1. Ceci signifie avoir toutes les relations pertinentes en début de liste, donc ordonnées de manière optimale.

#### 3.3.4.2 Résultats

Nous présentons dans cette section, les résultats de l'évaluation de nos approches présentées en section 3.3.3, pour différents seuils donnés. Concrètement, nous avons produit 60 000 relations syntaxiques induites avec le corpus de test, que nous avons ensuite ordonnées avec nos approches en appliquant un seuil. Ce dernier permet de mesurer l'impact de nos approches, en fonction du nombre de relations syntaxiques induites à considérer. Ainsi, un seuil fixé à 5 000 relations signifie que nous calculons l'AUC sur les 5 000 premières relations ainsi classées.
Le tableau 3.8 présente les AUC obtenues avec l'utilisation des vecteurs sémantiques et la validation Web. Les AUC obtenues avec les vecteurs sémantiques ne sont pas très satisfaisantes avec des scores proches d'une distribution aléatoire ($AUC = 0.5$). La structure des vecteurs sémantiques peut expliquer ces résultats. En effet, les 873 concepts définissant ces vecteurs ne sont pas assez discriminants, notre corpus utilisant un vocabulaire assez homogène (corpus d'actualités). Cette faible dimension des vecteurs sémantiques ne permet donc pas de classer assez finement nos relations syntaxiques induites. La validation Web donne quant à elle de meilleures AUC. Pour un seuil réduit, inférieur à 20 000, les résultats restent néanmoins assez faibles (moins de 0.70). Cela signifie que pour

l'ensemble des relations syntaxiques évaluées, cette approche est meilleure (AUC de 0.81) mais que le classement des premières relations reste difficile. Ceci peut s'expliquer par le fait que les relations syntaxiques les plus populaires ne sont pas nécessairement les plus pertinentes.

Nous proposons alors d'appliquer la première approche (combinaison 1) effectuant une combinaison des vecteurs sémantiques et de la validation Web. Nous faisons varier le paramètre $k$ de 0 à 1 par intervalle de 0.1.

| *Seuil* | VW (k=0) | $k$ = 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | VS (k=1) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *5000* | 0,61 | 0,61 | 0,61 | 0,62 | 0,62 | 0,63 | 0,63 | **0,66** | 0,55 | 0,56 | 0,51 |
| *10000* | 0,65 | 0,66 | 0,66 | 0,66 | 0,67 | 0,67 | **0,67** | 0,67 | 0,64 | 0,57 | 0,52 |
| *15000* | 0,68 | 0,68 | 0,68 | 0,68 | 0,69 | 0,70 | 0,70 | **0,71** | 0,65 | 0,56 | 0,54 |
| *20000* | 0,70 | 0,70 | 0,71 | 0,72 | 0,73 | **0,74** | 0,73 | 0,71 | 0,66 | 0,58 | 0,54 |
| *25000* | 0,72 | 0,73 | 0,75 | 0,75 | 0,75 | **0,76** | 0,75 | 0,73 | 0,69 | 0,62 | 0,55 |
| *30000* | 0,74 | 0,76 | 0,77 | 0,78 | 0,78 | **0,78** | 0,77 | 0,75 | 0,71 | 0,63 | 0,55 |
| *35000* | 0,75 | 0,78 | 0,79 | **0,79** | 0,79 | 0,78 | 0,77 | 0,75 | 0,72 | 0,64 | 0,55 |
| *40000* | 0,76 | 0,79 | **0,79** | 0,79 | 0,79 | 0,78 | 0,77 | 0,75 | 0,71 | 0,65 | 0,56 |
| *45000* | 0,78 | **0,79** | 0,79 | 0,79 | 0,79 | 0,79 | 0,78 | 0,76 | 0,73 | 0,67 | 0,55 |
| *50000* | 0,79 | **0,80** | 0,80 | 0,79 | 0,78 | 0,75 | 0,74 | 0,72 | 0,69 | 0,64 | 0,54 |
| *55000* | 0,80 | **0,80** | 0,79 | 0,78 | 0,75 | 0,73 | 0,71 | 0,69 | 0,66 | 0,62 | 0,54 |
| *60000* | **0,81** | 0,79 | 0,78 | 0,76 | 0,74 | 0,72 | 0,70 | 0,68 | 0,65 | 0,61 | 0,54 |

TABLE 3.9 – AUC obtenues avec la première combinaison.

| *Seuil* | *VW* | *Comb. 1* | *Comb. 2* |
|---|---|---|---|
| *5000* | 0,61 | **0,66** | **0,83** |
| *10000* | 0,65 | **0,67** | **0,82** |
| *15000* | 0,68 | **0,71** | **0,83** |
| *20000* | 0,70 | **0,71** | **0,83** |
| *25000* | 0,72 | **0,73** | **0,83** |
| *30000* | 0,74 | **0,75** | **0,83** |

| *Seuil* | *VW* | *Comb. 1* | *Comb. 2* |
|---|---|---|---|
| *35000* | 0,75 | 0,75 | **0,83** |
| *40000* | 0,76 | 0,75 | **0,83** |
| *45000* | 0,78 | 0,76 | **0,83** |
| *50000* | 0,79 | 0,72 | **0,82** |
| *55000* | 0,80 | 0,69 | **0,82** |
| *60000* | 0,81 | 0,68 | **0,81** |

TABLE 3.10 – AUC obtenues avec la seconde combinaison.

Le tableau 3.9 montre les résultats obtenus avec cette approche. Intuitivement, cette combinaison devrait donner des résultats pertinents pour des valeurs de $k$ faibles, ce qui privilégie l'approche VW. Néanmoins, cette supposition ne se vérifie pas pour tous les seuils. En effet, nous obtenons, pour la première moitié des seuils considérés, de meilleures AUC que celles obtenues pour la validation Web, avec des valeurs de $k$ autour de 0.7. Ces améliorations sont cependant peu significatives, de l'ordre de 3%. Pour une valeur de $k$ supérieure à $0,5$, nous obtenons des AUC assez proches de celles obtenues avec la validation Web. Ces résultats nous amènent à penser que l'approche utilisant les vecteurs sémantiques peut se révéler pertinente pour certaines relations. Ainsi, nous allons évaluer la seconde approche qui selectionne dans un premier temps et de manière globale des relations retenues par les vecteurs sémantiques. Le tableau 3.10 présente la seconde combinaison aux meilleurs scores obtenus pour de faibles seuils (les plus difficiles à améliorer) pour les approches précédentes : la validation Web et la première combinaison avec $k = 0.7$. La figure 3.3 présente quant à elle, les courbes ROC correspondantes au seuil 5000.

Pour la seconde combinaison, nous allons classer les 60 000 relations avec les vecteurs

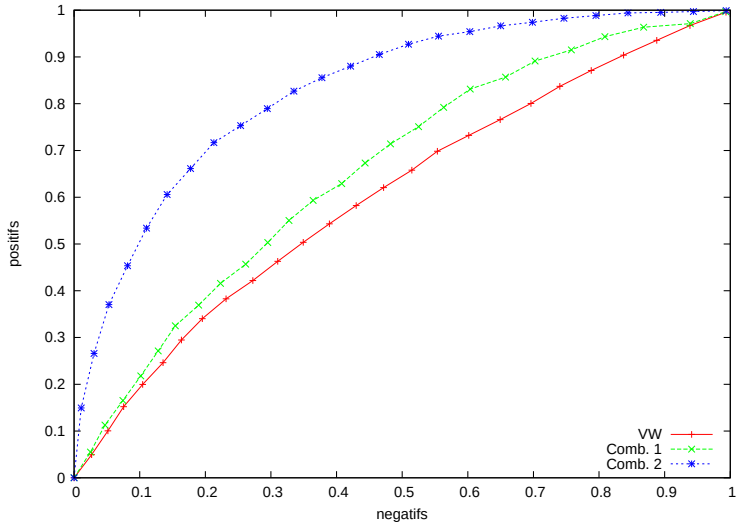


FIGURE 3.3 – Courbes ROC obtenues avec la validation Web (VW), la première combinaison (Comb. 1) et la seconde combinaison (Comb. 2) au seuil 5 000.

sémantiques. Puis, nous classons les $n$ premières avec la validation Web (la valeur de $n$ propre au paramètre de la seconde combinaison correspond ici au seuil).

Nous obtenons avec la seconde combinaison de meilleures AUC que les approches précédentes, quelque 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 0.83 avec la seconde combinaison. 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 le score obtenu par la seconde combinaison n'est pas dépendant du choix du seuil car les AUC restent relativement constantes (AUC variant de 0.81 à 0.83).

### 3.3.5 Discussion

L'aire sous la courbe ROC (AUC) est une bonne indication 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 l'expert. Nous proposons alors de calculer pour un faible seuil, soit les 1 000 premières relations induites, la précision des approches VW et la seconde combinaison. La précision qui calcule la proportion de relations induites correctes retournées par le système. La figure 3.4 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. La figure 3.4 montre que celle-ci est nettement meilleure avec la seconde combinaison pour les 300 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

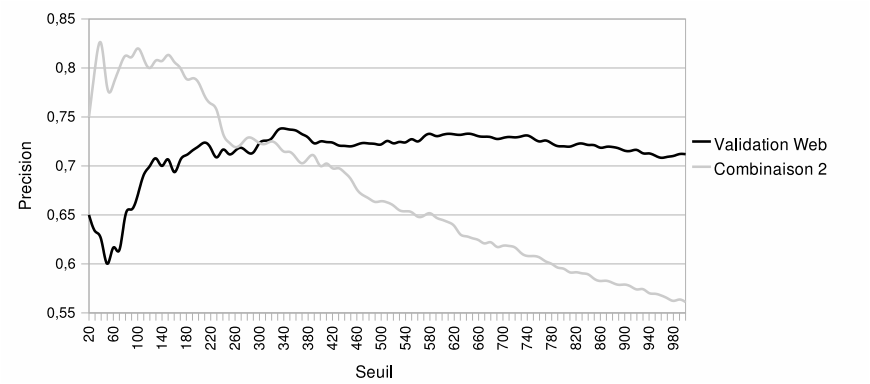


FIGURE 3.4 – Courbe lift comparant le classement de la VW et de la seconde combinaison.

avec les courbes ROC. Nous avons alors cherché à savoir si les relations syntaxiques placées en tête de liste avec la seconde combinaison étaient les mêmes que celles classées en tête avec la validation Web. Autrement dit, est-ce que les relations privilégiées par la seconde combinaison sont les relations les plus populaires sur le Web ?

Nos expérimentations ont montré que les relations en tête de liste avec la seconde combinaison ne sont pas les mêmes que celles en tête du classement effectué avec la validation Web. Par exemple, sur les *300* meilleures relations issues de la combinaison 2, avec VW, *37 (12 %)* sont placées entre $[0, 300[$, *39 (13 %)* entre $[300, 600[$, *39 (13 %)* entre $[600, 900[$ et *186 (62 %)* au rang plus élevé. Citons par exemple la relation syntaxique *détenir_arme*, classée $150^{ieme}$ avec la seconde combinaison qui est classée $1159^{ieme}$ avec la validation Web. Cela nous indique que la seconde combinaison permet de déterminer des relations moins fréquentes sur le Web, n'étant pas en tête de liste du classement avec la validation Web. Ainsi, nous déterminons et validons des pépites de connaissances, pouvant être plus discriminantes et plus intéressantes que des relations fréquentes qui n'apportent pas d'informations nouvelles.

**Conclusion**

Nous avons présenté dans cette section des solutions permettant de réduire l'implication humaine dans la validation de relations syntaxiques, qui sont dans notre cas, des relations dites induites. Ces relations ne sont originalement pas présentes dans le corpus, et peuvent permettre par exemple d'enrichir des ontologies. Nous les déterminons en considérant la proximité des verbes et les objets respectifs de ces verbes. Dès lors, nous proposons plusieurs approches afin de proposer à l'expert les relations induites les plus pertinentes.

Nous envisageons dans de futurs travaux d'étendre nos approches aux relations syntaxiques *originales* et d'appliquer d'autres mesures de proximité. Nous envisageons également, afin de mesurer la qualité et la cohérence des concepts formés par nos approches, de les soumettre à un expert.

Enfin, certains mots qui sont regroupés peuvent avoir des relations sémantiques qui diffèrent des liens de synonymie. Dans nos futurs travaux présentés dans le chapitre 4, nous souhaitons nous intéresser à ces autres liens sémantiques (notamment des liens d'hyperonymie/hyponymie). Dans ce cadre, nous proposerons d'utiliser, entre autres, un certain nombre de marqueurs linguistiques qui ont commencé à être étudiés en section 2.2.2.2 du chapitre 2.

Après avoir étudié l'*induction de relations verbales*, nous allons nous intéresser à l'*induction de termes nominaux*. Dans ces travaux, nous ne souhaitons pas seulement construire des syntagmes pertinents. En effet, l'objectif est aussi de proposer des syntagmes qui puissent interpeller (accrocher) le lecteur dans le cadre du titrage automatique. C'est la raison pour laquelle, dans la section 3.4, nous avons proposé une méthode de construction automatique de syntagmes qui se décline également en deux phases : construction des candidats (cf. section 3.4.2), filtrage de ces candidats (c.f. section 3.4.3).

## 3.4 Induction des syntagmes pour le titrage automatique

Ce travail de construction de titres courts issu de la Thèse de Cédric Lopez, fait suite aux précédents travaux consistant à extraire les syntagmes pertinents dans les textes qui peuvent se révéler pertinents pour constituer un titre (voir section 2.2.2.1 du chapitre 2). L'étude présentée ici, qui a été publiée dans [17, 15], consiste à construire automatiquement des titres sur la base de différents mots répartis dans les données textuelles. La section ci-dessous s'appuie essentiellement sur le travail présenté à TALN'2011 (Conférence sur le Traitement Automatique des Langues Naturelles) [17].

### 3.4.1 Introduction et contexte

Dans cette section, nous proposons une approche de construction automatique de titres courts (TC) français par des méthodes de Fouille du Web. À partir de patrons syntaxiques issus de nos analyses statistiques portées sur les titres réels, nous formons des TC candidats. Le principal problème rencontré est que plusieurs TC peuvent être pertinents pour un même texte (ou section de texte). Ils peuvent varier en fonction de leur taille (en nombre de mots), de leur forme ou bien du sujet mis en avant. Les TC candidats seront donc soumis à une validation en deux phases : (1) cohérence des candidats par rapport au texte, (2) cohérence des candidats par rapport au web. Les candidats font ensuite l'objet d'une contextualisation dynamique, indiquant ainsi le titre candidat le plus pertinent pour la partie de texte traitée. L'évaluation indique que les TC déterminés par notre approche sont pertinents.

L'objectif de l'approche de construction automatique de titres appelée Catit est de proposer des titres pertinents, en relation avec le contenu sémantique du texte à titrer. La section 3.4.4 discute d'un autre critère important à prendre en compte dans ce domaine d'étude, l'*accroche*.

Nous proposons un processus global composé de différentes étapes (cf. figure 3.5) qui

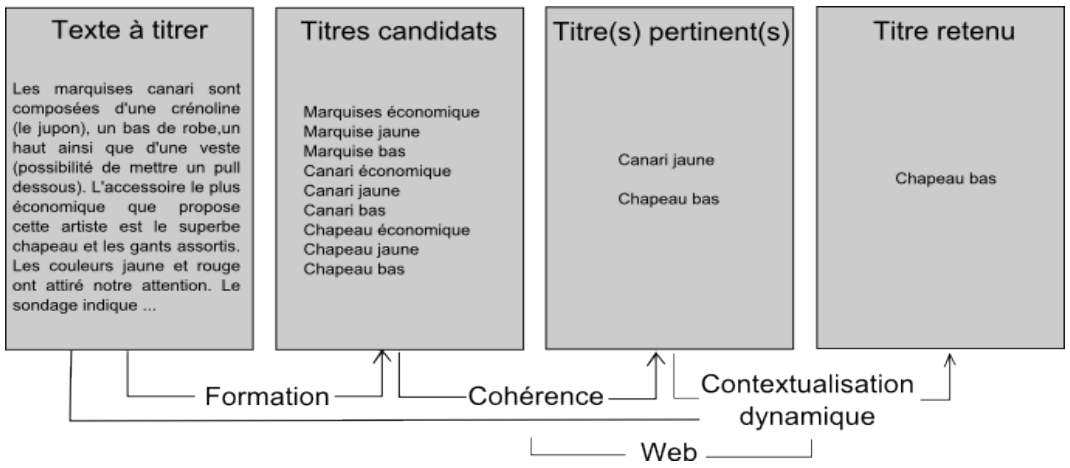


FIGURE 3.5 – Processus global de titrage automatique

seront détaillées dans les sections suivantes :

1. Formation des titres candidats (section 3.4.2) : Un ensemble de titres candidats est proposé automatiquement à partir des données extraites du texte et respectant les patrons syntaxiques déterminés lors de nos études préliminaires (section 3.4.2.1).
2. Cohérence des titres candidats (section 3.4.3.1) : Parmi les titres candidats formés à l'étape précédente, nous nous intéressons à leur cohérence par rapport au texte à titrer ainsi qu'à leur cohérence par rapport au Web.
3. Contextualisation dynamique des titres candidats (section 3.4.3.2) : Le contexte du texte et le contexte web de chaque titre candidat sont comparés afin de sélectionner le plus pertinent.

### 3.4.2 Construction automatique de titres courts

#### 3.4.2.1 Analyse préliminaire

Avant de construire les titres courts une analyse préliminaire permet de déterminer les patrons les plus adaptés.

Les articles journalistiques contiennent des sous-titres pouvant être simplement informatifs, mais aussi utilisant la présence de tournures humoristiques, l'emploi d'expressions, de citations. Nous considérons les sous-titres d'articles journalistiques comme des titres courts. Ainsi, notre étude statistique est réalisée sur les sous-titres d'articles journalistiques afin de déterminer ces patrons.

La base de données Factiva rassemble le texte intégral de plus de 8000 sources parmi lesquelles Le Monde est à disposition. Notre corpus d'étude a été constitué à partir de Factiva, sélectionnant 200 articles journalistiques français issus du quotidien Le Monde (novembre 2010) et contenant au moins un sous-titre. Afin que les résultats ne soient

pas biaisés par les éventuelles erreurs provoquées par le choix d'un étiqueteur morpho-syntaxique, les sous-titres ont été analysés manuellement, selon 4 patrons morphosyntaxiques contenant des noms communs (NC), adjectifs (ADJ) et mots outils (MO : articles, déterminants, prépositions, etc).

- 12% des sous-titres sont de la forme "$NC$" (ex. : " Objectifs ")
- 43% des sous-titres sont de la forme "$NC\ \ ADJ$" ou "$ADJ\ \ NC$" (ex. : "Paramètres sociopolitiques")
- 14% des sous-titres sont de la forme "$NC\ \ MO\ \ NC$" (ex. : "Hausse du budget")
- 26% des sous-titres contiennent quatre mots ou plus (ex. : "Les villepinistes s'élèvent contre la décision")

Compte tenu de ces résultats, nous décidons de nous intéresser plus particulièrement à la construction automatique de titres de la forme "$NC\ \ ADJ$" et "$ADJ\ \ NC$", qui couvrent 43% des sous-titres (ST) d'articles journalistiques issus de Le Monde. La section suivante consiste à construire des titres candidats de la forme "$NC\ \ ADJ$" et "$ADJ\ \ NC$".

#### 3.4.2.2 Combinaison d'approches statistiques et grammaticales pour la construction de candidats au titrage

La formation des titres candidats s'appuie sur le score TF-IDF (cf. section 2.1.1 du chapitre 2). Dans la suite, on notera $TF-IDF_X$ la valeur du TF-IDF obtenue pour $X$.

L'objectif est d'extraire les noms communs (NC) et adjectifs (ADJ) pertinents du texte à titrer. Après étiquettage du texte non lemmatisé [5] via le TreeTagger, à chaque nom commun extrait est attribué un score correspondant au TF-IDF (noté $TF-IDF_{NC}$), permettant de classer les noms communs (NC) par ordre de pertinence, de "saillance". En revanche, à chaque adjectif (ADJ) extrait est attribué un score correspondant au TF simple ($TF_{ADJ}$). En effet, moins l'adjectif est spécifique, et plus la probablilité qu'il puisse être le qualificatif d'un nom commun est élevée.

Dans le texte à titrer, les trois noms communs de plus haut TF-IDF et les dix adjectifs de plus haut TF sont extraits. Cette limite est due au nombre de requêtes limitées sur les moteurs de recherche [Keller and Lapata, 2003].

Soit $i$ le nombre de NC retenus ($i \in [\![1;3]\!]$) et $j$ le nombre de ADJ retenus ($j \in [\![1;10]\!]$). Tous les couples "$NC_i\ \ ADJ_j$" sont alors construits, un maximum de 30 titres candidats sont alors proposés. Parmi eux, tous ne sont pas cohérents, en particulier concernant la grammaticalité (ex. : "chapeau belle"). La section suivante permet de déterminer la cohérence des titres candidats.

### 3.4.3 Approche CATIT pour la sélection des titres candidats

Alors que des couples potentiellement pertinents ont été construits, il est alors nécessaire de sélectionner les titres les plus adaptés.

Dans une premier temps, deux critères de sélection statistiques seront appliqués pour

5. Nous verrons dans la suite qu'il est primordial de ne pas lemmatiser dans notre cas

évaluer une **certaine forme de cohérence** des titres construits (section 3.4.3.1). Par la suite, des sélections fondées sur des informations contextuelles seront mises en place pour sélectionner les titres les plus adaptés à la thématique des textes à titrer (section 3.4.3.2).

#### 3.4.3.1 Sélection statistique

**Méthode fondée sur le positionnement**

La pertinence des termes composant chaque titre candidat par rapport au texte est assurée par l'utilisation du TF-IDF lors de leur formation (cf. section 3.4.2). De cette façon, les noms communs et adjectifs les plus pertinents pour le titrage sont extraits.

Nous utilisons un autre critère de cohérence des titres candidats par rapport au texte, qui est la distance (en nombre de mots) entre les NC et les ADJ. Cette distance, notée $Dist_{NC-ADJ}$, est calculée pour chaque candidat puis utilisée dans le calcul du coefficient de distance (formule (3.2)).

$$Coef_{Dist} = \frac{1}{1 + Dist_{NC-ADJ}} \tag{3.2}$$

Si dans le texte, le candidat "$NC \quad ADJ$" apparaît, on aura $Dist_{NC-ADJ} = 0$ et $Coef_{Dist}$ atteindra son maximum. Le candidat "$NC \quad ADJ$" sera donc privilégié pour son utilisation en tant que titre. Cette distance est appliquée en tant que coefficient au score défini pour chaque candidat dans la suite.

**Méthode fondée sur la méthode** $DeMT$

Un critère de cohérence par rapport au Web permet de valider la cohérence des titres candidats (TC) en se fondant sur le Web (cf. mesures décrites dans le chapitre 1). Cette méthode permet notamment de mesurer la dépendance entre le nom commun et l'adjectif composant un titre candidat, d'où l'intérêt que ces derniers ne soient pas lemmatisés. On privilégie ainsi automatiquement un couple "$NC \quad ADJ$" bien construit (par exemple, "chapeau bas") par rapport à un couple mal construit (par exemple, *chapeau basse*), cette dépendance entre nom et adjectif sur le Web étant largement induite par les accords en genre et en nombre entre ces termes. Mais outre ces situations qui peuvent être identifiées par des règles grammaticales classiques, l'objectif reste plus large afin de valider la dépendance sémantique entre le nom et l'adjectif.

De manière plus précise, afin de mesurer cette dépendance, nous avons appliqué $DeMT_{Dice}$, $DeMT_{IM}$ et $DeMT_{IM3}$ décrites dans le chapitre 1. Le numérateur de ces fonctions de rang correspond alors au nombre d'occurrences des candidats de la forme "NC ADJ", noté *nb*. Cette dernière fonction sera également expérimentée dans la suite de cette étude.

La comparaison de ces trois mesures est effectuée à partir de 20 articles journalistiques issus de Le Monde. Pour chaque article, 30 titres candidats de la forme "$NC \quad ADJ$" ont

été formés. Les scores indiqués dans le tableau 1 correspondent au nombre de titre(s) candidat(s) à la fois pertinent(s) par rapport au texte et grammaticalement corrects, parmi les cinq de plus haut score. Au total, ce sont 400 titres qui ont été manuellement expertisés.

| Mesures | $DeMT_{Dice}$ | $DeMT_{IM}$ | $DeMT_{IM3}$ | $nb$ |
|---|---|---|---|---|
| Total | 42 | 36 | 41 | 32 |

TABLE 3.11 – Évaluation de la cohérence des résultats selon différentes mesures.

Cette évaluation indique que $Dice$, $IM$ et $IM^3$ obtiennent des résultats parfois proches avec toutefois un meilleur résultat pour $Dice$ et $IM^3$. La simple utilisation du nombre de résultats bruts retournés par Google est la moins performante par rapport à notre application. Compte tenu de ces résultats, nous choisissons la mesure de $DeMT_{Dice}$ dans la suite de notre travail.

Afin de prendre en compte les titres de la forme "$ADJ\ \ NC$", nous retenons la valeur maximum obtenue entre $DeMT_{Dice}(ADJ, NC)$ et $DeMT_{Dice}(NC, ADJ)$. Par exemple, on retiendra *beau chapeau* plutôt que "chapeau beau", le premier obtenant un score plus élevé que le second.

Finalement, plusieurs candidats cohérents par rapport au texte et par rapport au Web, peuvent arriver en tête du classement. Parmi ces titres candidats nous devons déterminer quel est le plus pertinent pour son utilisation en tant que titre, en tenant compte du contexte de chacun d'entre eux.

#### 3.4.3.2 Contextualisation dynamique

Pour un même document, plusieurs titres candidats peuvent être proposés. Afin de déterminer le titre le plus pertinent, nous comparons le contexte du texte à titrer avec le contexte dans lequel se retrouvent ces candidats sur le Web. Suite à la soumission d'une requête (via une API Google), le moteur de recherche Google présente les résultats sous forme d'une liste de sites Web. Pour chacun de ces sites, un aperçu du contenu de la page web est présenté (entre 10 et 30 mots), justifiant le résultat retourné en mettant en gras les termes initialement présents dans la requête. Le document utilisé pour la détermination du contexte Web de chaque titre candidat est la concaténation des 10 premiers aperçus (limite imposée par Google) d'une requête donnée. En ce qui concerne le contexte du texte, il est déterminé à partir du texte à titrer.

Pour déterminer le contexte Web et le contexte du texte, nous utilisons le modèle vectoriel de Salton [Salton et al., 1975]. Pour chaque nom commun et adjectif des documents (texte et document web), on détermine le TF qui constitue les coordonnées du vecteur contextuel (VCT pour le texte et VCW pour le Web). Finalement, à chaque titre candidat est associé un VCW. Si le vocabulaire associé à un contexte de titre candidat (VCW) est proche du vocabulaire du texte à titrer (VCT), alors nous privilégions ce candidat.

Pour chaque titre candidat, la similarité cosinus est utilisée entre deux vecteurs couvrant tous les couples possibles de la forme ($VCT_{Texte}$, $VCW_{Cand}$). Ainsi, les candidats

retenus sont ceux dont le contexte textuel est le plus "proche" du contexte Web.

Dans la section suivante, nous proposons une mesure globale réunissant la notion de cohérence des titres candidats et de contextualisation.

#### 3.4.3.3 Mesure globale

En s'appuyant sur les méthodes précédemment définies, nous mettons en place une mesure globale CATIT (Construction Automatique de TITres), permettant de mettre en avant les titres pertinents par rapport aux titres non pertinents, tenant compte à la fois de la cohérence des titres candidats par rapport au Web et au texte, ainsi que de leur contexte (dynamique). Cette mesure globale permet de fournir une fonction de rang globale prenant en compte tous les concepts.

Soit $TI_{Cand}$ la fonction appliquée à un titre candidat, qui est le produit du TF-IDF du nom commun (NC) et du TF-IDF de l'adjectif (ADJ) (formule (3.3)).

$$TI_{Cand} = TF.IDF_{NC_{Cand}} \times TF.IDF_{ADJ_{Cand}} \tag{3.3}$$

La prise en considération de $TI_{Cand}$ dans CATIT permet de tenir compte de la pertinence de l'information contenu dans les termes composant les titres candidats.

$$CATIT(Cand) = \begin{cases} Coef_{Dist} \times TI_{Cand} \times log_2(1 + cos(VCT_{Texte}, VCW_{Cand})), \\ \qquad\qquad si\ DeMT_{Dice}(Cand) > K \\ \\ Coef_{Dist} \times TI_{Cand} \times log_2(DeMT_{Dice}(Cand)) \\ \qquad\qquad sinon. \end{cases}$$

$DeMT_{Dice}(Cand)$ est toujours compris entre 0 et 1. Ainsi, avec l'utilisation de la fonction logarithme, les titres incohérents (inférieurs au seuil $K \in \mathbb{R}$ comparé à la mesure Web) seront toujours négatifs. Par ailleurs, le classement des candidats (via $DeMT_{Dice}$) des titres négatifs sera aussi maintenu, grâce au $log_2$ qui est une fonction strictement croissante ($Coef_{Dist}$ et $TI_{Cand}$ sont toujours positifs).
Au contraire, les titres cohérents (supérieurs au seuil $K$) seront toujours positifs, grâce au 1 qui correspond au maximum de la valeur $DeMT_{Dice}$ possible.

Enfin, le classement par proximité contextuelle ($cos(VCT_{Texte}, VCW_{Cand})$) respecte l'ordre établi par le cosinus. Nous faisons intervenir la distance $Coef_{Dist}$ permettant de privilégier, parmi les candidats cohérents et contextuellement pertinents, ceux qui sont constitués de termes proches dans le texte. Finalement, les titres candidats obtenant un résultat positif sont jugés pertinents par notre mesure. Le candidat obtenant le plus haut score est retenu pour son utilisation en tant que titre.

Le choix du seuil de pertinence $K$ est crucial. Celui ci est discuté dans la section suivante. Avec une valeur adaptée, nous évaluons notre mesure CATIT .

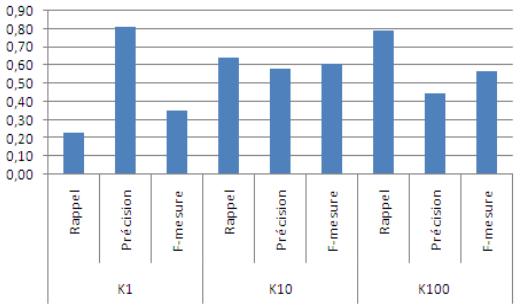


FIGURE 3.6 – Détermination de $K$

### 3.4.4 Expérimentations

Cette section est dédiée à l'évaluation des titres construits par notre approche selon plusieurs critères. Les évaluations permettant de déterminer le seuil $K$ puis la pertinence de notre approche CATIT ont été effectuées par le doctorant directement concerné par ce travail. Notons qu'une application dédiée à l'évaluation de nos différentes méthodes de titrage est en cours développement. Elle permettra de confronter les titres fournis par les méthodes POSTIT (section 2.2.2.1) et CATIT (section 3.4) à différents évaluateurs [Eugenio and Glass, 2004].

#### 3.4.4.1 Détermination du seuil $K$

Les résultats apportés par la mesure CATIT dépendent fortement du seuil de pertinence $K$. Le comportement de ce seuil est analysé à partir des 10 premiers articles parus le 1er janvier 1994 dans le quotidien Le Monde, soient 900 titres évalués manuellement [6]. On ne cherchera pas à juger l'acceptabilité des trente candidats mais seulement leur grammaticalité. Différents seuils $K_N$ sont testés (avec $N \in \{1, 10, 100\}$), fondés sur la moyenne des valeurs retournées par la mesure de Dice (formule (3.4)).

Cette détermination de $K$ s'appuie sur la précision, le rappel et la F-mesure. Dans le cadre de ces mesures, un titre acceptable est un titre grammaticalement correct. Les résultats sont présentés à la figure 3.6.

$$K_N = \frac{moy(DeMT_{Dice}(Cand))}{N} \tag{3.4}$$

L'utilisation du seuil $K_1$ n'est pas pertinente pour notre mesure car son utilisation entraînerait un élagage prématuré de nombreux candidats pouvant se révéler pertinents (précision élevée mais rappel faible). De même, l'utilisation du seuil $K_{100}$ n'est pas pertinente pour notre mesure car de nombreux candidats incohérents sont conservés (précision faible mais rappel élevé). Finalement, les résultats (Figure 3.6) indiquent que le meilleur compromis entre précision et rappel est atteint avec $K_{10}$. Dans la suite, nous utiliserons donc le seuil $K_{10}$, que nous appliquerons lors de l'évaluation de CATIT.

6. 30 titres candidats × 10 articles × 3 seuils $K$

#### 3.4.4.2 Évaluation de CATIT

Les titres construits automatiquement doivent répondre aux mêmes caractéristiques que les titres réels. Le premier critère concerne l'information transmise par le titre, qui doit être en relation avec le texte traité. Si ce critère est constaté, nous conclurons que le titre est informatif (noté I). Le second critère concerne l'accroche. Un titre sera jugé accrocheur (noté A) s'il contient une tournure humoristique, une expression ou autre construction surprenant le lecteur, grammaticalement correct et informatif (en relation avec le texte). En effet, il ne sera pas convenable de juger un titre accrocheur s'il n'est pas en relation avec le texte. Par exemple, le titre "Chapeau bas" peut être considéré comme étant informatif (dans cet exemple, le texte rend hommage à un couturier qui propose entre autre des chapeaux) et accrocheur (emploi d'expression). Si le texte ne traitait pas de chapeaux et qu'il n'avait rien à voir avec l'expression "chapeau bas", on ne pourrait pas considérer le titre "chapeau bas" comme étant accrocheur, bien qu'il s'agisse d'une expression. Finalement, un objectif de cette évaluation est de détecter si nos titres sont "pertinemment accrocheurs".

Afin de tenir compte de ces critères dans l'évaluation, nous calculons la précision et et le rappel adaptés aux critères A et I prédéfinis (formules (3.6) et (3.5)). La F-mesure est également calculée [7].

$$Précision_{I(resp.A)} = \frac{Nb\ de\ titres\ I(resp.\ A)\ retenus}{Nombre\ total\ de\ titres\ retenus} \tag{3.5}$$

$$Rappel_{I(resp.A)} = \frac{Nb\ de\ titres\ I\ (resp.\ A)\ retenus}{Nb\ total\ de\ titres\ I(resp.A)} \tag{3.6}$$

L'évaluation est effectuée à partir d'articles journalistiques issus du journal quotidien Le Monde. Nous avons retenu les 20 premiers articles publiés le 1er janvier 1994. Ainsi, ce sont 600 titres issus de notre méthode CATIT, utilisant le seuil $K_{10}$ qui ont été évalués manuellement en fonction de I et A (soit 1200 expertises au total). 1460 requêtes sur le moteur de recherche ont été nécessaires.

Les résultats de cette évaluation concernant la précision et le rappel sont présentés en Table 3.12. En plus, pour chaque article, le titre de plus haut score retourné par CATIT , noté T1, est évalué. Nous notons "oui" lorsque le critère est respecté et "non" sinon. La présence du symbole "ensemble vide" indique qu'aucun titre parmi les 30 titres candidats correspond au critère demandé. Par exemple, parmi les 30 candidats construits à partir de l'article 1, aucun est informatif ou pertinent.

En ce qui concerne les titres informatifs, ils obtiennent une précision de 0.40 compensée par un rappel de 0.82 (cf. Table 3.12). Puisque les titres T1 sont informatifs dans 75% des cas (cf. Table 3.12), nous pouvons en déduire que le seuil $K$ doit être affiné afin de retenir moins de titres candidats. Le moteur de recherche Google ne tenant pas compte de la présence de ponctuation dans les requêtes, un taux élevé de candidats constituent un bruit non négligeable. Un exemple directement lié à ce problème est le titre T1 de l'article 14 qui est mal construit (*Conditionnelle Peines*). Notons que pour cet article,

7. Nous utilisons la formule générique avec $\beta = 1$

le deuxième titre de plus haut score est *Peines symboliques* qui est informatif et accrocheur. Par ailleurs, une erreur de la part de l'étiqueteur impacte fortement les résultats, surtout s'il s'agit d'une erreur concernant la détermination des trois noms communs (qui se répercute alors sur 10 titres candidats).

Du côté des titres accrocheurs, les mêmes difficultés sont rencontrées. De plus, nous avons constaté que très peu de candidats accrocheurs (maximum deux par article dans cette évaluation) sont construits, problème lié au nombre limité de noms communs et adjectifs retenus à la première étape de notre approche. Ceci explique une précision faible et un rappel élevé. Notons tout de même que, malgré la relative rareté des titres candidats accrocheurs, 30% des titres construits par notre méthode sont accrocheurs.

Enfin, l'évaluation indique que 75% des titres T1 construits automatiquement par Catit sont informatifs et 30% sont accrocheurs (cf. Table 3.12). Ainsi, parmi les titres informatifs proposés, 40% sont accrocheurs, ce qui constitue un point positif pour notre approche. Finalement, nous comparons les titres T1 déterminés selon la méthode de titrage par Extraction de Syntagmes Nominaux (ESN) [14]. L'évaluation de ESN indique que seulement 60% des titres de la forme "nom adjectif" ou "adjectif nom" sont informatifs et 5% sont accrocheurs (cf. Table 3.13).

| | T1 | | Rappel | | Précision | | F-mesure | |
|---|---|---|---|---|---|---|---|---|
| | I | A | I | A | I | A | I | A |
| Article 1 | non | non | x | x | x | x | x | x |
| Article 2 | oui | oui | 0,75 | 0,50 | 0,50 | 0,33 | 0,60 | 0,40 |
| Article 3 | oui | oui | 1,00 | 1,00 | 0,21 | 0,14 | 0,35 | 0,25 |
| Article 4 | oui | non | 1,00 | 1,00 | 0,31 | 0,31 | 0,48 | 0,47 |
| Article 5 | oui | x | 0,86 | x | 0,50 | x | 0,63 | x |
| Article 6 | oui | oui | 0,83 | 1,00 | 0,50 | 0,40 | 0,63 | 0,57 |
| Article 7 | oui | x | 0,80 | x | 0,22 | x | 0,35 | x |
| Article 8 | oui | oui | 0,67 | 1,00 | 0,57 | 0,17 | 0,62 | 0,29 |
| Article 9 | oui | x | 1,00 | x | 0,38 | x | 0,55 | x |
| Article 10 | oui | x | 0,89 | x | 0,47 | x | 0,62 | x |
| Article 11 | non | non | 0,89 | x | 0,53 | x | 0,67 | x |
| Article 12 | oui | non | 1,00 | x | 0,33 | x | 0,50 | x |
| Article 13 | oui | oui | 0,83 | 1,00 | 1,00 | 0,20 | 0,91 | 0,33 |
| Article 14 | non | non | 0,75 | 0,50 | 0,33 | 0,11 | 0,46 | 0,18 |
| Article 15 | oui | non | 0,75 | x | 0,21 | x | 0,33 | x |
| Article 16 | non | non | x | x | x | x | x | x |
| Article 17 | non | non | 0,50 | x | 0,10 | x | 0,17 | x |
| Article 18 | oui | oui | 0,50 | 1,00 | 0,25 | 0,13 | 0,33 | 0,22 |
| Article 19 | oui | non | 0,80 | 1,00 | 0,44 | 0,11 | 0,57 | 0,20 |
| Article 20 | oui | non | 1,00 | x | 0,27 | x | 0,43 | x |
| Total | 75% | 30% | 0,82 | 0,89 | 0,40 | 0,21 | 0,51 | 0,32 |

Table 3.12 – Evaluation de Catit

| | ESN | | Catit | |
|---|---|---|---|---|
| Critères | I | A | I | A |
| Évaluation | 60% | 5% | 75% | 30% |

Table 3.13 – ESN versus Catit

### 3.4.5 Discussion

La construction automatique de titres est une tâche complexe car des titres à la fois cohérents, grammaticalement corrects, informatifs et accrocheurs doivent être construits puis choisis parmi une liste de titres ne respectant pas ces critères.

Pour ce type d'approche nécessitant une évaluation cognitive des résultats obtenus, un protocole d'évaluation automatique est extrêmement difficile à mettre en place. À titre d'exemple, comment peut-on évaluer automatiquement qu'un titre est accrocheur ? L'évaluation de la pertinence est tout aussi complexe et dépend souvent du domaine. Par exemple, les titres de courriels proposés par une méthode automatiques peuvent se révéler plus pertinents que les titres réels [14].

Nous avons donc préféré mettre en œuvre un protocole d'évaluation manuel qui se révèle beaucoup plus adapté dans le cadre de ce travail sur le titrage. Une application est en cours d'exécution [8] pour permettre à plusieurs évaluateurs de juger la pertinence et l'accroche des syntagmes proposés par notre système d'induction. La phase d'évaluation a commencé début octobre 2011 et rassemble déjà les avis de 65 participants. Le dépouillement des résultats sera effectué dans quelques semaines.

Les futurs travaux que nous souhaitons menons s'appuiront sur les structures verbales. En effet, dans certaines phrases les verbes présentent une information tout à fait essentielle à mettre en exergue (exemple, *"Le président a démissionné"*). Pour construire un titre nominal véhiculant l'information verbale, nous proposons de mettre en œuvre un processus de nominalisation du verbe qui devient la tête du syntagme construit (exemple, *"a démissionné"* $\rightarrow$ *"démission"*), puis de considérer le sujet comme modifieur. Nous ajouterons enfin les prépositions qui seront validées par des approches de fouille du web (mesure $DeMT$ présentée dans le chapitre 1). Ceci permettra de construire le titre *Démission du président"* issue du fragment *"Le président a démissionné"*. Par ailleurs la gestion des Entités Nommées est indispensable pour consolider ces travaux en cours d'étude.

## 3.5 Discussion générale

En introduction de ce mémoire, nous avons précisé que les applications de fouille de textes peuvent être mises en œuvre en corpus (cf. chapitre 2) ou hors corpus (cf. chapitre 1). Dans cette étude liée aux syntagmes induits, nous rencontrons également une telle situation. En effet, le processus d'induction dans le cadre de la découverte d'Entités Nommées (cf. section 3.2) est complètement indépendant des corpus d'où sont issus les termes. A contrario, l'induction des syntagmes pour la classification conceptuelle (cf. section 3.3) et le titrage (cf. section 3.4.2) est très dépendante des corpus.

Dans le chapitre d'introduction à ce mémoire, nous avons également montré de quelle manière les approches de fouille de textes peuvent utiliser des informations endogènes et/ou exogènes. Dans le cas des inductions syntagmatiques qui proposent des associations

8. http ://www.lirmm.fr/∼lopez/Titrage_general/evaluation_web2/

hypothétiques, seules des informations exogènes peuvent être exploitées pour les valider. Pour cette tâche, nos méthodes utilisent des informations Web via la mesure $DeMT$ qui a été adaptée aux différentes problématiques présentées dans ce chapitre.

Afin de construire les associations hypothétiques de base (par exemple, dans le cadre de la construction des titres), nous associons des termes qui ne sont pas naturellement en relation. Une limite de cette approche provient du fait que ces associations sont seulement construites à partir des termes présents dans les données textuelles. Actuellement, nous étendons nos méthodes afin de prendre en compte de nouvelles connaissances exogènes afin d'être moins dépendant des corpus. Dans ce cadre, des méthodes de titrage fondées sur le principe de reformulation en exploitant de nouvelles connaissances sémantiques sont en cours d'étude.

# Chapitre 4

# Conclusion et Perspectives

| *Thèmes de Recherche* | *Types de travaux* | *Années* |
|---|---|---|
| Extraction (80%) | | Exploitation (20%) |
| Analyse de données SMS | **Projet CS/MSH-M/UM3 – *SMS4SCIENCES***<br>membre | 2011-2012 |
| Extraction (70%) | | Exploitation (30%) |
| Classification de données OCR | **Collaboration industrielle** – ITESOFT (Aimargues)<br>responsable scientifique<br>*Stage M2 Recherche* | 2007-2008 |
| Extraction (50%) | | Exploitation (50%) |
| Cube de textes | **Collaboration académique** – LIRMM/CEMAGREF<br>*Stage Ingénieur CNAM* | 2009-2011 |

## 4.1 Bilan

En introduction de ce mémoire, les processus de fouille de textes ont été placés sous l'angle de deux phases successives de traitement : l'*extraction* et l'*exploitation* des descripteurs linguistiques. Bien que les contributions de recherche soient souvent focalisées sur une des deux phases, afin de mettre en place un processus complet d'une application ciblée, il est souvent nécessaire de traiter rigoureusement ces deux phases.

Dans ce mémoire, le traitement des descripteurs linguistiques a été discuté selon leur situation ou non en corpus (chapitres 1 et 2). Pourtant, il n'est pas rare que les approches de traitement soient parfaitement entremêlées. Par exemple, lorsque que des méthodes de traitement de chaînes de caractères sont appliquées, les connaissances contextuelles liées aux corpus permettent d'enrichir les informations initiales. Il en est de même pour l'utilisation de connaissances *endogènes* et *exogènes* décrites dans le chapitre 1. La combinaison de l'ensemble de ces approches est parfaitement bien illustrée dans le chapitre 3 qui prend en compte toutes les techniques proposées dans les chapitres 1 et 2.

Les travaux présentés en perspectives ont trait à l'étude de textes réputés complexes. Outre le traitement des descripteurs linguistiques afin de dégager des liens sémantiques entre eux (cf. section 4.2), nous proposons de nous intéresser également à leur organisation sous forme d'entrepôts de données (cf. section 4.3) afin de découvrir des connaissances nouvelles.

Dans le travail prospectif proposé dans ce chapitre, la première contribution (section

4.2) se focalise clairement sur la première phase de ce processus (phase d'extraction). La contribution du travail lié à l'organisation des données textuelles dans les entrepôts est quant à elle orientée sur la problématique de l'exploitation des descripteurs, bien que leur extraction au préalable soit nécessaire.

Enfin, dans nos futurs travaux, nous souhaitons, également, fouiller les données au sein d'un paradigme plus large en y intégrant des données autres que textuelles (cf. section 4.4). Ce travail d'envergure, qui reposera sur des thématiques de recherche nouvelles, nécessitera des contributions aux niveaux des deux phases d'un processus de fouille de textes.

## 4.2 Nouveaux descripteurs

Ce mémoire met en relief le fait que l'identification des descripteurs pertinents est une étape souvent décisive pour des tâches de fouille de textes. Ceci est un problème difficile pour des données textuelles complexes dont quelques exemples sont décrits ci-dessous.

Dans le cadre de la thèse d'Hassan Saneifar, nous avons utilisé des méthodes d'apprentissage supervisé permettant de déterminer les segments logiques à partir des données logs [35]. Cette étape préliminaire est cruciale pour les tâches de Recherche d'Information décrites dans le chapitre 2. Pour appliquer les méthodes de détection de segments dans les fichiers logs, il est nécessaire d'avoir une représentation adaptée de ces données textuelles spécifiques. Dans ce cadre, nous avons proposé une représentation de type n-grammes de caractères (cf. chapitre 1). Classiquement, les n-grammes s'appuient sur le contenu même des données textuelles (les mots ou les lettres). Dans le contexte de l'étude des fichiers logs, une telle représentation se révèle trop spécifique afin de déterminer les différents segments. Nous avons alors défini un nouveau type de n-grammes appelé vs-grammes généralisés qui s'appuie essentiellement sur la représentation des informations structurelles des fichiers (ponctuations, symboles, mise en page, etc).

Au cours d'une collaboration avec la société ITESOFT et l'encadrement d'un stage Recherche, nous nous sommes également intéressés à des types de données nécessitant le traitement sous forme de n-grammes de caractères. L'objectif de ces travaux était de classer des documents numériques issus d'un processus de rétro-conversion d'OCR (Reconnaissance Optique de Caractères) [Junker and Hoch, 1997] qui sont souvent bruités. Ces derniers doivent être répartis en plusieurs catégories comme les factures ou les attestations. Nous avons alors proposé la méthode HYBRED (HYBrid REpresentation of Documents) [12] pour représenter les textes issus d'un traitement OCR et qui permet d'améliorer les méthodes de classification comme les machines à support vectoriel (SVM) [Joachims, 1998]. La première étape de la méthode consiste à sélectionner les mots ayant une étiquette grammaticale spécifique (par exemple, noms, adjectifs, verbes) qui sont souvent davantage porteurs de sens pour des tâches de classification thématique [1]. L'étape suivante consiste à déterminer des frontières. Ce concept de frontière fait référence aux travaux de [Bourigault, 1993] qui effectuent une extraction de syntagmes nominaux sur la base d'un repérage préliminaire de marqueurs linguistiques (par exemple, "préposition +

adjectif possessif", "préposition + article indéfini"). Dans notre cas, les mots apportant peu d'informations déterminent les frontières. Même si notre méthode est différente, l'objectif reste le même que [Bourigault, 1993] car nous cherchons à prendre en considération les groupes de mots pertinents situés entre les frontières. L'étape suivante de l'approche HYBRED extrait les n-grammes de caractères (cf. chapitre 1) situés entre les frontières. Le choix de l'utilisation des n-grammes de caractères dans ce contexte, est lié au fait que cette approche est moins sensible au bruit dans les données [Jalam and Chauchat, 2002]. En effet, les documents que nous traitons comportent une part de bruit liée à diverses déformations lors de la reconnaissance de documents. Par exemple, il est possible que le mot *feuille* soit lu *teuille* par un outil OCR. Une approche fondée sur les 3-grammes prendra en compte les autres n-grammes tout à fait corrects ("eui", "uil", etc) même si le premier n-gramme ("teu") se révèle bruité. Enfin, la dernière étape de l'approche HYBRED effectue un filtrage statistique sur la base du TF-IDF (cf. chapitre 2).

Dans de futurs travaux, nous souhaitons davantage approfondir l'hybridation des différentes approches afin de traiter d'autres types de données bruitées et à faible contenu textuel tels que les tweets (projet en collaboration avec la société Web Report) et les SMS (projet sud4sciences [1] qui a débuté en 2011). Le vocabulaire utilisé dans ce type de message présente des spécificités lexicales, syntaxiques et graphiques. Par exemple, dans ces messages, le mot *pkoi* qui représente une forme contractée du mot *pourquoi*, est souvent fréquent. D'autres variantes lexicales, qui sont en général synonymes, sont également employées (*pkooi*, *pquoi*, etc). De telles connaissances sémantiques-lexicales sont tout à fait cruciales pour les méthodes d'étiquetage spécifiques aux tweets récemment proposées par [Gimpel et al., 2011].

Les mesures lexicales décrites dans le chapitre 1 sont inefficaces pour déterminer la synonymie des mots courts typiques des messages tweet et SMS. Ainsi, il est nécessaire de développer de nouvelles approches. Dans ce contexte, pour construire des dictionnaires spécifiques à ce type de données, certaines méthodes s'appuient sur l'exploitation de corpus alignés (corpus bruts et transcrits manuellement) [Fairon and Paumier, 2006, Beaufort et al., 2010]. Leur traitement est cependant une tâche difficile. En effet, la simple segmentation en *mot* peut se révéler complexe [2].

Après avoir identifié les mots, dans nos futurs travaux, nous souhaitons proposer des approches nouvelles afin de déterminer les liens sémantiques dans ce type de données complexes.

Dans le cadre du Web 2.0 et des communications de type SMS, pour déterminer la proximité lexicale entre des termes proches, nous souhaitons mettre en place des approches en deux étapes sur la même base que celles décrites dans le chapitre 3. La première étape consiste à déterminer les termes candidats proches en s'appuyant sur la proximité lexicale mais aussi phonétique (*café* et *kfé* par exemple). Ensuite des mesures

1. http ://sud4science.org/

2. Pour une telle tâche d'alignement, [Beaufort et al., 2010] définissent le mot comme la plus longue séquence qui ne possède pas le même séparateur de part et d'autre de l'alignement (après avoir effectué un alignement par caractère au préalable).

de fouille du Web décrites dans le chapitre 1 pourront être mises en place. Ces mesures qui exploitent les résultats retournés par les moteurs de recherche devront combiner les recherches exactes sur le Web (chaînes de caractères) aux opérateurs classiques ($AND$, $OR$). Une telle combinaison est essentielle, la première méthode fournissant, en général, une valeur de précision élevée alors que la seconde privilégie le rappel. Par ailleurs, la prise en compte d'informations temporelles [3] dans cette mesure peut être pertinente pour ce type de données parfaitement évolutive. Par exemple, le vocabulaire utilisé dans les SMS et tweets, est régulièrement enrichi lexicalement voire syntaxiquement. Ce point sera étudié en comparant les descripteurs des SMS collectés en 2004, 2008, 2009, 2010 dans le cadre du projet sms4sciences [4] avec (1) les SMS collectés cette année (du 15 septembre au 15 décembre 2011) dans notre Région [5], (2) nos données tweets disponibles.

Dans un premier temps , nous pourrons nous intéresser aux liens de synonymie entre les descripteurs. Par la suite, nous souhaitons dans le cadre d'une thèse qui a débuté en octobre 2011, nous intéresser aux différents types de liens sémantiques comme l'hyponymie/hyperonymie de manière plus globale. Pour cela, nous souhaitons nous appuyer sur des informations statistiques, linguistiques, temporelles tout en nous focalisant, dans un premier temps, sur les marqueurs étudiés dans le projet PEPS RESENS (cf. chapitre 2).

Dans ces différents travaux, nous projetons également d'exploiter les informations géographiques [Eisenstein et al., 2010] qui sont cruciales pour l'étude des données SMS ou tweet. Les liens qui existent entre *données textuelles* et *informations géographiques* seront prochainement discutées lors du workshop GeoDoc'2012 que nous organisons dans le cadre de la conférence PAKDD [6]. Dans ce contexte, nous avons étudié une nouvelle représentation des données textuelles pour en faciliter l'analyse en considérant différents types d'informations (géographiques, temporelles, et autres). Cette nouvelle représentation est décrite dans la section suivante.

## 4.3 Nouvelle représentation des données textuelles

Afin de faciliter l'analyse des données, nous proposons de stocker les textes dans des entrepôts (datawarehouse) [Codd et al., 1993]. Pour faciliter l'analyse, des opérateurs (Roll-Up, Drill-Down) permettent de naviguer au travers des hiérarchies et ainsi agréger les données en fonction des requêtes utilisateurs.

Récemment, de nouvelles approches fondées sur des cubes de textes proposent d'utiliser les technologies OLAP (On-Line Analytical Processing) pour analyser et extraire des connaissances [Lin et al., 2008, Pérez-Martínez et al., 2008, Zhang et al., 2009, Pujolle et al., 2008]. Dans cette même lignée, nous avons proposé un modèle permettant d'intégrer les textes dans des entrepôts de données textuelles [7]. Nous avons alors appliqué ce

3. De telles informations sont disponibles, par exemple, via le moteur de recherche Exalead.
4. http ://www.sms4science.org/
5. Plus de 30000 SMS ont déjà été récoltés en trois semaines.
6. http ://www.lirmm.fr/∼mroche/GeoDoc2012/

modèle aux données issues de twitter. Nous nous sommes plus particulièrement intéressés aux messages liés à la thématique de la Santé.

Dans un tel contexte, différents opérateurs de manipulation des données doivent être définis afin d'identifier des tendances, rechercher les top-k mots ou termes les plus significatifs sur une période de temps, les plus représentatifs d'une ville ou d'un pays, d'un certain mois, d'une certaine année. Les hiérarchies pour définir les informations géographiques et temporelles ont alors un fort impact dans la phase d'analyse. Ainsi, dans [7], nous avons proposé la mesure *TF-IDF adaptatif* qui permet d'identifier les mots les plus significatifs selon le niveau des hiérarchies du cube (par exemple, à partir de la dimension localisation). Ceci permet par exemple de mettre en exergue qu'un mot relatif à un symptôme, une maladie ou une épidémie peut se révéler discriminant pour certaines villes sans être caractéristique à un niveau de granularité plus élevé (par exemple en considérant un ensemble de pays). Illustrons cette situation avec l'exemple suivant. Au cours de l'année 2011, une épidémie liée à la bactérie *Escherichia Coli* s'est développée en Europe. Nous pouvons relever le fait que le terme *durchfall* (traduction du mot *diarrhées* en allemand) était sur-représenté dans des requêtes google dans les environs de Kiel (Ville du Nord de l'Allemagne) au milieu de l'année 2011 mais moins utilisé dans d'autres villes [7]. Ceci peut naturellement s'expliquer par le fait que la maladie était circonscrite à plusieurs villes du Nord de l'Allemagne. La mise en avant d'un critère de discriminance sur les mots relativement à un critère géographique peut, dans ce cas, être décisif pour les professionnels du domaine de la Santé et/ou de l'Épidémiologie.

La mesure de base proposée dans [7] devra être enrichie afin de proposer plusieurs types de navigation. Nous pourrons envisager une navigation en profondeur (par exemple, pour rechercher les tendances des villes d'un pays défini) ou en largeur (par exemple, pour analyser les tendances des villes par rapport au monde).

Par ailleurs, lors de la navigation plusieurs descripteurs peuvent être ambigus, c'est-à-dire être présents à différents endroits dans la hiérarchie. Dans nos travaux liés au domaine médical, nous avons rencontré un tel problème. En effet, pour l'analyse des tweets utilisant un vocabulaire médical, nous nous sommes appuyés sur le thesaurus MeSH (Medical Subject Headings) pour naviguer (généraliser/spécialiser les termes). Cependant, dans ce thésaurus de nombreuses instances sont polysémiques rendant la navigation difficile. Ainsi, nous avons proposé des méthodes de désambiguïsation qui utilisent la mesure *DeMTC* (cf. chapitre 1) afin de choisir à partir de quel nœud ou feuille appliquer nos opérateurs d'agrégation [7]. Pour une telle tâche de désambiguïsation, nous expérimentons actuellement l'utilisation d'informations contextuelles plus riches (cf. chapitre 2). Ceci engendre des difficultés dues au contexte réduit des tweets (140 caractères maximum).

Afin de mener à bien ces différents traitements pour modéliser au mieux nos entrepôts, il sera nécessaire d'avoir des connaissances lexicales et/ou syntaxiques propres aux tweets.

7. Informations issues de requêtes via *Google Trends* : http ://www.google.com/trends

Il sera alors essentiel de développer les premières perspectives proposées dans ce chapitre (cf. section 4.2).

Par ailleurs, pour proposer une analyse plus fine des données, nous souhaitons prendre en compte des aspects liés aux sentiments et donc intégrer une nouvelle dimension "sentiment" dans nos cubes de textes. Pour cela, nous devrons prendre en compte les différents dictionnaires liés au sentiment [Esuli and Sebastiani, 2006, Cambria et al., 2010] ou les construire selon les domaines d'étude (cf. section 1.2.3 du chapitre 1).

Enfin, notre modèle est focalisé sur un type de texte particulier. Il pourrait être utile de rassembler dans notre entrepôt de données textuelles différentes formes de documents (tweets, dépêches d'agence, articles journalistiques, rapports de l'INVS [8]). En effet, l'intégration de différents types de textes proposant des points de vues différents mais complémentaires ne peut qu'améliorer l'analyse. Dans un tel contexte, les mesures d'agrégation devront prendre en compte les caractéristiques de chaque type de textes (spécificités lexicales, stylistiques), taille des textes, etc.

De manière globale, il semble intéressant d'intégrer différentes données textuelles afin d'améliorer les tâches d'analyse. À moyen terme il semble également essentiel de prendre en compte d'autres types de données (images, vidéos, etc) comme nous le discuterons dans la section suivante.

## 4.4 Nouveau paradigme

Dans les travaux menés ces dernières années, je me suis plus spécifiquement focalisé sur l'extraction et l'induction de descripteurs linguistiques dans les textes. Je me suis alors attaché au traitement de textes réputés complexes.

Ainsi, dans les travaux menés avec la société Adamentium [9] dans les années 2006/2007, je me suis plus spécifiquement intéressé à la classification de textes dans le cadre de la problématique du *contrôle parental*. Bien que m'étant focalisé sur la classification de ces textes spécifiques, cette entreprise avait alors pour ambition de proposer des combinaisons d'approches liées aux images et textes. Cette problématique qui peut être étendue à un domaine plus vaste pourrait aussi associer d'autres supports.

En effet, avec le développement croissant du Web 2.0 et des données inhérentes, un autre mode de communication a vu le jour provoqué par le développement et surtout l'enlacement des différents supports de communication. En effet, aujourd'hui, les seuls descripteurs linguistiques présents au sein d'une ressource Web sont souvent nécessaires mais plus suffisants. Ainsi, il n'est pas rare que l'humain ne puisse pas analyser un texte (court) sans exploiter les informations associées (par exemple, les image, les vidéo, le son). Ainsi, il semble tout à fait pertinent de proposer de nouvelles approches de fouille

8. INstitut de Veille Sanitaire.
9. Société qui a été depuis rachetée.

qui associent l'extraction et l'exploitation de descripteurs par des méthodes de TAL aux descripteurs issus du traitement d'image, du traitement de la parole, etc. Dans ce cadre, les textes seront les données centrales à prendre en compte dans un processus global. Les différentes sources d'information permettront, par exemple, de placer un propos issu d'un texte dans son contexte. Ceci peut se révéler crucial pour la problématique de désambiguisation qui reste toujours sensible en fouille de textes.

Un tel travail que nous pouvons appeler *fouille de textes multimédia* peut être très difficile à mener car, outre la prise en compte des différentes données, il est nécessaire d'adapter et/ou combiner les méthodologies et traitements associés. À plusieurs reprises, dans ce manuscrit, nous avons fait référence à l'hétérogénéité des données textuelles. Ceci devient un verrou encore plus difficile si nous nous intéressons à la fouille de textes multimédia.

# Sélection de mes principales publications


[1] I. Bayoudh, N. Béchet, and M. Roche. Blog classification : Adding linguistic knowledge to improve the k-nn algorithm. In *Proceedings of Intelligent Information Processing (IIP)*, pages 68–77, 2008.

[2] N. Béchet, J. Chauché, V. Prince, and M. Roche. Corpus and Web : Two Allies in Building and Automatically Expanding Conceptual Classes. **Informatica**, 34(3) :279–286, 2010.

[3] N. Béchet, M. Roche, and J. Chauché. How the ExpLSA approach impacts the document classification tasks. In *Proceedings of IEEE International Conference on Digital Information Management (ICDIM)*, pages 241–246, 2008.

[4] N. Béchet, M. Roche, and J. Chauché. Comment valider automatiquement des relations syntaxiques induites. In *Actes de la conférence Extraction et Gestion des Connaissances (EGC), article nominé parmi les meilleurs articles académiques d'EGC'09*, pages 169–180, 2009.

[5] N. Béchet, M. Roche, and J. Chauché. Towards the Selection of Induced Syntactic Relations. In *Proceedings of 31st European Conference on Information Retrieval (ECIR), LNCS, Springer-Verlag (Poster)*, pages 786–790, 2009.

[6] Z. Bellahsene, S. Benbernou, H. Jaudoin, F. Pinet, O. Pivert, F. Toumani, S. Bernard, P. Colomb, R. Coletta, E. Coquery, F. De Marchi, F. Duchateau, M.-S. Hacid, A. HadjAli, and M. Roche. Forum : a flexible data integration system based on data semantics. **SIGMOD Record**, 39(2) :11–18, 2010.

[7] S. Bringay, N. Béchet, F. Bouillot, P. Poncelet, M. Roche, and M. Teisseire. Towards an on-line analysis of tweets processing. In *Proceedings of Database and Expert Systems Applications (DEXA), LNCS, Springer-Verlag*, volume 2, pages 154–161, 2011.

[8] F. Duchateau, Z. Bellahsene, and M. Roche. A context-based measure for discovering approximate semantic matching between schema elements. In *Proceedings of IEEE Research Challenges in Information Science (RCIS)*, pages 9–20, 2007.

[9] F. Duchateau, Z. Bellahsene, and M. Roche. Improving quality and performance of schema matching in large scale. **Ingénierie des Systèmes d'Information (ISI)**, 13(5) :59–82, 2008.

[10] A. Harb, M. Plantié, M. Roche, G. Dray, F. Trousset, and P. Poncelet. Détection d'opinion. comment déterminer les adjectifs d'opinion d'un domaine donné. **Documents Numériques**, 11(1-2) :37–61, 2008.

[11] R. Kessler, N. Béchet, J.-M. Torres Moreno, M. Roche, and M. El-Bèze. Job offer management : How improve the ranking of candidates. In *Proceedings of Foundations of Intelligent Systems (ISMIS), LNCS, Springer-Verlag*, pages 431–441, 2009.

[12] S. Laroum, N. Béchet, H. Hamza, and M. Roche. Classification automatique de documents bruités à faible contenu textuel. **Numéro spécial de la revue RNTI, Fouille de Données Complexes**, E-18, 2010.

[13] D. Li, A. Laurent, P. Poncelet, and M. Roche. Extraction of unexpected sentences : A sentiment classification assessed approach. **Intelligent Data Analysis**, 14(1) :31–46, 2010.

[14] C. Lopez, V. Prince, and M. Roche. Automatic titling of electronic documents by noun phrase extraction. In *Proceedings of Soft Computing and Pattern Recognition (SOCPAR)*, 2010.

[15] C. Lopez, V. Prince, and M. Roche. Automatic generation approach of short titles. In *Proceedings of Language and Technology Conference (LTC)*, 2011.

[16] C. Lopez, V. Prince, and M. Roche. Recherche documentaire par titrage automatique. In *Actes d'INFORSID*, pages 217–232, 2011.

[17] C. Lopez and M. Roche. Approche de construction automatique de titres courts par des méthodes de fouille du web. In *Actes de Traitement Automatique des Langues Naturelles (TALN)*, pages 39–50, 2011.

[18] A. Mela, M. Roche, and M.A. Bekhtaoui. Mixer les moyens pour extraire les gloses. In *Actes de la conférence Extraction et Gestion des Connaissances (EGC)*, pages 95–106, 2011.

[19] M. Plantié, M. Roche, G. Dray, and P. Poncelet. Is a voting approach accurate for opinion mining ? In *Proceedings of Data Warehousing and Knowledge Discovery (DaWaK), LNCS, Springer-Verlag*, pages 413–422, 2008.

[20] V. Prince and M. Roche, editors. *Information Retrieval in Biomedicine : Natural Language Processing for Knowledge Integration.* **Medical Information Science Reference, IGI Gobal**, 460 pages, 2009.

[21] M. Roche. *Intégration de la construction de la terminologie de domaines spécialisés dans un processus global de fouille de textes.* PhD thesis, Université Paris 11, Décembre 2004.

[22] M. Roche. How statistical information from the web can help identify named entities. In *Proceedings of International Conference on Web Information Systems (WEBIST), Session Web and Text Mining*, 2011.

[23] M. Roche and Y. Kodratoff. Pruning Terminology Extracted from a Specialized Corpus for CV Ontology Acquisition. In *Proceedings of onToContent Workshop - OTM'06, LNCS, Springer-Verlag*, pages 1107–1116, 2006.

[24] M. Roche and Y. Kodratoff. Text and web mining approaches in order to build specialized ontologies. **Journal of Digital Information**, 10(4), 2009.

[25] M. Roche and P. Poncelet, editors. *Fouille de Données d'Opinions.* **Revue des Nouvelles Technologies de l'Information (RNTI)**, Volume E-17, 202 pages, 2009.

[26] M. Roche and V. Prince. *AcroDef* : A quality measure for discriminating expansions of ambiguous acronyms. In *Proceedings of CONTEXT, LNCS, Springer-Verlag*, pages 411–424, 2007.

[27] M. Roche and V. Prince. Managing the Acronym/Expansion Identification Process for Text-Mining Applications. **International Journal of Software and Informatics**, 2(2) :163–179, 2008.

[28] M. Roche and V. Prince. A web-mining approach to disambiguate biomedical acronym expansions. **Informatica**, 34(2) :243–253, 2010.

[29] B. Rosoor, L. Sebag, S. Bringay, and M. Roche. Quand un tweet détecte une catastrophe naturelle... In *Actes de la conférence Veille Stratégique Scientifique et Technologique (VSST)*, pages 283–286, 2010.

[30] A. Sallaberry, N. Pecheur, S. Bringay, M. Roche, and M. Teisseire. Sequential patterns mining and gene sequence visualization to discover novelty from microarray data. **Journal of Biomedical Informatics**, *Elsevier*, 44(5) :760 – 774, 2011.

[31] H. Saneifar, S. Bonniol, A. Laurent, P. Poncelet, and M. Roche. Recherche de passages pertinents dans les fichiers logs par enrichissement de requêtes. In *Actes des Journées Francophones sur les Ontologies (JFO)*, 2009.

[32] H. Saneifar, S. Bonniol, A. Laurent, P. Poncelet, and M. Roche. Terminology extraction from log files. In *Proceedings of Database and Expert Systems Applications (DEXA), LNCS, Springer-Verlag*, pages 769–776, 2009.

[33] H. Saneifar, S. Bonniol, A. Laurent, P. Poncelet, and M. Roche. Passage retrieval in log files : An approach based on query enrichment. In *Proceedings of Advances in Natural Language Processing (IceTAL), LNCS, Springer-Verlag*, pages 357–368, 2010.

[34] H. Saneifar, S. Bonniol, A. Laurent, P. Poncelet, and M. Roche. Recherche de passages pertinents dans les fichiers logs par enrichissement de requêtes. In *Actes de la COnférence en Recherche d'Infomations et Applications (CORIA)*, pages 239–254, 2010.

[35] H. Saneifar, S. Bonniol, P. Poncelet, and M. Roche. Identification des divisions logiques de fichiers logs. In *Actes des Rencontres de la Société Francophone de Classification (SFC)*, 2011.

[36] C. Serp, E. Cazal, A. Laurent, and M. Roche. TERVOTIQ : un système de vote pour l'extraction de la terminologie d'un corpus en français médiéval. In *Proceedings of Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, volume 2, pages 1069–1080, 2008.

[37] C. Serp, A. Laurent, M. Roche, and M. Teisseire. La quête du graal et la réalité numérique. **Corpus**, 7 :173–189, 2008.

# Bibliographie


[Agrawal and Srikant, 1994] Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. *VLDB'94*.

[Andreevskaia and Bergler, 2006] Andreevskaia, A. and Bergler, S. (2006). Semantic tag extraction from wordnet glosses. In *Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation*.

[Audibert, 2003] Audibert, L. (2003). étude des critères de désambiguïsation sémantique automatique : résultats sur les cooccurrences. In *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN)*, pages 33–44.

[Aussenac-Gilles and Bourigault, 2003] Aussenac-Gilles, N. and Bourigault, D. (2003). Construction d'ontologies à partir de textes. In *Actes de Traitement Automatique des Langues Naturelles (TALN)*, volume 2, pages 27–47.

[Aussenac-Gilles and Jacques, 2008] Aussenac-Gilles, N. and Jacques, M. (2008). Designing and Evaluating Patterns for Relation Acquisition from Texts with CAMELEON. *Terminology, Pattern-Based approaches to Semantic Relations*, 14(1) :45–73.

[Azé, 2003] Azé, J. (2003). *Extraction de Connaissances dans des Données Numériques et Textuelles*. Thèse de Doctorat, Univ. de Paris 11.

[Baxendale, 1958] Baxendale, B. (1958). Man-made index for technical literature - an experiment. *IBM Journal of Research and Development.*, pages 354–361.

[Beaufort et al., 2010] Beaufort, R., Roekhaut, S., Cougnon, L., and Fairon, C. (2010). A hybrid rule/model-based finite-state framework for normalizing sms messages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 770–779. Association for Computational Linguistics.

[Bestgen, 2004] Bestgen, Y. (2004). Analyse sémantique latente et segmentation automatique de textes. In *Proceedings of the International Conference on Statistical Analysis of Textual Data (JADT)*, volume 1, pages 171–181.

[Bourigault, 1993] Bourigault, D. (1993). Analyse syntaxique locale pour le repérage de termes complexes dans un texte. *T.A.L.*, 34(2) :105–118.

[Bourigault, 2002] Bourigault, D. (2002). UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de Traitement Automatique des Langues Naturelles (TALN)*, pages 75–84.

[Bourigault et al., 2004] Bourigault, D., Aussenac-Gilles, N., and Charlet, J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle*, 18(1) :87–110.

[Bourigault and Fabre, 2000] Bourigault, D. and Fabre, C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaire*, 25 :131–151.

[Bourigault and Jacquemin, 1999] Bourigault, D. and Jacquemin, C. (1999). Term extraction + term clustering : An integrated platform for computer-aided terminology. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, pages 15–22.

[Brill, 1994] Brill, E. (1994). Some advances in transformation-based part of speech tagging. In *Conference on Artificial Intelligence (AAAI), Vol. 1*, pages 722–727.

[Brody and Diakopoulos, 2011] Brody, S. and Diakopoulos, N. (2011). Using Word Lengthening to Detect Sentiment in Microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 562–570.

[Brun and Hagège, 2004] Brun, C. and Hagège, C. (2004). Intertwining deep syntactic processing and named entity detection. In *Proceedings of Advances in Natural Language Processing, 4th International Conference (EsTAL)*, pages 195–206.

[Cacheda et al., 2010] Cacheda, F., V. Carneiro, D. F., and Formoso, V. (2010). Performance evaluation of large-scale information retrieval systems scaling down. In *Proceedings of the International Workshop on Large-Scale and Distributed Systems for Information Retrieval – SIGIR*.

[Cambria et al., 2010] Cambria, E., Speer, R., Havasi, C., and Hussain, A. (2010). Senticnet : A publicly available semantic resource for opinion mining. In *AAAI Fall Symposium Series*.

[Chang et al., 2002] Chang, J., Schtze, H., and Altman, R. (2002). Creating an online dictionary of abbreviations from medline. *Journal of the American Medical Informatics Association*, 9 :612–620.

[Chauché, 1984] Chauché, J. (1984). Un outil multidimensionnel de l'analyse du discours. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 11–15.

[Church and Hanks, 1990] Church, K. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. In *Computational Linguistics*, volume 16, pages 22–29.

[Cilibrasi and Vitanyi, 2007] Cilibrasi, R. and Vitanyi, P. M. B. (2007). The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3) :370–383.

[Clas, 1994] Clas, A. (1994). Collocations et langues de spécialité. *Meta*, 39(4) :576–580.

[Claveau and Sébillot, 2003] Claveau, V. and Sébillot, P. (2003). Apprentissage symbolique pour l'acquisition de ressources linguistiques. In *Actes de l'atelier "Acquisition, apprentissage et exploitation de connaissances sémantiques pour l'accès au contenu textuel" de la plateforme AFIA*.

[Clech and Zighed, 2003] Clech, J. and Zighed, D. (2003). Data mining et analyse des cv : Une expérience et des perspectives. In *Proceedings of Extraction et Gestion de Connaissances (EGC)*, pages 189–200.

[Codd et al., 1993] Codd, E., Codd, S., and Salley, C. (1993). Providing OLAP (on-line analytical processing) to user-analysts : An IT mandate. In *White Paper*.

[Daille, 1994] Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques.* PhD thesis, Université Paris 7.

[Daille et al., 2000] Daille, B., Fourour, N., and Morin, E. (2000). Catégorisation des noms propres : une étude en corpus. *Cahiers de Grammaire*, 25 :115–129.

[David and Plante, 1990] David, S. and Plante, P. (1990). De la nécéssité d'une approche morpho syntaxique dans l'analyse de textes. In *Intelligence Artificielle et Sciences Cognitives au Quebec*, volume 3, pages 140–154.

[Davidiv et al., 2010] Davidiv, D., Tsur, O., and Rappoport, A. (2010). Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of International Conference on Computational Linguistics (COLING).*

[Doan et al., 2002] Doan, A., Madhavan, J., Domingos, P., and Halvey, A. (2002). Learning to map ontologies on the semantic web. In *Proceedings of WWW Conference.*

[Downey et al., 2007] Downey, D., Broadhead, M., and Etzioni, O. (2007). Locating complex named entities in web text. In *Proceedings of theInternational Joint Conference on Artificial Intelligence (IJCAI)*, pages 2733–2739.

[Duan et al., 2010] Duan, Y., Jiang, L., Qin, T., Zhou, M., and Shum, H. (2010). An empirical study on learning to rank of tweets. In *Proceedings of International Conference on Computational Linguistics (COLING).*

[Dulac-Arnold et al., 2011] Dulac-Arnold, G., Denoyer, L., and Gallinari, P. (2011). Text classification : A sequential reading approach. In *Proceedings of the European Conference on Information Retrieval (ECIR)*, pages 411–423.

[Eisenstein et al., 2010] Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1277–1287.

[Esuli and Sebastiani, 2006] Esuli, A. and Sebastiani, F. (2006). SentiWordNet : A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resourcesand Evaluation (LREC)*, pages 417–422.

[Eugenio and Glass, 2004] Eugenio, B. D. and Glass, M. (2004). The kappa statistic : A second look. *Computational Linguistics*, 30(1) :95–101.

[Euzenat et al., 2004] Euzenat, J. et al. (2004). State of the art on ontology matching. Technical Report KWEB/2004/D2.2.3/v1.2, Knowledge Web.

[Euzenat and Shvaiko, 2007] Euzenat, J. and Shvaiko, P. (2007). *Ontology Matching.* Springer-Verlag.

[Even and Enguehard, 2002] Even, F. and Enguehard, C. (2002). Extraction d'informations à partir de corpus dégradés. In *Proceedings of 9ème conference sur le Traitement Automatique des Langues Naturelles (TALN'02)*, pages 105–115.

[Facca and Lanzi, 2005] Facca, F. M. and Lanzi, P. L. (2005). Mining interesting knowledge from weblogs : a survey. *Data Knowl. Eng.*, 53(3) :225–241.

[Fairon and Paumier, 2006] Fairon, C. and Paumier, S. (2006). A translated corpus of 30,000 french SMS. In *Proceedings of Language Resources and Evaluation Conference (LREC).*

[Faure, 2000] Faure, D. (2000). *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.

[Faure and Nédellec, 1999] Faure, D. and Nédellec, C. (1999). Knowledge acquisition of predicate argument structures from technical texts using machine learning : The system ASIUM. In *In Proc of the European Workshop, Knowledge Acquisition, Modelling and Management, LNAI*, pages 329–334.

[Ferri et al., 2002] Ferri, C., Flach, P., and Hernandez-Orallo, J. (2002). Learning decision trees using the area under the ROC curve. In *Proceedings of 9th International Conference on Machine Learning (ICML)*, pages 139–146.

[Fort et al., 2009] Fort, K., Ehrmann, M., and Nazarenko, A. (2009). Vers une méthodologie d'annotation des entités nommées en corpus. In *Actes de Traitement Automatique du Langage Naturel (TALN).*

[Frantzi et al., 2000] Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms : the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2) :115–130.

[Gimpel et al., 2011] Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for twitter : Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 42–47.

[Ginsberg et al., 2009] Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, pages 1012–1014.

[Hamon and Nazarenko, 1998] Hamon, T. and Nazarenko, A. (1998). Using general semantic information to help the terminology structuration. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, pages 675–680.

[Heid, 1998] Heid, U. (1998). Towards a corpus-based dictionary of German noun-verb collocations. In *Proceedings of the Euralex International Congress*, pages 301–312.

[Heiden and Guillot, 2003] Heiden, S. and Guillot, C. (2003). Capitalisation des savoirs par le web : une application de la tei pour l'encodage et l'exploitation des textes de la base de français médiéval. *Ancien et moyen français sur le Web, enjeux méthodologiques et analyse du discours).*

[Ho-Dac et al., 2004] Ho-Dac, L.-M., Jacques, M.-P., and Rebeyrolle, J. (2004). Sur la fonction discursive des titres. *S. Porhiel and D. Klingler (Eds). L'unité texte, Pleyben, Perspectives.*, pages 125–152.

[Hu and Liu, 2004] Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of KDD'04, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA.

[Illouz, 1999] Illouz, G. (1999). Méta étiqueteur adaptatif : Vers une utilisation pragmatique des ressources linguistiques. In *Actes de Traitement Automatique des Langues Naturelles (TALN).*

[Jackiewicz et al., 2010] Jackiewicz, A., Hunston, S., and El-Bèze, M., editors (2010). *Opinions, sentiments et jugements d'évaluation.* Traitement Automatique des Langues (TAL), Volume 51, Numéro 3.

[Jacquemin, 1997] Jacquemin, C. (1997). Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. In *Mémoire d'Habilitation à Diriger des Recherches en informatique fondamentale, Université de Nantes.*

[Jacques and Aussenac-Gilles, 2006] Jacques, M. and Aussenac-Gilles, N. (2006). Variabilité des performances des outils de tal et genre textuel. cas des patrons lexico-syntaxiques. *Traitement Automatique des Langues*, 47(1).

[Jacques and Rebeyrolle, 2004] Jacques, M. and Rebeyrolle, J. (2004). Titres et structuration des documents. *Actes International Symposium : Discourse and Document.*, pages 125–152.

[Jalam and Chauchat, 2002] Jalam, R. and Chauchat, J. (2002). Pourquoi les n-grammes permettent de classer des textes ? recherche de mots-clefs pertinents à l'aide des n-grammes caractéristiques. In *6th International Conference on Textual Data Statistical Analysis, France*, pages 381–390.

[Jarmasz and Szpakowicz, 2003] Jarmasz, M. and Szpakowicz, S. (2003). Roget's thesaurus and semantic similarity. In *Conference on Recent Advances in Natural Language Processing*, pages 212–219.

[Joachims, 1998] Joachims, T. (1998). Text categorization with support vector machines : learning with many relevant features. In *Proceedings of European Conference on Machine Learning (ECML)*, pages 137–142.

[Joshi et al., 2006] Joshi, M., Pakhomov, S., Pedersen, T., and Chute, C. G. (2006). A comparative study of supervised learning as applied to acronym expansion in clinical reports. In *Proceedings of the Annual Symposium of the American Medical Informatics Association*, pages 399–403.

[Junker and Hoch, 1997] Junker, M. and Hoch, R. (1997). Evaluating OCR and Non-OCR Text Representations for Learning Document Classifiers. In *Proceedings of the 4th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1060–1066. IEEE Computer Society.

[Kamps et al., 2004] Kamps, J., Marx, M., Mokken, R., and de Rijke, M. (2004). Using wordnet to measure semantic orientation of adjectives. *In Proceedings of LREC 2004, the 4th International Conference on Language Resources and Evaluation*, IV :174–181.

[Kastner and Monz, 2009] Kastner, I. and Monz, C. (2009). Automatic single-document key fact extraction from newswire articles. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 415–423. Association for Computational Linguistics.

[Kefi, 2006] Kefi, H. (2006). *Ontologies et aide à l'utilisateur pour l'interrogation de sources multiples et hétérogènes.* PhD thesis, Université de Paris 11.

[Keller and Lapata, 2003] Keller, F. and Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational linguistics*, 29(3) :459–484.

[Lallich and Teytaud, 2004] Lallich, S. and Teytaud, O. (2004). évaluation et validation des règles d'association. *Numéro spécial "Mesures de qualité pour la fouille des*

*données", Revue des Nouvelles Technologies de l'Information (RNTI)*, RNTI-E-1 :193–218.

[Landauer and Dumais, 1997] Landauer, T. and Dumais, S. (1997). A solution to plato's problem : The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2) :211–240.

[Larkey et al., 2000] Larkey, L., Ogilvie, P., Price, M., and Tamilio, B. (2000). Acrophile : An automated acronym extractor and server. In *Proceedings of the Fifth ACM International Conference on Digital Libraries*, pages 205–214.

[Larousse, 1992] Larousse, T. (1992). *Thésaurus Larousse - des idées aux mots, des mots aux idées.* Ed.Larousse, Paris.

[Laurens, 1999] Laurens, M. (1999). La description des collocations et leur traitement dans les dictionnaires. *Romaneske*, 4.

[Levenshtein, 1966] Levenshtein, V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10 :707.

[Lin et al., 2008] Lin, C. X., Ding, B., Han, J., Zhu, F., and Zhao, B. (2008). Text Cube : Computing IR Measures for Multidimensional Text Database Analysis. In *Proceedings of Int. Conf. on Data Mining (ICDM'08)*, pages 905–910.

[Lin, 1998] Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of 15th International Conf. on Machine Learning*, pages 296–304. Morgan Kaufmann.

[Lv and Zhai, 2009] Lv, Y. and Zhai, C. (2009). Adaptive relevance feedback in information retrieval. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*, pages 255–264.

[Maedche and Staab, 2001] Maedche, A. and Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16 :72–79.

[Maedche and Staab, 2002] Maedche, A. and Staab, S. (2002). Measuring similarity between ontologies. In *Proceedings of Knowledge Engineering and Knowledge Management (EKAW)*, pages 251–263.

[Màrquez et al., 1999] Màrquez, L., Padro, L., and Rodriguez, H. (1999). Improving tagging accuracy by using voting taggers. In *Proceedings of NLP+IA/TAL+AI ?98.*

[Mela, 2004] Mela, A. (2004). Linguistes et "talistes" peuvent coopérer : repérage et analyse des gloses. *Revue Française de Linguistique Appliquée, "Linguistique et informatique : nouveaux défis", B. Habert (resp.)*, 9(1).

[Mela, 2005] Mela, A. (2005). Le repérage automatique des gloses de nomination seconde. *Langues et langage, "Les marqueurs de la glose", A. Steuckardt (resp.), Publications de l'Université de Provence.*

[Mel'čuk et al., 1999] Mel'čuk, I. A., Arbatchewsky-Jumarie, N., Elnitsky, L., and Lessard, A. (1984, 1988, 1992, 1999). *Dictionnaire explicatif et combinatoire du francais contemporain.* Presses de l'Université de Montréal, Montréal, Canada. Volume 1, 2, 3, 4.

[Miller, 1995] Miller, G. (1995). Wordnet : A lexical database for english. In *Communications of the ACM.*

[Monceaux, 2002] Monceaux, L. (2002). *Adaptation du niveau d ?analyse des interventions dans un dialogue – application à un système de question-réponse*. PhD thesis, Université Paris 11.

[Nadeau and Sekine, 2007] Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1) :3–26.

[Navarro, 1999] Navarro, G. (1999). A guided tour to approximate string matching. *ACM Computing Surveys*, 33 :2001.

[Navigli, 2009] Navigli, R. (2009). Word sense disambiguation : A survey. *ACM Comput. Surv.*, 41(2).

[Nenadic et al., 2003] Nenadic, G., Spasic, I., and Ananiadou, S. (2003). Terminology-Driven Mining of Biomedical Literature. *Bioinformatics*, 19(8) :938–943.

[Nouvel and Soulet, 2011] Nouvel, D. and Soulet, A. (2011). Annotation d'entités nommées par extraction de règles de transduction. In *Proceedings of Extraction et Gestion des Connaissances (EGC)*, pages 119–130.

[Nyberg et al., 2010] Nyberg, K., Raiko, T., Hyvönen, E., and Tiinanen, T. (2010). Document classification utilising ontologies and relations between documents. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs (MLG)*.

[Okazaki and Ananiadou, 2006] Okazaki, N. and Ananiadou, S. (2006). Building an abbreviation dictionary using a term recognition approach. *22*, Bioinformatics(24) :3089–3095.

[Orasan et al., 2008] Orasan, C., Cristea, D., Mitkov, R., and Branco, A. H. (2008). Anaphora resolution exercise : an overview. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

[Paik et al., 1994] Paik, W., Liddy, E., Yu, E., and McKenna, M. (1994). Categorizing and standardizing proper nouns for efficient information retrieval. In *Corpus Processing for Lexical Acquisition, MIT Press, chap. 4*.

[Pattabhi et al., 2010] Pattabhi, T., Rao, R., and Devi, S. L. (2010). How to get the same news from different language news papers. In *Proceedings of Fourth International Workshop on Cross Lingual Information Access – COLING Conference*, pages 11–15.

[Pérez-Martínez et al., 2008] Pérez-Martínez, J. M., Llavori, R. B., Cabo, M. J. A., and Pedersen, T. B. (2008). Contextualizing data warehouses with documents. *Decision Support Systems*, 45(1) :77–94.

[Petrovic et al., 2006] Petrovic, S., Snajder, J., Dalbelo-Basic, B., and Kolar, M. (2006). Comparison of collocation extraction measures for document indexing. In *Proceedings of Information Technology Interfaces (ITI)*, pages 451– 456.

[Peñalver Vicea, 2003] Peñalver Vicea, M. (2003). Le titre est-il un désignateur rigide ? *Dialnet, Vol. 2*, pages 251–258.

[Porter, 1980] Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3) :130–137.

[Prince and Labadié, 2007] Prince, V. and Labadié, A. (2007). Text segmentation based on document understanding for information retrieval. In *Natural Language Processing and Information Systems, LNCS, Springer*, pages 295–304.

[Pujolle et al., 2008] Pujolle, G., Ravat, F., Teste, O., and Tournier, R. (2008). Fonctions d'agrégation pour l'analyse en ligne (OLAP) de données textuelles. fonctions top_kwk et avg_kw opérant sur des termes. *Ingénierie des Systèmes d'Information*, 13(6) :61–84.

[Qamar and Gaussier, 2009] Qamar, A. and Gaussier, E. (2009). Online and batch learning of generalized cosine similarities. In *Proceedings of International Conference on Data Mining (ICDM)*, pages 926–931.

[Rahm and Bernstein, 2001] Rahm, E. and Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *VLDB Journal : Very Large Data Bases*, 10(4) :334–350.

[Rebeyrolles, 2000] Rebeyrolles, J. (2000). *Forme et fonction de la définition en discours.* PhD thesis, Université de Toulouse-le-Mirail, Toulouse II.

[Rehder et al., 1998] Rehder, B., Schreiner, M., Wolfe, M., Laham, D., Landauer, T., and Kintsch, W. (1998). Using latent semantic analysis to assess knowledge : Some technical considerations. In *Discourse Processes*, volume 25, pages 337–354.

[Robertson et al., 1994] Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M. (1994). Okapi at trec-3. In *Proceedings of TREC (Text REtrieval Conference).*

[Sakaki et al., 2010] Sakaki, T., Okazaki, M., and Matsuon, Y. (2010). Earthquake shakes twitter users : real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web (WWW)*, pages 851–860.

[Salton and Buckley, 1988] Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management 24*, page 513 à 523.

[Salton et al., 1975] Salton, G., Wong, A., and Yang, C. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11) :613–620.

[Schmid, 1994] Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49.

[Sclano and Velardi, 2007] Sclano, F. and Velardi, P. (2007). Termextractor : a web application to learn the shared terminology of emergent web communities. In *Proceedings of the 3rd International Conference on Interoperability for Enterprise Software and Applications (I-ESA 2007)*, Funchal, Portugal.

[Scott and Matwin, 1999] Scott, S. and Matwin, S. (1999). Feature engineering for text classification. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*, pages 379–388.

[Sebastiani, 2002] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1) :1–47.

[Sjobergh, 2003] Sjobergh, J. (2003). Combining pos-taggers for improved accuracy on swedish text. In *Proceedings of the Nordic Conference of Computational Linguistics (NoDaLiDa).*

[Smadja, 1993] Smadja, F. (1993). Retrieving collocations from text : Xtract. *Computational Linguistics*, 19(1) :143–177.

[Smadja et al., 1996] Smadja, F., McKeown, K. R., and Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons : A statistical approach. *Computational Linguistics*, 22(1) :1–38.

[Sokolova and Lapalme, 2008] Sokolova, M. and Lapalme, G. (2008). Verbs speak loud : Verb categories in learning polarity and strength of opinions. In *In Proceedings of Conference of the Canadian Society for Computational Studies of Intelligence*, pages 320–331.

[Sokolova and Lapalme, 2011] Sokolova, M. and Lapalme, G. (2011). Learning opinions in user-generated web content. *Natural Language Engineering*, 17(4) :541–567.

[Stein, 2003] Stein, A. (2003). Part of speech tagging and lemmatisation of old french texts. In *http ://www.unistuttgart. de/lingrom/stein/forschung/altfranz/afrlemma.pdf.*

[Stevenson et al., 2009] Stevenson, M., Guo, Y., Alamri, A., and Gaizauskas, R. (2009). Disambiguation of biomedical abbreviations. In *Proceedings of the BioNLP 2009 Workshop*, pages 71–79, Boulder, Colorado. Association for Computational Linguistics.

[Taboada et al., 2006] Taboada, M., Anthony, C., and Voll, K. (2006). Creating semantic orientation dictionaries. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC).*

[Taboada et al., 2011] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2) :267–307.

[Thanopoulos et al., 2002] Thanopoulos, A., Fakotakis, N., and Kokkianakis, G. (2002). Comparative Evaluation of Collocation Extraction Metrics. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 620–625.

[Tranier et al., 2004] Tranier, J., Baraër, R., Bellahsene, Z., and Teisseire, M. (2004). Where's charlie : Family-based heuristics for peer-to-peer schema integration. In *Proceedings of IDEAS*, pages 227–235.

[Turney, 2001] Turney, P. (2001). Mining the Web for synonyms : PMI–IR versus LSA on TOEFL. *Proceedings of the 12th European Conference on Machine Learning (ECML), LNCS*, 2167 :491–502.

[Turney, 2002] Turney, P. (2002). Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews. *In Proceedings of 40th Meeting of the Association for Computational Linguistics*, pages 417–424.

[Turney and Pantel, 2010] Turney, P. D. and Pantel, P. (2010). From frequency to meaning : Vector space models of semantics. *Journal of Artificial Intelligence Research (JAIR)*, 37 :141–188.

[Vinet, 1993] Vinet, M.-T. (1993). L'aspet et la copule vide dans la grammaire des titres. *Persee*, 100 :83–101.

[Vivaldi et al., 2001] Vivaldi, J., Màrquez, L., and Rodríguez, H. (2001). Improving term extraction by system combination using boosting. In *Proceedings of the 12th European Conference on Machine Learning (ECML)*, pages 515–526.

[Voll and Taboada, 2007] Voll, K. and Taboada, M. (2007). Not all words are created equal : Extracting semantic orientation as a function of adjective relevance. In *Proceedings of the 20th Australian Joint Conference on Artificial Intelligence.*

[Voorhees, 1999] Voorhees, E. M. (1999). The trec-8 question answering track report. In *Proceedings of Text REtrieval Conference (TREC-8)*, pages 77–82.

[Weissenbacher and Nazarenko, 2007] Weissenbacher, D. and Nazarenko, A. (2007). Identifier les pronoms anaphoriques et trouver leurs antécédents : l'intérêt de la classification bayésienne. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*, pages 145–155, France. ATALA.

[Wiegand et al., 2010] Wiegand, M., Balahur, A., Roth, B., Klakow, D., and Montoyo, A. (2010). A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*.

[Wilks, 1998] Wilks, Y. (1998). Language processing and the thesaurus. In *National Language Research Institute*.

[Xu and Huang, 2007] Xu, J. and Huang, Y. (2007). Using svm to extract acronyms from text. *Soft Comput.*, 11(4) :369–373.

[Yamanishi and Maruyama, 2005] Yamanishi, K. and Maruyama, Y. (2005). Dynamic syslog mining for network failure monitoring. In *KDD '05 : Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 499–508, New York, NY, USA. ACM.

[Yeates, 1999] Yeates, S. (1999). Automatic extraction of acronyms from text. In *New Zealand Computer Science Research Students' Conference*, pages 117–124.

[Zajic et al., 2002] Zajic, D., Door, B., and Schwarz, R. (2002). Automatic headline generation for newspaper stories. *Workshop on Text Summarization (ACL 2002 and DUC 2002 meeting on Text Summarization). Philadelphia.*

[Zhang et al., 2009] Zhang, D., Zhai, C. X., and Han, J. (2009). Topic cube : Topic modeling for OLAP on multidimensional text databases. In *Proceedings of the SIAM Int. Conference on Data Mining*, pages 1123–1134.