Traitement Automatique des Langues et Recherche d'Information en langue arabe dans un domaine de spécialité : Apport des connaissances morphologiques et syntaxiques pour l'indexation

Siham Boulaknadel

École Centrale de Nantes **Université de Nantes** **École des Mines de Nantes**

ÉCOLE DOCTORALE STIM

« SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DES MATÉRIAUX »

Année 2008

N° attribué par la bibliothèque

Traitement Automatique des Langues et Recherche d'Information en langue arabe dans un domaine de spécialité :

Apport des connaissances morphologiques et syntaxiques pour l'indexation

THÈSE DE DOCTORAT

Discipline : INFORMATIQUE

Présentée
et soutenue publiquement par

Siham Boulaknadel

Le 18 Octobre 2008, devant le jury ci dessous

Président :	José Martinez	LINA, Univ.Nantes
Rapporteurs :	Josiane Mothe, Professeur	IRIT, Univ.Toulouse
	Abdelfatah Hamdani, Professeur	IERA
Examinateurs :	Béatrice Daille, Professeur	LINA, Univ.Nantes
	Driss Aboutajdine, Professeur	FSR, Univ.Mohammed V
	Elqadi Abderrahim, Professeur Assistant	EST, Meknès

Directeur de thèse : Pr. Béatrice Daille / Pr. Driss Aboutajdine

Laboratoire: LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE.
CNRS FRE 2729. 2, rue de la Houssinière, BP 92 208 . 44 322 Nantes, CEDEX 3. N° ED 503- 020

TRAITEMENT AUTOMATIQUE DES LANGUES ET RECHERCHE D'INFORMATION EN LANGUE ARABE : APPORT DES CONNAISSANCES MORPHOLOGIQUES ET SYNTAXIQUES POUR L'INDEXATION

Siham BOULAKNADEL

⋈

favet neptunus eunti

Université de Nantes

Siham BOULAKNADEL

Traitement automatique des langues et Recherche d'information en Langue arabe : apport des connaissances morphologiques et syntaxiques pour l'indexation

xxiii+

Ce document a été préparé avec LaTeX2$_\varepsilon$ et la classe `these-LINA` version 0.92 de l'association de jeunes chercheurs en informatique LOGIN, Université de Nantes. La classe `these-LINA` est disponible à l'adresse :

`http://www.sciences.univ-nantes.fr/info/Login/`

Impression : memoire.tex – 2/2/2009 – 12:25

Révision pour la classe : $Id: these-LINA.cls,v 1.3 2000/11/19 18:30:42 fred Exp

Résumé

La Recherche d'Information a pour objectif de fournir à un utilisateur un accès facile à l'information qui l'intéresse, cette information étant située dans une masse de documents textuels. Afin d'atteindre cet objectif, un système de recherche d'information doit représenter, stocker et organiser l'information, puis fournir à l'utilisateur les éléments correspondant au besoin d'information exprimé par sa requête. La plupart des systèmes de recherche d'information (SRI) utilisent des termes simples pour indexer et retrouver des documents. Cependant, cette représentation n'est pas assez précise pour représenter le contenu des documents et des requêtes, du fait de l'ambiguïté des termes isolés de leur contexte. Une solution à ce problème consiste à utiliser des termes complexes à la place de termes simples isolés. Cette approche se fonde sur l'hypothèse qu'un terme complexe est moins ambigu qu'un terme simple isolé. Notre thèse s'inscrit dans le cadre de la recherche d'information dans un domaine de spécialité en langue arabe. L'objectif de notre travail a été d'une part, d'identifier les termes complexes présents dans les requêtes et les documents. D'autre part, d'exploiter pleinement la richesse de la langue en combinant plusieurs connaissances linguistiques appartenant aux niveaux morphologique et syntaxique, et de montrer comment l'apport de connaissances morphologiques et syntaxiques permet d'améliorer l'accès à l'information. Ainsi, nous avons proposé une plate-forme intégrant divers composants dans le domaine public ; elle conduit à montrer l'apport significatif et tranché de plusieurs de ces composants. En outre, nous avons avons défini linguistiquement les termes complexes en langue arabe et nous avons développé un système d'identification de termes complexes sur corpus qui produit des résultats de bonne qualité en terme de précision, en s'appuyant sur une approche mixte qui combine modèle statistique et données linguistiques

Remerciements

Le travail de cette thèse a été réalisé au sein du Laboratoire de Recherche en Informatique et Télécommunications (LRIT) de la Faculté des Sciences de Rabat. Il a été effectué dans le cadre du programme de la co-tutelle, en collaboration avec le laboratoire d'Informatique de Nantes Atlantique (LINA) de l'université de Nantes, France.

Je tiens tout d'abord à exprimer ma profonde gratitude à Monsieur Driss Aboutajdine, professeur à la Faculté des Sciences de Rabat et responsable du LRIT, pour m'avoir encadré avec un intérêt constant et une grande compétence, pour sa disponibilité, son soutien, ses conseils, et les encouragements qui m'ont permis de mener à bien ce travail.

J'exprime ma profonde reconnaissance à Madame Béatrice Daille, professeur à l'université de Nantes, pour son aide précieuse, les efforts qu'elle a prodigués pour l'accomplissement de ce travail, ainsi pour la qualité de l'encadrement qu'elle m'a assuré.

Je tiens aussi à remercier Monsieur Abderrahim El Qadi, professeur assistant à l'école Supérieure de Technologie de Meknès, pour son co-encadrement, pour les discussions fructueuses que nous avons eues et pour l'intérêt qu'il a bien voulu porter à mon travail.

Que Monsieur José Martinez, professeur à l'université de Nantes, trouve ici l'expression de mes remerciements les plus sincères d'avoir accepté de présider cette thèse.

Je suis très honorée par la présence de Madame Josiane Mothe, professeur à l'université de Toulouse, France, et Monsieur Abdelfatah Hamdani, professeur à l'Institut des Etudes et de Recherches pour l'Arabisation. Qu'ils trouvent ici mes sincères remerciements d'avoir accepté d'être rapporteurs de ce travail.

J'exprime également mes remerciements à tous les membres du Comité Scientifique de la coopération franco-marocaine dans le domaine des STIC (programme géré par l'INRIA du côté français).

Je remercie Les responsables du Centre National pour la Recherche Scientifique et Technique (CNRST) pour m'avoir permis d'effectuer ce travail dans de bonnes conditions matérielles.

Je remercie toutes les personnes qui ont participé de manière directe ou indirecte à la concrétisation de ce travail et plus particulièrement mon amie Fadoua Ataa-Allah, qui m'a accompagné au cours de mes années de thèse. Qu'elle trouve ici une expression de ma reconnaissance.

Je voudrais aussi remercier tous mes collègues du laboratoire LRIT qui ont rempli ces années de complicité de moments agréables ainsi que l'équipe TALN dont la compagnie en contexte professionnel est réellement enrichissante.

Je remercie ma famille qui a su manifester son soutien et m'entourer d'affection pendant les moments difficiles.

Sommaire

Table des matières

CHAPITRE 1

Introduction

L'évolution très rapide d'Internet a conduit à révéler la RI au grand jour, notamment par le biais des moteurs de recherche. La profusion de données numériques disponibles a rendu indispensables des moyens de recherche performants et automatiques, permettant à tout un chacun de trouver une information précise. Un systèmes de recherche d'information (SRI) doit faire face à trois types de défis à savoir, la gestion d'un volume important d'informations, la présence de multiples supports et, finalement, le caractère plurilingue de la Toile qui représente un enjeu considérable. Dans ce contexte, l'importance grandissante d'autres langues que l'anglais a suscité le développement d'outils et de techniques automatiques afin de permettre leur traitement informatique. Ce besoin n'est pas marginal. En septembre 2007[1], la proportion d'internautes naviguant en langue arabe était estimée à 17,4 %. Sur cette base, nous estimons que l'utilisation de la langue arabe sur le Web va atteindre des valeurs comparables à celle des langues européennes.

En comparaison de l'anglais ou d'autres langues indo-européennes, la langue arabe présente des caractéristiques singulières. Ainsi, son traitement automatique doit faire face à :

- la nature agglutinante de la langue : l'ensemble des morphèmes collés à l'unité lexicale[2] véhiculent plusieurs informations morphosyntaxiques.
- la richesse flexionnelle de l'arabe
- l'absence de voyellation de la majorité des textes arabes écrits : ce phénomène entraîne un nombre important d'ambiguités morphologiques. En arabe, chaque lettre doit prendre un signe de voyellation et de surcroît les voyelles finales sont porteuses de certains traits morpho-syntaxiques comme la déclinaison, le mode, le cas.

Face à ces défis et sous l'impulsion des campagnes d'évaluation TREC-2001 [55], diverses approches [4] se tournent vers des représentations plus riche des documents manipulés dont l'objectif est d'améliorer les perfomances d'un SRI en langue arabe .

Notre étude s'inscrit dans le cadre de l'indexation pour la recherche d'information dans un domaine de spécialité en langue arabe. Nous nous sommes intéressées à un domaine particulier car l'extraction des termes complexes est plus significative qu'en domaine général. L'un de nos objectifs est l'amélioration des performances des SRI dans un domaine de spécialité en langue arabe par la prise en compte des unités lexicales complexes qui sont plus précises que les unités lexicales réduites à un seul terme. De plus, traiter le problème de leur variation permettra d'introduire de la flexibilité dans la procédure d'appariement. Différentes réalisations linguistiques seront regroupées et considérées équivalentes si elles portent le même contenu informationnel. Dans leur grande majorité, les systèmes actuels de RI en langue arabe se contentent d'exploiter les unités simples et n'effectuent peu de traitements de nature linguistique [85] [4]. Ces traitements se limite à la troncature des unités lexicales extraites du corpus et à l'utilisation d'anti-dictionnaire de la langue arabe.

[1] http://www.internetworldstats.com/

[2] Selon Mel'cuk [93], une unité lexicale est une entité trilatérale composée de (i): un sens, (ii): une forme phonique/graphique et (iii): un ensemble de traits de combinatoire (le syntactique)

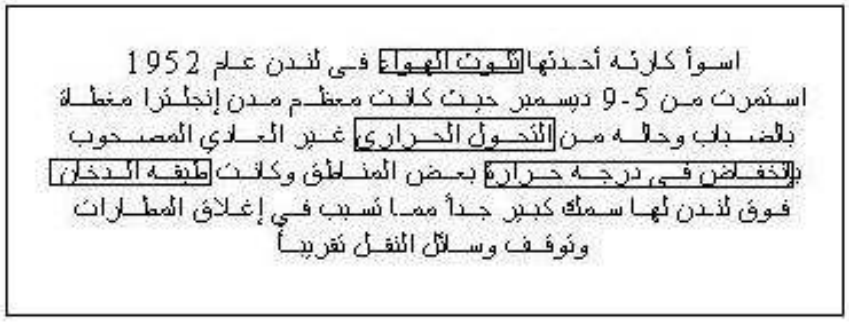

Figure 1.1 – Unités lexicales représentatives du document

Notre approche d'extraction de connaissances textuelles prend en considération les combinaisons des unités lexicales au niveau de l'analyse de texte. Contrairement à la plupart des travaux qui analyse le texte sous la forme de chaînes de caractères atomiques, notre but est de traiter le texte en conservant les rapports syntagmatiques qu'entretiennent ces unités lexicales. Ainsi, nous considérons le document illustré dans la Figure 1.1, nous avons encadré, à titre d'exemple, des unités lexicales (pollution de l'air, transfert thermique) capables de représenter le contenu du document.

Dans nos travaux, nous définissons ces unités et nous montrons que se sont en fait des termes complexes qui peuvent être repérés dans le texte en utilisant les rapports syntagmatiques entre les termes. Notre méthodologie d'extraction de termes complexes se base donc sur deux points de vue : un point de vue statistique et un point de vue linguistique.

Le point de vue linguistique concerne les combinaisons des éléments textuels au niveau du texte. C'est un niveau qui est très proche de la syntaxe et qui prend en considération les rapports syntagmatiques entre les différentes unités textuelles. Il permet de mettre en évidence les combinaisons linguistiquement correctes donc qui sont susceptibles d'être sémantiquement plus riches. Notre but n'est pas une analyse en vue de la compréhension de la langue naturelle mais une reconnaissance de la structure linguistique des unités textuelles capables de véhiculer le sens contenu dans un texte.

Le point de vue statistique concerne l'évaluation des mesures statistiques comme un critère pour mesurer l'importance des termes extraîts ainsi que leur pouvoir de représenter le contenu textuel des documents. Ce critère mesure le pouvoir évocateur d'un terme et permet de comparer les termes entre eux.

1.1 Organisation de la thèse

Dans le deuxième chapitre, nous présentons les mécanismes traditionnels de la RI. Nous y dressons une synthèse des principales techniques exploitées par les SRI pour représenter les documents et requêtes, mettre en correspondance leurs contenus et retourner à l'utilisateur les documents dont le contenu est le plus proche de celui de sa requête. Le chapitre trois propose une synthèse des contributions possibles des techniques issues du TAL pour une application en RI à travers un tour d'horizon des diverses tentatives déjà réalisées dans ce vaste domaine. Nous dressons un bilan de l'apport du TAL à la RI et plus particulièrement à évaluer l'intérêt en RI de combiner des informations linguistiques multi-niveaux (d'ordre morphologique, syntaxique et sémantique) plus à même d'exploiter la richesse de la langue. Le chapitre quatre décrit les principales caractéristiques de la langue arabe et recense un nombre de ressources linguistiques comprenant des lexiques monolingues et multilingues ainsi que des corpus de langue générale et des corpus de spécialité ainsi que des outils linguistiques à savoir les analyseurs morphologiques et les racineurs. Le chapitre cinq présente une étude linguistique menée sur corpus afin d'établir les spécifications linguistiques nécessaires à l'acquisition des termes complexes. Nous caractérisons ensuite les différentes variations que peuvent subir ces termes. Pour la découverte des termes complexes, nous utilisons plusieurs types de méthodes : l'analyse partielle qui permet une formalisation linguistique des spécifications linguistiques que nous privilégions par rapport à une analyse par frontières. Nous décrivons les régles morphologiques et nous abordons la technique choisie pour filtrer les termes complexes, les modèles statistiques, l'évaluation de ces modèles, et la justification du choix final de n'en retenir qu'un seul. le chapitre six propose une plate forme pour évaluer l'intérêt de combiner en RI des informations linguistiques appartenant à la fois au niveau morphologique, syntaxique de la langue, de sélectionner les schémas de pondération qui améliorent la performance de la méthode LSA pour la recherche d'information dans un corpus spécialisé en langue arabe et de comparer la performance du modèle vectoriel avec celle du modèle LSA. Le système que nous proposons pour l'évaluation de l'intégration de connaissances s'inspire du travail de [123], puisque nous ré-utilisons l'idée de cette architecture qui nous paraît pertinente pour représenter plusieurs informations linguistiques sous la forme de descripteurs. Nous terminons en mettant en exergue les principales contributions de nos travaux et en proposant quelques pistes de recherche ouvertes à l'issue de ces derniers.

CHAPITRE 2
Recherche d'information

Ce chapitre présente un tour d'horizon du domaine de Recherche d'information (RI). Il decrit tout d'abord le processus général de RI, pour retrouver parmi un ensemble de documents ceux qui répondent précisemment à la requête d'un utilisateur. Il introduit ensuite les diverses méthodes de representation textuelle des documents et requêtes communément utilisées. Enfin, il présente un aperçu sur les techniques d'évaluation utilisées pour juger de la pertinence des systèmes de recherche d'information.

2.1 Introduction

Le but de la recherche d'information (RI) est de développer des systèmes capables de retrouver parmi un ensemble de documents ceux qui répondent au mieux à la requête d'un utilisateur. Pour cela, il est important de constituer une représentation du contenu du document et de la requête afin de procéder à un appariement plus pertinent entre eux. L'approche souvent adoptée en RI textuelle est plutôt de chercher des représentants qui correspondent généralement, dans le cadre de l'indexation automatique, à un ensemble d'unités lexicales [1] extraits des documents et requêtes, nommés termes d'indexation. L'indexation consiste donc à associer à chaque document (ou à chaque requête) un descripteur (également nommé index) formé de l'ensemble des termes d'indexation extraits de son contenu.
Pour établir une correspondance entre documents et requêtes, représentés par des descripteurs, les SRI se basent sur des modèles de RI. Ils permettent :

- d'offrir une interprétation aux descripteurs en donnant une représentation interne des textes et des questions basée sur les termes d'indexation ;
- de définir les stratégies à adopter pour comparer les représentations des documents et des requêtes. Leur comparaison donne lieu à un score qui traduit leur degré de ressemblance ;
- de proposer éventuellement des méthodes de classement des résultats retournés à l'utilisateur.

Une fois les représentations des documents et des requêtes mises en correspondance, le système retourne à l'utilisateur la liste des documents répondant à sa requête. Ainsi, des méthodes et des mesures d'évaluation sont nécessaires pour estimer la validité des résultats retournés par le système. Une partie de ce chapitre y est consacrée.

2.2 Processus de recherche d'information

Le processus de RI a pour but d'établir une correspondance pertinente entre l'information recherchée par l'utilisateur, représentée généralement par le biais d'une requête, et l'ensemble des documents

[1] C'est l'unité de base de la lexicologie, c'est-à-dire l'unité élémentaire du lexique d'une langue L, qui doit être décrite dans un dictionnaire par une entrée séparée

disponibles. Il s'articule autour de deux étapes essentielles : les phases d'indexation et de recherche. Le processus complet est représenté en figure C.1.
L'étape d'indexation se base sur l'analyse des documents et des requêtes afin de créer une représentation de leur contenu textuel qui soit utilisable par le SRI. Chaque document (et requête) est alors associé à un descripteur représenté par l'ensemble des termes d'indexation extraits.
La phase de recherche a pour objectif d'apparier les documents et la requête de l'utilisateur en comparant leurs descripteurs respectifs. Elle se base sur un formalisme précis défini par un modèle de RI. Les documents présentés en résultat à l'utilisateur, et considérés comme les plus pertinents, sont ceux dont les termes d'indexation sont les plus proches de ceux de la requête.

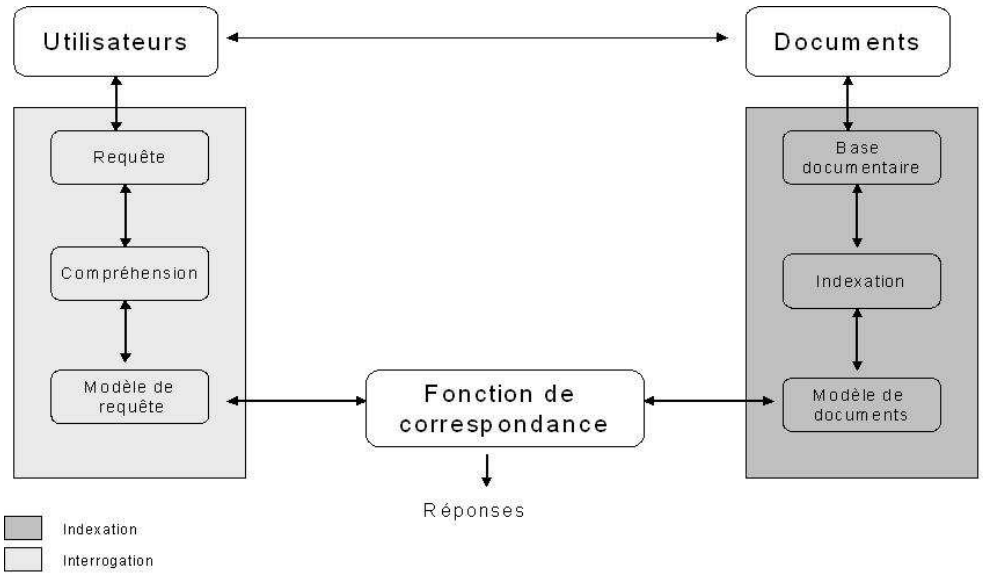

Figure 2.1 – Système de recherche d'information

2.3 Modèles de RI

Comme nous l'avons vu, le but d'un SRI demeure dans sa capacité à établir une correspondance entre un document et une requête. De nombreux modèles ont été proposés en RI, ils sont généralement regroupés autour des trois familles suivantes [105]:

- les modèles ensemblistes qui considèrent le processus de recherche comme une succession d'opérations à effectuer sur des ensembles d'unités lexicales contenues dans les documents,
- les modèles algébriques au sein desquels la pertinence d'un document par rapport à une requête est envisagée à partir de mesures de distance dans un espace vectoriel,

- les modèles probabilistes qui représentent la RI comme un processus incertain et imprécis où la notion de pertinence peut être vue comme une probabilité de pertinence.

2.3.1 Modèles ensemblistes

Nous nous intéressons ici uniquement au principal représentant des modèles inspirés de la logique booléenne et de la théorie des ensembles pour modéliser l'appariement entre une requête et les documents de la collection : le modèle booléen classique.
Le modèle booléen est le modèle le plus ancien et également le plus simple en RI. Un document est représenté par l'ensemble d'unités lexicales qu'il contient. Une requête est représentée comme une formule logique portant sur la présence ou l'absence d'unités lexicales reliées par des connecteurs (le ou $\vee$, le et $\wedge$, le non $\neg$).
Le modèle booléen fait correspondre à chaque connecteur une opération ensembliste portant les documents de la base. Si l'on note D la base documentaire, et D_q l'ensemble des documents de la base correspondant à la requête q, on définit récursivement :

requête	ensemble réponse
q = t avec t un terme	D_q=D_t l'ensemble des documents contenant t
q = $q_1 \wedge q_2$	$D_q = D_{q_1} \cap D_{q_2}$
q = $q_1 \vee q_2$	$D_q = D_{q_1} \cup D_{q_2}$
q = $q_1 \neg q_2$	D $\setminus D_{q_1}$

Ainsi, une requête : orange $\wedge$ (ville $\vee$ cité) $\wedge$ ($\neg$ (réseau $\vee$ opérateur $\vee$ mobile)) retourne à l'utilisateur les documents contenant obligatoirement l'unité lexicale "orange" et l'un des deux unités lexicales "ville" ou "cité" mais qui ne contiennent en aucun cas les unités lexicales "réseau", "opérateur" et "mobile". Les limites de ce modèle résultent directement de la représentation choisie. Ainsi, étant donné à un document, la requête est soit vraie soit fausse. En termes ensemblistes, cela se traduit par la discrimination d'un document à partir de l'absence ou la présence d'une seule unité lexicale dans ce dernier.

2.3.2 Modèles algébriques

Couramment employé en RI, les modèles algébriques considèrent les documents et les requêtes comme faisant partie d'un même espace vectoriel, et leur appariement est fait suivant une mesure algébrique de similarité. Parmi les différentes variantes de ce type de modèle, le plus connu est le modèle vectoriel.

Modèle vectoriel

Ce modèle [115] répond à beaucoup de problèmes posés par le modèle booléen. Il représente un document par un vecteur indiquant l'utilisation ou non d'une unité lexicale servant à l'indexation. Le calcul de similarité correspond au nombre d'unités lexicales en commun entre la requête et chaque document de la base et est utilisé pour trier les documents. Ainsi, le premier document présenté est celui qui est potentiellement le plus adéquat à répondre à la requête posée. Un document sera donc représenté par un vecteur tel que :

$$D_i \ = \ (t_{i1} \ t_{i2} \ ... \ t_{im}) \tag{2.1}$$

Avec m unités lexicales dans l'ensemble d'indexation et t_{ik} poids du terme k.

Le processus de recherche associé à ce modèle consiste à calculer la similarité entre une requête et les documents réponses. La requête représentée dans le modèle aura une représentation identique à celle utilisée pour un document :

$$R = (t_{i1}\ t_{i2}\ ...\ t_{im}) \text{ avec } t_{ik} \neq 0 \text{ si } t_{ik} \text{ est présent dans } R \tag{2.2}$$

Finalement, tous les vecteurs des documents de la collection peuvent être rassemblés dans une matrice dont les lignes et les colonnes représentent respectivement les documents et les termes d'indexation. Cette matrice est appelée matrice d'occurrences. Les matrices d'occurrences sont le plus souvent très creuses puisque les termes d'indexation formant les bases de l'espace vectoriel apparaîssent rarement en totalité dans l'ensemble des documents. En plus la taille de la matrice est proportionnelle à la taille de l'espace vectoriel et au nombre de documents de la collection, et est généralement très grande ce qui rend les coûts calculatoires des opérations effectuées sur cette matrice parfois prohibitifs.
Une autre conséquence associée à la taille de l'espace est connue sous le nom de problème des hautes dimensionnalités, qui se traduit par le fait que plus un espace est grand plus la distance entre deux objets les plus éloignés est voisine de la distance séparant les objets les plus proches de cet espace. Ainsi, on cherche donc souvent à réduire la dimension de l'espace de représentation pour éviter ce phénomène même si en pratique il ne semble pas trop affecter les SRI.

Modèle GVSM

Le modèle Generalised Vector Space Model (GVSM) [129] est une extension du modèle vectoriel classique. Il a été developpé dans le souci de répondre aux critiques selon lesquelles les unités lexicales ne sont pas de bonnes bases pour l'espace vectoriel classique puisqu'ils ne sont pas indépendants. L'idée de base de ce modèle est de se placer dans un espace de représentation dual où les documents servent à des unités lexicales et non plus l'inverse. L'avantage est que les documents formant les bases de l'espace sont plus facilement considérés comme indépendants les uns des autres [18].

Modèle LSA

Le modèle de l'analyse de la sémantique latente [33] est une extension du modèle vectoriel [115]. Il propose de transformer la représentation traditionnelle en une représentation plus sémantique, qui a pour but de favoriser le rapprochement de documents et requêtes sémantiquement similaires.
Partant du principe qu'une représentation vectorielle traditionnelle basée uniquement sur les unités lexicales contient trop de bruit. Il propose en se basant sur une décomposition en valeurs singulières de la matrice pondérée classique d'occurrences des termes d'indexation dans les documents de la collection, de créer un espace vectoriel plus petit où les dimensions ne sont plus représentées par les unités lexicales mais par une combinaison linéaire de ces dernières. Ces combinaisons sont susceptibles de mieux faire ressortir les affinités sémantiques latentes entre les unités lexicales ou entre unités textuelles (phrases, paragraphes ou documents).
La méthode LSA utilise une matrice A (unités lexicales * unités textuelles) qui est composée des vecteurs d'unités lexicales des unités textuelles comme pour le modèle vectoriel standard. Ensuite, la matrice A est décomposée en valeurs singulières, cette décomposition dont le symbolisme est représenté ci-dessous va permettre de créer un nouvel espace vectoriel :

$$A = U_0 S_0 V_0' \tag{2.3}$$

Avec, $A = [a_{ij}]$, a_{ij} est la fréquence d'apparition de l'unité lexicale i dans le document j,
S_0 est la matrice diagonale de valeurs singulières,
U_0 et V_0 sont orthogonales,
t est le nombre de lignes de A,
d est le nombre de colonnes de A,
m est de l'ordre de A (=min(t,d)).

La représentation par unités lexicales contient beaucoup de bruits, ces bruits se retrouvent dans les dimensions de S_0 qui ont des valeurs faibles. Le modèle LSA supprime ces dimensions de valeurs faibles (en les remplaçant par la valeur 0), ce qui diminue la dimension de S_0 à k, cette matrice modifiée est appelée maintenant S. Par conséquent, les matrices U_0 et V_0' nettoyées deviennent U et V'.
Ainsi un nouvel espace vectoriel est obtenu:

$$A = U_0 S_0 V_0' = \tilde{A} = USV' \tag{2.4}$$

Où la matrice $\tilde{A}$ de rang k, la plus proche de A par la méthode des moindres carrés est unique. Ce modèle permet donc de représenter les documents dans un espace de dimension k. Il permet, de façon symétrique, de représenter les unités lexicales des vecteurs qui sont une indication du profil de co-occurrence d'une unité lexicale dans les documents.

2.3.3 Modèles probabilistes

Les modèles probabilistes, dont une présentation très complète est faite par [121], tentent quant à eux de modéliser la notion de pertinence. Le modèle probabiliste représente la probabilité de la pertinence d'un document D par rapport à une requête R. Le but de cette fonction de similarité dans ce modèle est d'essayer de séparer les documents pertinents des non pertinents au sein d'une collection. L'idée de base, dans ce modèle, est de tenter de déterminer les probabilités P(R/D) et P(NR/D) pour une requête donnée. Cette probabilité signifie : si on retrouve le document D, quelle est la probabilité qu'on obtienne l'information pertinente et non pertinente. [111] énonce que la présentation des documents à l'utilisateur dans l'ordre décroissant des probabilités est optimale dans un cadre de RI. Selon ce principe, les documents retournés sont ceux dont la probabilité de pertinence est supérieure à la probabilité de non pertinence.

2.3.4 Description détaillée du modèle vectoriel

Dans cette sous section, nous détaillons le modèle vectoriel que nous utilisons dans nos expérimentations. Nous présentons le fonctionnement du modèle en soulignant ses divers paramètres, valables pour d'autres modèles de représentation, tels que le choix des termes d'indexation, les schémas de pondérations et les différentes mesures de similarités existantes.

Choix des termes d'indexation

Le choix de termes d'indexation est une étape très importante car ces derniers constituent la structure de l'espace dans lequel seront représentés les documents. Ils doivent être le plus discriminant possible. Ces termes d'indexation ne devraient pas être trop nombreux car ce sont eux qui vont déterminer la taille de l'espace vectoriel et donc la complexité des calculs de similarité.

Analyse basée sur les fréquences d'occurrences des unités lexicales

Elle consiste à choisir les termes d'indexation en fonction de leur fréquence d'apparition dans les textes. Il se base sur des méthodes numériques qui trouvent principalement leurs origines dans la loi de Zipf. Les travaux de Zipf montrent que si les unités lexicales sont rangés par ordre décroissant de leur fréquence d'apparition au sein d'un texte (ou d'une collection), il existe alors une relation entre le rang de ces termes d'indexation et leur fréquence. Cette relation peut s'exprimer par la relation suivante :

rang * (fréquence de l'unité lexicale/nombre d'unités lexicales) = constante

qui signifie que si le rang d'une unité lexicale est multiplié par le nombre de fois où il apparaît dans les textes, on aura tendance à trouver un nombre constant. Par exemple, si l'unité lexicale la plus fréquente d'un texte (rang = 1) apparaît 1000 fois, la deuxième unité lexicale aura tendance à se trouver 500 fois dans le texte et ainsi de suite. À la fin de cette liste, on trouvera 1000 unités lexicales n'ayant été utilisées qu'une seule fois dans le texte. La loi de Zipf est l'une des premières à avoir montré que les unités lexicales dans les documents ne s'organisent pas de manière aléatoire.

Pondération

Une fois les termes d'indexation sont choisis, il est intéressant de montrer que tel terme est plus important que les autres pour décrire un document. Cela consiste à transformer l'occurrence d'une unité lexicale dans un document par une combinaison de pondérations locales L(i,j), indiquant l'importance de l'unité lexicale i dans l'unité textuelle j; pondérations globales G(i), indiquant l'importance de l'unité lexicale i dans l'ensemble des unités textuelles de la collection ; et normalisations N(j), pénalisant la variation de la longueur entre les différentes unités textuelles de la collection. Nous présentons ci-dessous quelques formules mettant en pratique ces trois niveaux de pondération.

1. Pondération locale L(i,j)
   La pondération locale d'une unité lexicale cherche à mesurer l'importance d'une unité lexicale au sein d'un document. Cette pondération est fonction de la fréquence de l'unité lexicale dans le document, notée tf (term frequency). L'idée est que si une unité lexicale apparaît souvent dans un document, il est plus pertinent pour décrire le contenu du document qu'une unité lexicale n'apparaissant que rarement.
   Le tableau 2.1 présente quelques mesures classiques, outre la fréquence simple de l'unité lexicale, on trouve une pondération binaire qui remplace toute fréquence d'unité lexicale supérieure ou égale à 1 par 1. En utilisant cette pondération, on se ramène donc au cas ensembliste.
   La pondération normalisée quant à elle permet de prendre en compte non seulement la fréquence de l'unité lexicale t_i dans d_j, mais aussi de mesurer son importance relativement aux autres unités lexicales de d_j. Le poids w_i d'une unité lexicale est ainsi compris entre 0 et 1.
   La pondération logarithmique [16] a pour but de diminuer l'influence des grandes valeurs. Si l'on veut au contraire donner plus de poids aux unités lexicales très fréquentes, il est possible d'utiliser des pondérations telles que la fréquence au carré.

2. Pondération globale G(i)
   Le calcul d'une pondération globale permet d'exploiter l'ensemble des documents. Il est basé sur le nombre de documents de la collection dans lesquels l'unité lexicale considérée apparaît. Les mesures les plus utilisées est la fréquence inverse, notée idf présentée dans le tableau 2.2.

Signification	Formule
binaire	$w_i = \begin{cases} 1 & \text{si} \quad tf(t_i) \geq 0 \\ 0 & \text{sinon} \end{cases}$
fréquence	$w_i = tf(t_i)$
fréquence normalisée	$w_i = \frac{tf(t_i)}{maxtf(t)}$
logarithme de fréquence	$w_i = \begin{cases} 1 + log(tf(t_i)) & \text{si} \quad tf(t_i) \geq 0 \\ 0 & \text{sinon} \end{cases}$
fréquence au carré	$w_i = (tf(t_i))^2$

Table 2.1 – *Formules de pondération locale*

Quatre pondérations globales bien connues sont : *Normal*, *Gfldf*, *Idf*, et Entropie [41]. Chacune est définie en termes de fréquence des unités lexicales $tf(t_{ij})$, de fréquence des documents *dfi* (le nombre des documents auxquels l'unité lexicale i appartient), et de fréquence globale *gfi* (le nombre total de fois où l'unité lexicale i apparaît dans la collection), avec N le nombre de documents et M le nombre des unités lexicales dans le corpus.
La pondération "*Normal*" normalise les longueurs de chaque ligne à 1, elle a pour effet de donner un poids élevé aux unités lexicales peu fréquentes et elle ne dépend que de la somme des fréquences au carré et pas de la distribution de ces fréquences.
GfIdf et *Idf* pondèrent les unités lexicales par le nombre des documents différents dans lesquels ils apparaissent. La différence entre les deux c'est que *GfIdf* augmente le poids des unités lexicales fréquentes.
L'entropie est la seule méthode qui tient compte de la distribution des unités lexicales dans les documents. L'incertitude moyenne ou « entropie » d'une unité lexicale est donnée par $\sum_j^N \frac{p_{ij} \log p_{ij}}{\log(N)}$. Soustraire cette quantité à une constante permet d'attribuer un poids minimum aux unités lexicales qui sont distribuées de la même façon dans tous les documents et un poids maximum aux unités lexicales qui sont concentrées dans quelques documents.

3. Normalisation
   Deux principales raisons rendent l'utilisation de la normalisation nécessaire:
   - Les hautes fréquences : les longs documents emploient habituellement les mêmes unités lexicales à plusieurs reprises. Par conséquent, la fréquence de ces unités lexicales peut être grande, augmentant la contribution moyenne de ses derniers dans le calcul de similarité.
   - Le nombre des unités lexicales : généralement, dans les longs documents le vocabulaire est plus riche et varié que dans les courts, ce qui augmente le nombre des unités lexicales en communs entre une requête et un long document; d'où l'augmentation de la similarité requête-document et la chance de récupération des documents longs au détriment des plus courts.
4. Combinaison des pondérations
   La combinaison des pondérations s'appuie sur trois facteurs : un facteur de pondération locale qui quantifie l'importance de l'unité lexicale dans le document, un facteur de pondération globale

Signification	Formule
normal	$\frac{1}{\sum_j^N tf(t_{ij})}$
GfIdf	$\frac{gf_i}{df_i}$
Idf	$\log_2 \frac{N}{df_i}$
entropie	$1 - \sum_j^N \frac{p_{ij} \log p_{ij}}{\log(N)}$ $\quad p_{ij} = \frac{tf(t_{ij})}{gf_i}$

Table 2.2 – *Formules de pondération globale*

qui mesure la représentativité de l'unité lexicale dans l'ensemble de la collection de documents, et un facteur de normalisation qui prend en considération la taille du document. Les schémas de pondération classiques utilisés dans la littérature sont : la pondération *Tfc*, la pondération *Ltc* et la pondération Okapi BM-25 [31].

- *Tfc* : $\frac{f_{ij} * df_i}{\sqrt{\sum_{k=1}^{M} (f_{kj} * df_k)^2}}$

La pondération Tfxidf ne prend pas en considération le fait que les unités textuelles peuvent être de différentes longueurs; ainsi la pondération Tfc est semblable au Tfxidf à la différence que la *Tfc* emploie la longueur normalisée;

- *Ltc* : $\frac{\log(f_{ij}+1) * df_i}{\sqrt{\sum_{k=1}^{M} (\log(f_{kj}+1) * df_k)^2}}$

La pondération *Ltc* est une approche légèrement différente, qui emploie le logarithme de la fréquence d'une unité lexicale, de ce fait elle réduit les effets de grandes différences dans les fréquences.

- Okapi-BM25 : $\frac{3*(\log N - \log df_i)*f_{ij}}{2*(0.25+(0.75*\frac{N*dl_j}{\sum_{k=1}^{N} dl_k}))+f_{ij}}$

où dl est la longueur du document L'Okapi BM25 repose sur le calcul des poids des unités lexicales avec prise en compte de la fréquence d'apparition des unités lexicales dans le document et la requête ainsi qu'un facteur de correction tenant compte de la longueur de document.

Mesures de similarité

La mesure de similarité entre documents, établie sur la base de leur représentation dans l'espace vectoriel, a elle aussi fait l'objet de nombreuses études. Nous présentons ici brièvement les mesures les plus couramment utilisées ; pour une revue plus complète, se référer notamment à [12]. On notera par la suite D et Q deux vecteurs (pondérés), représentant dans le cadre de la RI un document d et une requête q. Mesures ensemblistes Les mesures ensemblistes comparent la proximité de deux vecteurs en utilisant seulement l'information de la présence ou de l'absence d'une unité lexicale dans un document (correspondant au facteur binaire de pondération locale). Les mesures les plus utilisées dans ce cadre sont les coefficients de Dice et de Jaccard. Ainsi, la distance issue du coefficient de Dice s'écrit :

$$\delta_{Dice}(D, Q) = 2\frac{|D \cap Q|}{|D| + |Q|} \tag{2.5}$$

De Jaccard s'écrit :

$$\delta_{Jaccard}(D,Q) = \frac{|D \cap Q|}{|D \cup Q|} \tag{2.6}$$

Mesures géométriques Les mesures géométriques sont considérées comme une extension des mesures ensemblistes dans le cas où les pondérations des termes d'indexation sont prises en compte. Ainsi aux opérations ensemblistes correspondent les opérations vectorielles telles que le produit vectoriel (noté par $\cdot$) ou le calcul de norme.
La mesure cosinus est la plus utilisée pour le calcul de similarité, c'est le cosinus de l'angle entre les vecteurs représentant les documents :

$$\delta_{cos}(D,Q) = \frac{D \cdot Q}{\|D\|\|Q\|} = \frac{\sum_{i=1}^{n} w_{d,i} w_{q,i}}{\sqrt{\sum_{i=1}^{n} w_{d,i}^2 \sum_{i=1}^{n} w_{q,i}^2}} \tag{2.7}$$

L'intérêt du cosinus est qu'il normalise les vecteurs par la longueur des documents.
D'autres mesures géométriques classiques, sont utilisées comme la distance euclidienne L2.

$$\delta_{L2}(D,Q) = \|D - Q\|_{L2} = \sqrt{\sum_{i=1}^{n} (w_{d,i} - w_{q,i})^2} \tag{2.8}$$

et la distance L1 :

$$\delta_{L1}(D,Q) = \|D - Q\|_{L1} = \sum_{i=1}^{n} (|w_{d,i} - w_{q,i}|) \tag{2.9}$$

Si les vecteurs des documents sont normalisés par la norme euclidienne, la distance euclidienne est monotone par rapport à la mesure du cosinus :

$$\frac{\delta_{L2}(D,Q)^2}{2} = 1 - \delta_{cos}(D,Q) \tag{2.10}$$

Les deux mesures seront donc dans ce cas équivalentes pour le classement des document retournés. Mesures distributionnelles Dans le cas où les vecteurs D et Q sont normalisés suivant la norme L1, et peuvent donc être interprétés comme des distributions de probabilité [87]. Des mesures pour la dissimilarité entre distributions de probabilité sont également proposées.
On peut par exemple définir une distance χ^2 dont quelques propriétés intéressantes sont données par [108] :

$$\delta_{\chi^2}(D,Q) = \sqrt{\sum_{i=1}^{n} \rho_i (w_{d,i} - w_{q,i})^2} \tag{2.11}$$

Avec

$$\rho_i = \frac{N}{\sum_{d \in N} w_{d,i}}$$

Dans ce cadre, on peut utiliser la divergence de Kullback-Leibler (KL), ou bien encore la Divergence de Jensen-Shannon. Une présentation plus développée de ces mesures est détaillée dans les travaux de [12].

2.3.5 Critères d'évaluation des SRI

La pertinence est une connaissance très complexe à évaluer. Ainsi, elle dépend fortement de l'utilisateur, qui est le seul à savoir si le document retourné par le système correspond à sa recherche initiale. Il est néanmoins essentiel de disposer de techniques d'évaluation solides qui, en définissant des mesures précises, permettent de juger l'efficacité des SRI à retrouver des documents pertinents, quels que soient les méthodes d'indexation, de recherche ou les modèles qu'ils implémentent.

Rappel et précision

Les deux mesures communément utilisées pour évaluer un système de recherche d'information sont le taux de précision et celui de rappel [114]. Ces deux mesures peuvent être définies par :

$$précision = \frac{nombre\ total\ de\ documents\ pertinents\ retrouvés\ par\ le\ système}{nombre\ total\ de\ documents\ retrouvés\ par\ le\ système}$$

$$rappel = \frac{nombre\ total\ de\ documents\ pertinents\ retrouvés\ par\ le\ système}{nombre\ total\ de\ documents\ pertinents\ dans\ la\ collection}$$

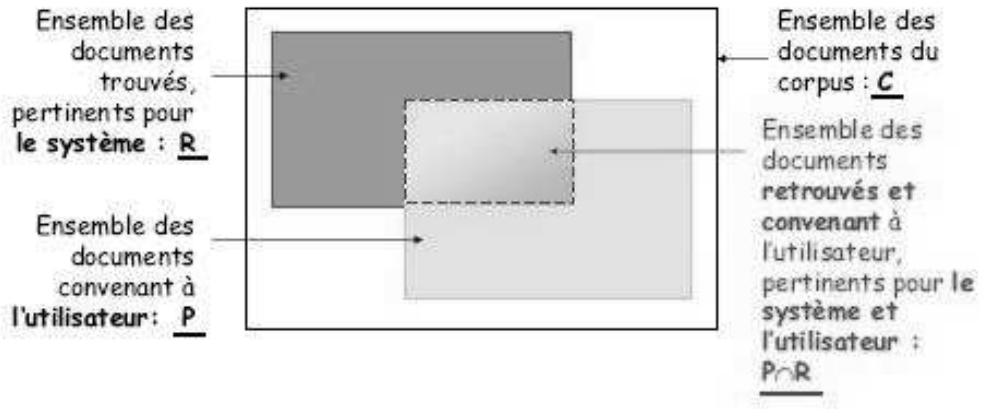

Figure 2.2 – Rappel/Précision

Courbe Rappel/Précision

Les performances d'un SRI peuvent être représentées par une courbe Rappel/Précision. Lorsque les valeurs exactes de rappel ne peuvent pas être atteintes, Il est fréquent d'employer une interpolation sur ces courbes, qui consiste à lisser la courbe initiale pour qu'elle soit décroissante. La valeur interpolée de la précision pour un point de rappel i est la précision maximale obtenue pour un point supérieur ou égal à 1. L'avantage est de définir la précision sur des valeurs standardisées.

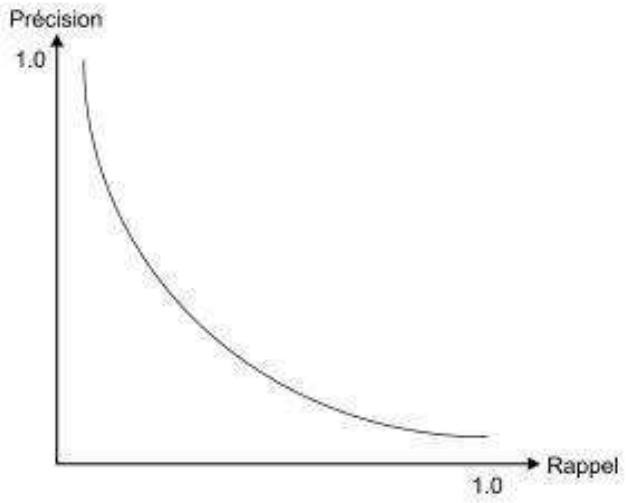

Figure 2.3 – Exemple de courbe précision/rappel

Le calcul de la précision et du rappel s'effectue pour chaque élément de la liste des documents retrouvés par le SRI. Pour évaluer la performance globale d'un système, il est utile de disposer d'une mesure unique qui réunie en une seule grandeur la performance du SRI.

Mesures globales

La précision moyenne interpolée IAP (Interpolated Average Precision) est une mesure décrivant la précision globale du système évalué sur une requête. Elle consiste à calculer la précision des résultats sur onze points de rappel qui vaut 0, 10, 20, 30, 40, 50, 60, 70, 80, 90 et 100%. Si ces points ne sont pas atteints, les mesures sont alors interpolées. La moyenne de ces 11 précisions forme la précision moyenne interpolée.

La R-précision calcule, pour une requête, la précision obtenue pour un nombre donné de documents retrouvés par le système. Ce nombre est fixé pour chaque requête en fonction du nombre de documents pertinents dans la collection. La R-précision est intéressante lorsque la collection comporte un nombre considérable de documents pertinents.

La F-mesure [125] est définie comme la combinaison pondérée du taux de rappel et du taux de précision :

$$F = \frac{2P * R}{P + R}$$

Où P et R représentent respectivement les résultats de précision et de rappel.

2.4 Conclusion

Le but de ce chapitre était de présenter le domaine de la recherche d'information, de décrire plus particulièrement les principales étapes à savoir l'indexation et la recherche, d'introduire les principaux modèles sur lesquels se basent les SRI, et de présenter les méthodes d'évaluation adoptées pour attester de la validité des mécanismes implémentés au coeur de ces systèmes.

En décrivant dans ce chapitre les mécanismes de RI, nous avons pu faire ressortir plusieurs limites pouvant expliquer ces résultats limites qui n'excèdent pas les 30 %-40%. La première est associée aux méthodes exploitées pour représenter les contenus textuels. Le passage du document ou de la requête en texte intégral à une représentation en " sac de mots ", telle qu'elle est présentée par la plupart des modèles de RI, implique des pertes d'informations conséquentes. Cette représentation ignore les relations susceptibles d'exister entre les unités lexicales. Or, dans la langue, les unités lexicales ne sont pas agencés les uns à la suite des autres par hasard ; chaque phrase a une structure significative ; chaque unité lexicale entretient des relations avec les autres.
La seconde limite concerne le processus d'appariement qui peut se résumer à une simple comparaison d'unités lexicales. Les faiblesses d'un tel processus de mise en correspondance s'articulent autour de deux problèmes. Le premier problème, est la polysémie des unités lexicales ("avocat" qui désigne à la fois le fruit, la fonction). L'ambiguïté qui en découle conduit à une baisse de précision des systèmes puisqu'elle implique la récupération de documents non pertinents. Le deuxième problème est la possibilité offerte par le langage naturel de formuler de différentes manières une même idée. Un document pertinent peut ainsi contenir des termes " sémantiquement " proches de ceux de la requête mais toutefois différents (des synonymes : voiture vs automobile ou encore des unités lexicales ayant une forme morphologique différente : indexer vs indexation).
En fonction de ces difficultés associées à la complexité du langage naturel, une solution souvent évoquée est d'employer des unités plus fines que les chaînes graphiques pour représenter les documents et requêtes. Ces nouvelles unités sont obtenues par une analyse linguistique des contenus textuels réalisée à l'aide des techniques du traitement automatique des langues (TAL), et permet aux SRI un appariemment plus pertinent des documents et requêtes

CHAPITRE 3
Impact du TAL en RI

Les techniques du traitement automatique des langues permettent d'extraire des textes des informations plus riches que de simples unités lexicales. Ces informations de nature morphologique, syntaxique et sémantique ont été partiellement utilisées en RI pour améliorer les méthodes d'appariement, les représentations des contenus des documents et requêtes et le processus de recherche. Ce chapitre établit un tour d'horizon sur l'impact de ces différentes informations linguistiques issues par des techniques du TAL sur les systèmes de recherche d'information et sur leurs performances.

3.1 Impact des connaissances morphologiques en recherche d'information

L'avantage d'une analyse morphologique, en RI appliquée aux documents et requêtes, est de reconnaître que les unités lexicales : créer, créateur, créatrice, quoique graphiquement différents, représentent différentes formes d'une même unité lexicale, appelées variantes morphologiques. L'appariement de ces formes, peut ainsi apparaître pertinent.

Dans cette section, nous nous intéressons au traitement de la variation morphologique en RI. Nous faisons état des expériences qui ont exploité des informations morphologiques au sein des systèmes, en découpant notre analyse suivant les différents outils d'analyse qu'il est possible d'appliquer sur les documents et les requêtes.

3.1.1 Traitement de la variation morphologique en RI

Les différentes expériences réalisées n'utilisent pas les mêmes types de traitements morphologiques pour évaluer l'intérêt d'employer un niveau d'analyse morphologique des documents et requêtes en RI. Dans le but d'établir un bilan efficace, nous présentons dans un premier temps l'apport en RI d'une procédure de racinisation ensuite nous nous intéressons à l'impact des analyseurs morphologiques.

Utilisation d'un racineur

Il y a deux façons principales d'appliquer une procédure de racinisation (stemming) dans un SRI. La première consiste à l'utiliser lors de la phase d'indexation. Les termes des documents et des requêtes sont ramenés à leurs racines et l'appariement entre les termes de la requête et ceux des documents s'effectue sur cette base. La deuxième consiste à l'employer pour la tâche d'extension de requêtes. Ces dernières sont alors enrichies à l'aide de termes morphologiquement liés à ceux qu'elles contiennent, généralement par le biais de familles morphologiques. Les travaux de [88] et [65] mesurant l'apport de la racinisation pour l'anglais mènent à des conclusions globalement décevantes puisqu'aucune amélioration de résultats n'est remarquée par rapport à un SRI traditionnel. Les mêmes observations sont obtenues par [51],

qui comparent l'impact respectif de quatre types de racineurs (les algorithmes de Porter, de Lovins, un racineur qui supprime simplement les marques du pluriel et un racineur à base de dictionnaires). Leur conclusion est que les performances apportées par ce traitement restent insuffisantes. Les expériences de [82] se montrent prometteuses puisque la racinisation conduit à une augmentation des résultats située entre 1,3% et 45,3% selon les collections et les techniques de racinisation utilisées. Les améliorations les plus conséquentes sont obtenues dans le cas de documents courts (environ 45 unités lexicales) associés à des requêtes courtes (comportant 7 unités lexicales en moyenne). D'autres expériences relatives à l'utilisation de la racinisation apparaissent encourageantes pour la langue arabe [85]. Il découle de ces différentes expériences que l'apport de la procédure de racinisation est tributaire. Ainsi dans certains cas, la racinisation s'avère très bénifique alors que dans d'autres travaux, elle peut entraîner une dégradation des performances. Dans le but d'avoir une idée générale sur l'avantage des connaissances morphologiques au sein des SRI, nous nous intéressons aux systèmes qui intègrent une analyse morphologique s'effectuant par le biais de lemmatiseurs et d'analyseurs dérivationnels. Notre choix de ne pas faire de séparation dans notre présentation entre ces deux types d'outils se confirme par le fait qu'au sein des diverses expériences que nous allons décrire, ces traitements sont souvent couplés en raison de leur complémentarité.

Utilisation d'analyseurs morphologiques flexionnels et dérivationnels

Les analyseurs flexionnels et dérivationnels peuvent intervenir sur deux niveaux d'un SRI. Ils peuvent être appliqués pour l'indexation des documents et requêtes. L'appariement entre document et requête s'effectue sur la base du lemme (pour les analyseurs flexionnels) ou sur la base de la racine (pour les analyseurs dérivationnels). Aussi, ils peuvent être appliqués pour l'extension des requêtes. Il s'agit d'acquérir les lemmes des termes des requêtes et d'enrichir ces dernières à l'aide des familles morphologiques constituées à partir d'analyse flexionnelle et dérivationnelle. Les expérimentations de [54] pour le français montrent l'influence de la morphologie flexionnelle en RI. Une intégration d'un module de lemmatisation présente une amélioration de la précision moyenne de 16%. Les expériences de [135] appliquées au domaine medical, montrent l'impact faible de la lemmatisation et de dérivation dans une tâche d'appariement entre requêtes et termes normalisés. L'emploi des connaissances flexionnelles et dérivationnelles améliore en moyenne les réponses à une requête. La flexion agit sur 6,6% des cas avec une hausse modeste, et la dérivation agit dans 2% des cas avec une augmentation plus nette. Les expériences de [126] pour l'espagnol emploient successivement ces deux types d'analyseurs. La première étape consiste à étiqueter morpho-syntaxiquement les unités lexicales et à obtenir les lemmes des textes à indexer. La deuxième étape consiste à remplacer les termes par le représentant de la famille morphologique à laquelle il appartient. Une augmentation du rappel est constatée par rapport à une approche de racinisation.
Les expériences présentées montrent l'avantage de recourir à des connaissances morphologiques pour améliorer les performances d'un SRI. L'apport des outils de racinisation et d'analyse flexionnelle et dérivationnelle en RI est tributaire d'un certain nombre de facteurs à savoir le type de collection utilisé (longueur des requêtes, taille des documents...) ou la langue prise en compte.

3.2 Impact des connaissances syntaxiques en recherche d'information

L'avantage de recourir à des connaissances syntaxiques en RI est d'aller au delà de la notion de chaînes de caractères, en prenant en compte principalement des unités lexicales, composées de plusieurs termes (*i.e pollution de l'air*), plus caractéristiques et moins ambiguës. Dans l'objectif d'avoir une idée

sur l'impact de ces connaissances en RI, nous nous intéressons aux résultats d'expériences qui les intègrent au sein des systèmes. Après avoir présenté des notions sur la syntaxe, nous abordons les différentes expérimentations qui utilisent ces connaissances en RI. Nous finissons en présentant les différentes adaptations nécessaires aux SRI visant à manipuler de telles connaissances.

3.2.1 Notions de syntaxe

A l'issue d'une analyse morphosyntaxique, les formes initiales présentes dans un énoncé sont substituées par une liste ordonnée d'éléments contenant un certain nombre d'informations, parmi lesquelles la catégorie grammaticale et, éventuellement, le lemme, les flexions ou d'autres connaissances dont la présence dépend de l'application souhaitée.

La syntaxe décrit comment ces éléments s'ordonnent pour créer des constituants. De ce fait, on représente souvent le résultat d'une analyse syntaxique de façon hiérarchique. De nombreux formalismes ont été proposés pour l'analyse syntaxique ; nous présentons ici trois types d'analyses syntaxiques à savoir l'analyse superficielle, la grammaire hors-contexte et l'analyse syntaxique en dépendance. L'analyse superficielle (ou partielle) a pour but de reconnaître les syntagmes simples, non récursifs, d'un énoncé, sans les lier les uns aux autres. Ainsi, les principaux attachements responsables des ambiguités citées ci-dessus, comme les attachements prépositionnels, ne sont pas traités (du moins dans un premier temps). L'idée est d'obtenir des résultats certes moins riches, mais plus rapides et plus sûrs. Elle est appropriée pour un certain nombre d'applications qui ne cherchent pas à définir des dépendances précises, comme par exemple la reconnaissance de syntagmes nominaux. Cette approche s'oppose à l'analyse profonde (ou complète), qui cherche à regrouper chaque phrase dans une unique représentation.

Une grammaire hors-contexte (ou CFG, pour context-free grammar) se compose d'un ensemble de règles de la forme :

$$\text{E} \longrightarrow \text{E1} \ldots \text{En}$$

qui expriment le fait que la séquence d'expressions E1 . . . En peut être réécrite par un nouvel identifiant unique E, en faisant abstraction des éléments qui l'entourent.

Règles		*Exemples*
S	→ *NP VP*	[Le loup]$_{NP}$ [sort de la forêt]$_{VP}$
NP	→ *Pronom*	il
NP	→ *Nom_Propre*	Paul
NP	→ *Det Adj? Nom Adj?*	[Le]$_{DET}$ [petit]$_{ADJ}$ [chaperon]$_{NOM}$ [rouge]$_{ADJ}$
NP	→ *NP PP*	[La fille]$_{NP}$ [de Minos et de Pasiphaé]$_{PP}$
VP	→ *Verbe PP*	[sort]$_{Verbe}$ [de la forêt]$_{PP}$
VP	→ *Verbe NP*	[mange]$_{Verbe}$ [le chat]$_{NP}$
PP	→ *Prep NP*	[de]$_{Prep}$ [la forêt]$_{NP}$
...	→ ...	...

Figure 3.1 – Règles hors-contexte permettant d'obtenir les constructions

L'application de ces règles, illustrées à la figure 3.1 par des regroupements de constituants linguistiques, est souvent représentée de façon arborescente, comme le montre la figure 3.2. Les CFG permettent

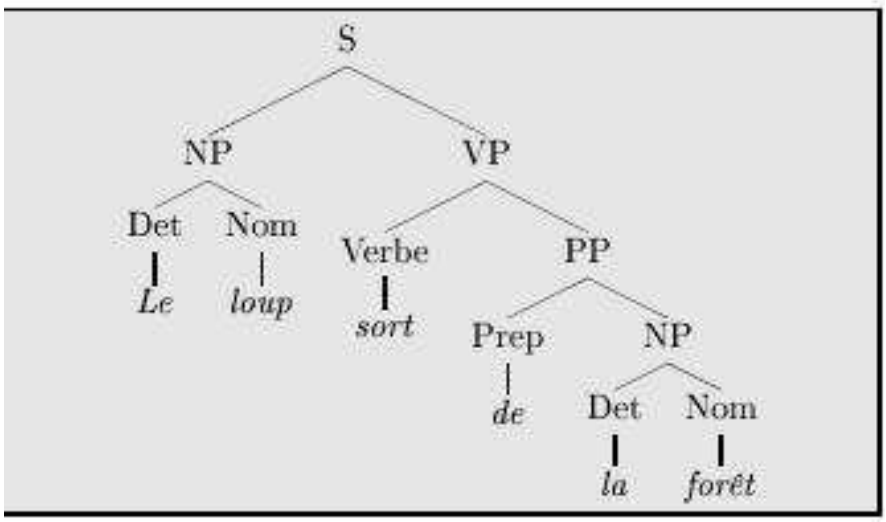

Figure 3.2 – Arbre syntaxique de la phrase "Le loup sort de la forêt"

de refléter les aspects hiérarchiques des constructions grammaticales par l'imbrication des constituants et la récursivité des règles.

L'analyse syntaxique en dépendance diffère surtout de l'analyse en constituants (comme les règles hors-contexte) par le mode de représentation. Les deux approches n'ont pas de différence au niveau de leur couverture ou de leur expressivité. L'idée vise à mettre en évidence les relations de dépendance entre les unités lexicales " têtes " (éléments centraux de syntagmes) et les unités lexicales " modifieurs " ou "expansion" qu'ils régissent, que ce soit globalement au niveau de la phrase (où le prédicat verbal constitue fréquemment le mot tête principal régissant en particulier le sujet) ou à l'intérieur d'un syntagme. Ainsi, selon [123], les groupes d'unités lexicales "information retrieval", "retrieval of information", "retrieve more information" et "information that is retrieved" ... peuvent être réduits à la relation retrieve+information où "retrieve" est l'élément tête et "information" son modifieur.

3.2.2 Utilisation des connaissances syntaxiques au sein d'un SRI

L'utilisation des connaissances syntaxiques en RI se limite généralement à la prise en compte de syntagmes. Tout d'abord, nous présentons comment ceux-ci pouvaient être exploités en RI, ensuite nous nous intéressons à leur apport sur les performances des systèmes.

Utilisation des syntagmes en RI

D'après le type d'analyse syntaxique appliquée aux textes, les syntagmes utilisés en RI peuvent être de diférentes formes . Tout d'abord, nous abordons les deux principaux types de syntagmes généralement pris en compte ensuite nous nous intéressons à leur intégration au sein des systèmes. Les syntagmes Un syntagme, c'est un concept linguistique : c'est un groupe de unités lexicales qui, ensemble, produisent

un sens unique. Ils peuvent être des termes complexes [95], que nous considérons comme toute unité lexicale constituée d'au moins deux unités lexicales pleines [1], auxquels peuvent s'associer des déterminants et des prépositions (en français par exemple la structure Nom Prép Nom (*i.e. pommes de terre*)). Les termes complexes remplacent ainsi les termes simples en tant que termes d'indexation. Leur avantage par rapport aux termes simples est qu'ils sont plus susceptibles de désigner des concepts puisqu'ils réfèrent un domaine de connaissance spécialisé.
L'extraction de ces termes n'exige pas obligatoirement une analyse syntaxique très poussée des textes. Certains techniques d'acquisition utilisent des indices numériques (aspect fréquentiel). l'aspect fréquentiel utilisé est d'autant plus fiable et performant que le corpus est volumineux. Ces techniques [23] [86] cherchent à associer les unités lexicales apparaîssant ensemble dans un texte de manière statistiqument significative. Elles reposent pour la plupart sur l'évaluation de la probabilité que des sequences d'unités lexicales apparaîssent ensemble dans une certaine fenêtre de texte plus souvent que le hasard ne l'aurait permis. D'autres techniques plus poussées proposent, en se basant sur des indices structurels (aspect symbolique) [24], d'extraire des combinaisons d'unités lexicales à structure syntagmatique connue (*i.e. la structure Nom Prép Nom*), qui revèlent souvent des résultats pertinents pour la RI [68].
En plus de l'extraction des termes complexes, autres travaux [26] [27] proposent de prendre en compte les variantes des termes complexes. Ces variantes, repérées par le biais d'une analyse syntaxique plus fine, sont normalisées, réduites à une forme unique qui est alors exploitée comme terme d'indexation.
D'autres travaux procédent à une analyse syntaxique plus sophistiquée (*i.e une analyse syntaxique en dépendances*), il est possible de mettre en évidence les relations de dépendances entre les unités lexicales présentes au sein d'un syntagme, et de ressortir des relations tête+modifieur (*i.e tête+expansion*). La normalisation des syntagmes en structure tête+modifieur présente l'avantage de réduire les différentes variantes en une seule et même forme. L'apport des relations tête+modifieur sur les performances des SRI qui les utilisent est fortement tributaire, nous le verrons à travers les expériences, et la qualité de l'analyse syntaxique mise en oeuvre. Intégration dans un SRI Les syntagmes peuvent être pris en considération dans des SRI de deux façons. Soit dans la phase d'indexation, ils sont exploités comme termes d'indexation. L'appariement entre les documents et requêtes devrait permettre de retrouver plus de documents qu'une mise en correspondance à l'aide de termes simples étant donné que les documents contenant uniquement les variantes doivent aussi être retournés. Pour les structures tête+modifieur, le but est de favoriser les documents dans lesquels les termes simples composant le syntagme entretiennent la même relation de dépendance que dans la requête. Ainsi, si dans la requête le terme matériel est en position modifieur dans un syntagme (*i.e installation de matériel*), il est alors possible de laisser de côté tous les documents où matériel est en position tête (*i.e dans matériel de fabrication*). En expansion de requêtes, il convient d'identifier toutes les variantes des syntagmes qu'elles contiennent et de les enrichir par elles.
Nous abordons les résultats de différentes expériences intégrant ces syntagmes en RI.

Résultats de l'utilisation des syntagmes en RI

Nous présentons ici les expériences qui intégrent au sein des SRI des termes complexes, puis celles qui utilisent des syntagmes structurés en tête + modifieur. Utilisation des termes complexes En ce qui concerne les termes complexes "statistiques" extraîts par le biais d'une approche numérique, diverses

[1] D'un point de vue terminologique, la définition des termes complexes est souvent plus compliquée et fait l'objet de nombreuses discussions. Par exemple, une distinction est souvent faite entre les syntagmes lexicalisés qui figent une construction syntaxique (i.e fil de fer barbelé) et les unités lexicales formées par composition (i.e bébé-éprouvette). Dans certaines approches, seuls les premiers correspondent véritablement aux termes complexes.

expériences telles que celles [115] [45] montrent, lors de la phase d'indexation, que l'intégration des termes complexes obtenus par repérage de cooccurrences améliore la précision moyenne des systèmes par rapport à des SRI n'utilisant que des termes simples. Ces améliorations sont cependant dépendantes de plusieurs paramètres. Nous en distinguons essentiellement trois. Le premier est lié à la taille des collections utilisées lors des expérimentations [45]. Ainsi, plus le nombre de documents est grand plus les méthodes numériques d'extraction sont crédibles, puisqu'elles s'appuient essentiellement sur la notion de fréquence. Le deuxième est associé au domaine de la base documentaire utilisée. Les termes complexes, vivement représentés en domaine spécialisé, sont en général moins fréquents dans les collections généralistes et tendent à ne pas être extraits par des méthodes d'acquisition numériques. Le dernier concerne la langue de la collection. Ainsi, les expériences de [53] sur le français ne permettent pas de montrer l'intérêt de ces termes en RI.

L'apport de termes complexes "syntaxiques" extraits par le biais de méthodes symboliques, est également très variable. Les résultats obtenus dans [45] [89] ou [53] pour le français ne montrent aucune amélioration par rapport à une indexation par termes simples. Une des raisons justifiant ces faibles résultats est liée respectivement au nombre de requêtes considéré trop faible et aux tailles des collections trop petites. D'autres expériences [38] [68] montrent que, dans certains cas, l'intégration des termes complexes syntaxiques dans un SRI l'emporte sur une indexation à base des termes simples ou des termes complexes statistiques.

D'après ces diverses expériences, il est difficile de conclure de manière déterminée sur l'avantage ou non d'employer des termes complexes en RI, voire sur la méthode d'acquisition priviligiée de ces termes. Il est certain que leur efficacité est associée à la qualité des éléments extraits, mais dépend également de la représentation de ces termes dans les index. Utilisation des syntagmes structurés tête + modifieur Un nombre de travaux ont exploité les syntagmes structurés en tête+modifieur en complément des termes simples. Ainsi, les expériences de [124] [62] [5] confirment l'avantage d'intégrer ces structures pour l'anglais puisqu'une amélioration perceptible des performances des SRI en terme de précision et rappel est remarquée par rapport à une indexation par termes simples. Des travaux similaires ont été menées sur d'autres langues, comme celles de Pohlmann et Kraaj [81] pour l'allemand, de Vilares et al. [126] pour l'espagnol, et révèlent une amélioration significative des performances. D'aprés ces expériences, il est apparaît intéressant d'intégrer des syntagmes structurés en tête+modifieur au sein des SRI puisqu'une hausse significative des performances des systèmes est observée (cf. également les travaux de Zhai et al. [133] attestant ces conclusions). L'utilisation de ces structures est cependant fortement sensible à la longueur des requêtes : une requête courte pourra avoir peu de chances de contenir des syntagmes de ce type, rendant ainsi leur prise en compte inutile.
D'autres types de structures syntaxiques ont été utilisées en RI [119], obtenues par le biais d'une analyse syntaxique complète de la phrase visant à rendre compte de dépendances plus complexes entre les termes que les relations tête+modifieur (*i.e les dépendances entre syntagmes*). L'avantage de l'utilisation de ces structures en RI n'a cependant pas été démontré.

L'apport des syntagmes est ainsi étroitement associé à la façon dont les SRI sont adaptés à leur prise en compte. Nous abordons à présent les modifications qu'il est essentiel d'opérer sur les systèmes pour pouvoir utiliser pleinement la richesse de ces connaissances syntaxiques.

3.2.3 Adaptation des SRI pour l'intégration des connaissances syntaxiques

Les SRI, initialement élaborés pour prendre en compte des termes simples, ne sont pas toujours appropriés à accueillir des connaissances linguistiques poussées. Deux mécanismes traditionnels en RI doivent être modifiés pour la prise en compte des syntagmes. Le premier, vise les mesures de pondération adoptées pour mesurer leur pouvoir de représentativité du contenu textuel. Ces mesures se basent, principalement sur la notion de fréquence. Ainsi, les mesures telles que le term frequency (tf) * inverse document frequency (idf) sont inadaptées à ces termes moins fréquents que les termes simples et qui sont alors sous-pondérés. Par conséquent, comme le constate [121], une mauvaise pondération peut entraîner une dégradation des résultats dans le processus de recherche.
A partir de cette constatation, un certain nombre de nouvelles mesures de pondération ont été proposées. Une première méthode consiste à pondérer l'expression (terme complexe) en fonction des poids de ses composantes. Les résultats obtenus ne sont pas uniformes comme prouvé dans [45] [90]. D'autres travaux ont utilisé une pondération dite "syntaxique" [61] [5] basée sur les catégories grammaticales des composantes des termes, en favorisant par exemple un certains types de syntagmes (comme les nominaux), ou en accordant plus d'importance à certains éléments (la tête du terme par exemple) [5].
Ainsi, la tête d'un syntagme est considérée comme une entité prépondérante et il est nécessaire de la privilégier par rapport aux autres entités. La tête d'un syntagme ne peut avoir que la catégorie grammaticale Substantif . Cette catégorie grammaticale se voit alors associée un poids plus important que les autres catégories. Les catégories grammaticales des syntagmes sont classées selon la hiérarchie suivante :

- La catégorie des substantifs est la catégorie la plus porteuse d'information.
- Les catégories d'adjectif, verbe à l'infinitif, participe passé et adverbe ont un poids moyen,
- Les catégories de proposition, conjonction et article ont un poids nul.

Le deuxième élément à prendre en considération est la façon dont les syntagmes sont intégrés au sein des index, et leur combinaison avec les termes simples.

Une première expérience consiste à isoler ces deux types d'informations. Ainsi, les travaux de [62] concluent, après avoir évalué différentes stratégies d'intégration au sein du modèle vectoriel, que l'indexation par des expressions complexes est plus performante que celle effectuée par des termes simples. Suivant la méthode de [48], il remplace une représentation traditionnelle, regroupant l'ensemble des termes d'indexation dans un seul vecteur, par une représentation en deux sous-vecteurs différents. Cette stratégie d'indexation donne un meilleur classement des documents en tête de liste et augmente la précision. L'application des deux sous-vecteurs est aussi utilisée pour l'expansion des requêtes, qui sont enrichies avec des termes issus de ces deux parties. [78] appliquent la même technique mais cette fois ci, ils montrent que les composantes des termes complexes doivent être ajoutées à l'index des termes simples.

Or [124] et [6] utilisent le même index pour les termes simples et complexes. Ces structures complexes sont appliquées en complément (et non en remplacement) des termes simples. L'avantage de cette méthode est de se baser sur les termes simples des termes complexes pour correspondre un document et une requête, au cas où ces termes complexes ne peuvent être employés pour l'appariement.
La prise en compte des termes complexes semble intéressante dans une optique de RI. Plus particulièrement, l'exploitation de ces structures semble pertinente pour offrir une description plus riche du contenu informationnel. Leur influence sur les performances d'un SRI dépend essentiellement de la qualité des informations extraites des documents et requêtes. Leur impact est associé aussi à la façon dont ils sont intégrés au sein des SRI, et aux adaptations opérées dans les modèles pour leur prise en compte.

3.3 Impact des connaissances sémantiques en RI

L'emploi d'une analyse sémantique des documents et requêtes en RI cherche à obtenir des informations sur le sens des unités lexicales et sur les relations que ces dernières entretiennent entre eux. Dans cette section, nous décrivons l'exploitation de connaissances linguistiques en RI par la prise en considération de connaissances sémantiques. Nous abordons les divers types d'informations utilisables, puis nous nous intéressons à leur intégration dans les SRI d'une part lors de la phase d'expansion des requêtes, et d'autre part lors de l'indexation.

3.3.1 Types de connaissances sémantiques utilisables en RI

Les connaissances sémantiques qui peuvent être intégrées au sein des SRI ont plusieurs origines. Elles peuvent tout d'abord être originaire de bases lexicales existantes, telle que WORDNET [47]. Cette ressource représente la caractéristique d'envelopper la plupart des unités lexicales comme les noms, verbes, adjectifs et adverbes de la langue anglaise et de rendre compte des relations sémantiques qu'ils entretiennent. Une telle base générale n'est cependant pas adaptée pour un domaine spécialisé et le choix d'exploiter des données plus restreintes se heurte à l'absence de telles ressources. Ansi une des solutions est d'utiliser des informations sémantiques acquises automatiquement à partir des collections de textes [24]. L'extraction de ces connaissances en corpus peut être réalisée à l'aide de méthodes numériques. Ces approches permettent par exemple de découvrir des associations d'unités lexicales, correspondant à des termes complexes. Elles présentent aussi la possibilité, en partant d'une analyse des unités lexicales qui partagent les mêmes propriétés contextuelles, de faire apparaître des classes à caractère conceptuel d'unités lexicales et de repérer des relations paradigmatiques (e.g. synonymie, hyperonymie...) entre ces unités.

3.3.2 Approches d'intégration des connaissances sémantiques

Une première approche pour utiliser des connaissances sémantiques en RI consiste à exploiter des relations sémantiques disponibles pour étendre les requêtes et accéder à des documents pertinents (en utilisant un synonyme d'un terme de la requête à titre d'exemple). L'expansion joue un rôle important dans la précision de la requête de l'utilisateur par l'ajout de termes à la requête, ce qui permet de cibler le sens et rend la requête par la suite moins ambigue.

La deuxième approche consiste à introduire des connaissances sémantiques lors de l'indexation des documents et des requêtes dans le but d'enrichir les représentations. Les connaissances sémantiques exploitées varient selon le type d'utilisation.

Utilisation de connaissances sémantiques en expansion de requête

Les caractéristiques des connaissances sémantiques extraites du corpus sont liées à leur méthode d'acquisition. Dans cette sous-section, nous nous intéressons à travers les travaux décrits à l'intégration des connaissances extraites à l'aide d'approches numériques ensuite nous abordons celle basée sur une approche symbolique. Exploitation de connaissances sémantiques extraites par une approche numérique Les expériences décrites ici, enrichissent les termes d'une requête d'utilisateur avec les termes qui co-occurent entre eux dans les documents. Deux stratégies d'expansion sont envisageables : la première consiste à étendre chaque terme de la requête, alors que la deuxième étend la requête dans sa globalité et les termes ajoutés doivent être proches de l'ensemble des termes de la requête.

Les expérimentations de [52] illustrent la première stratégie, sans conduire à une hausse significative

des résultats. Celles de [102] montrent une amélioration significative lors des expérimentations. [102] justifient leur faible résultat par le fait que les techniques appliquées pour l'acquisition des cooccurrences favorisent l'extraction de termes de même fréquence. Les expériences de [106] illustrent la seconde stratégie qui prend en compte non plus chaque terme de manière isolée mais la requête dans sa globalité. Ils montrent que cette méthode conclut à une amélioration de la performance des SRI comprise entre 20% et 30%. Exploitation de connaissances sémantiques extraites par une approche symbolique Les expérimentations menées par [25] utilisent des connaissances sémantiques extraites par des méthodes symboliques. Elles décrivent un enrichissement des noms contenus dans les requêtes par des verbes qui leur sont liés par une relation spécifique. L'idée est de considérer que les noms ne sont pas les seules catégories qui permettent un apport sémantique en reformulation. Les résultats montrent une amélioration significative des performances du SRI testé.

D'autres travaux proposent l'utilisation du lien nom-verbe. C'est le cas des travaux de [57] qui part du principe que l'une des manières de caractériser sémantiquement un nom est d'extraire l'ensemble des verbes utilisés avec lui pour recenser ce qu'il permet de faire. Le système identifie des liens nom-verbe en se basant sur une analyse syntaxique partielle et de l'utilisation des patrons. Cette approche s'avère interessante pour les requêtes courtes définies ambigues.

Utilisation des connaissances sémantiques pour l'indexation

Nous distinguons deux types d'utilisation de connaissances sémantiques pour l'indexation. L'indexation conceptuelle basée sur des ontologies, applicable sur des domaines spécialisés et faisant usage d'un formalisme de représentation de connaissances, et l'indexation sémantique qui exploite pour enrichir l'indexation des documents et requêtes, des connaissances sémantiques soit générales, soit acquises en corpus. Indexation conceptuelle Plusieurs travaux se sont intéressés à l'indexation conceptuelle, les expériences de [22] ont développé un système conceptuel nommé ELEN (gEnie Logiciel rEcherche d'iNformation) destiné à l'interrogation de logiciels qui se base sur une représentation des connaissances par graphes conceptuels de sowa [120] (l'appariement requête-documents est donc un appariement de graphes). Or, les travaux de [130] emploient un modèle fondé sur une organisation taxonomique de la connaissance. Une taxonomie conceptuelle organisée selon une relation, qui relie des concepts plus généraux aux plus spécifiques, est utilisée pour apparier entre les termes de la requête et les documents. Cette taxonomie est batie sur des expressions extraites des textes grâce à une base lexicale existante et d'une analyse morphologique et syntaxique des textes. Les auteurs montrent une amélioration de la précision et du rappel en adoptant l'indexation conceptuelle par rapport à un système d'indexation classique.
L'indexation conceptuelle apparaît intéressante et permet de résoudre certains problèmes du langage naturel à savoir la polysémie par exemple en utilisant une indexation basée sur la notion de concepts et sur les relations entre concepts. Cependant, elle necessite des ressources considérables et fait appel à des techniques complexes.

L'un des points faibles de cette approche est lié au coût de sa mise en oeuvre. En outre, les performances des SRI basés sur cette approche n'ont pas atteint un stade supérieur à celles des SRI classiques. Ces constatations nous amènent à prendre en compte un autre type d'indexation qui est l'indexation sémantique. Indexation sémantique Les systèmes basés sur une indexation sémantique exploitent des ressources soit acquises en corpus, soit construites pour enrichir la représentation des documents et requêtes.

Exploitation des connaissances sémantiques issues de ressources

La plupart des SRI utilisent la base lexicale WORDNET pour l'enrichissement de l'indexation par

l'ajout des synsets [2] leur correspondant dans la base.
Les expérimentations de [94] basés sur le principe d'indexation fondée sur les synsets, mènent à une augmentation de 29% de la précision. Généralement, une amélioration des performances des SRI est perceptible lorsque les deux types d'indexation (classique et à base sémantique) sont combinés.

Exploitation des connaissances sémantiques acquises en corpus

Les SRI basés sur une indexation sémantique exploitant des connaissances acquises en corpus, utilisent des informations de cooccurrences pour dériver le sens des termes et aboutir à un appariement fondé sur le sens et non sur les unités lexicales. Ainsi, les travaux de [107] proposent d'améliorer les performances d'un SRI, en ajoutant des connaissances sémantiques qui sont des informations de co-occurrence d'unités linguistiques des documents avec les termes d'indexation retenus [3]. Le SRI testé présente une augmentation très significative des performances pour un faible rappel. Les expériences de [117], proposent une méthode dans laquelle documents et requêtes sont traités pour extraire les termes d'indexation et leurs sens. Le sens de chaque terme repose sur l'exploration de son contexte et sur le principe que les occurrences d'une unité lexicale utilisées dans le même sens partagent le même contexte. Les résultats montrent une amélioration de 4% de la précision moyenne.
L'efficacité de l'utilisation de connaissances sémantiques en RI dépend d'un nombre de facteurs. L'un d'entre eux est leur mode d'acquisition (utilisant des connaissances issues de ressources généralistes construites, ou extraites à l'aide de méthodes numériques ou symboliques).
S'interroger sur la manière la plus performante d'intégrer ces connaissances sémantiques au sein des systèmes est aussi un point principal : en extension de requêtes ou au coeur du système. Le second choix nécessite d'adapter les modèles de RI pour prendre en considération des relations de dépendances entre termes.

3.4 Conclusion

Ce chapitre dresse un bilan de l'impact de différentes informations linguistiques issues par des techniques du TAL sur les systèmes de recherche d'information et sur leurs performances. A travers ce tour d'horizon sur les contributions possibles des techniques du TAL à la RI, nous avons constaté que ces techniques permettent d'acquérir des informations plus riche que les unités lexicales simples. Ces informations permettent une meilleure représentation du contenu textuel et du besoin de l'utilisateur.

Une première remarque que l'on peut faire est que les potentialités du TAL sont loin d'être exploitées . Sur le niveau syntaxique, seuls sont pris en considération les termes complexes. Sur le plan sémantique, rares sont les expérimentations qui s'intéressent à la notion de thème, qui cherchent à identifier le thème des documents et à établir une correspondance avec la thématique des informations recherchées par l'utilisateur. Cependant, les différentes expériences effectuées ne prennent généralement en considération qu'un seul niveau de langue (morphologique, syntaxique ou sémantique). Il serait intére-ssant de chercher à coupler ces connaissances pour évaluer si une caractérisation très riche des documents et requêtes a un apport en RI.

Il reste toutefois que les informations linguistiques exploitables est très dépendante des modèles de RI dans lesquels elles vont être intégrées. La plupart des SRI se base sur des modèles de représentation des documents et requêtes qui se repose sur des ensembles d'unités lexicales indépendants, ce qui n'est pas adaptés à des connaissances qui cherchent à établir des relations entre termes. Ceci conduira à la

[2] Ensemble de synonymes
[3] Plus détail est dans [107]

conception de nouveaux modèles de représentation capables d'exploiter la puissance des informations linguistiques.

CHAPITRE 4

La Langue Arabe : état de l'art

L'arabe (al ?arabiya en transcription traditionnelle) est la langue parlée à l'origine par les Arabes. C'est une langue sémitique (comme l'akkadien et l'hébreu). Au sein de cet ensemble, elle appartient au sous groupe du sémitique méridional. Du fait de l'expansion territoriale au Moyen Âge et par la diffusion du Coran, cette langue s'est répandue dans toute l'Afrique du nord et en Asie mineure.

Dire langue arabe, c'est donc parler d'un ensemble complexe dans lequel se déploient des variétés écrites et orales répondant à un spectre très diversifié d'usages sociaux, des plus savants aux plus populaires. Mais au delà de cette diversité, les sociétés arabes ont une conscience aiguë d'appartenir à une communauté linguistique homogène. Elles sont farouchement attachées à l'intégrité de leur langue, d'où l'importance de l'ASM qui constitue le terrain commun pour cette large population.

Par ses propriétés morphologiques et syntaxiques, le traitement automatique doit faire face à :

- la nature agglutinante de la langue : l'ensemble des morphèmes collés à l'unité lexicale[1] véhiculent plusieurs informations morphosyntaxiques.
- la richesse flexionnelle de l'arabe
- l'absence de voyellation de la majorité des textes arabes écrits : ce phénomène entraîne un nombre important d'ambiguités morphologiques. En arabe, chaque lettre doit prendre un signe de voyellation et de surcroît les voyelles finales sont porteuses de certains traits morpho-syntaxiques comme la déclinaison, le mode, le cas.

En outre des propriétes linguistiques, l'arabe recense un nombre de ressources linguistiques comprenant des lexiques monolingues et multilingues ainsi que des corpus de langue générale et des corpus de spécialité consacrés à une situation de communication ou à un domaine de la connaissance. L'arabe compte aussi un certain nombre d'outils linguistiques à savoir les analyseurs morphologiques ainsi que les racineurs basés essentiellement sur une procédure de désuffixation qui consiste à supprimer les suffixes qui différencient les flexions des unités lexicales (les formes conjuguées d'un verbe par exemple).

Dans ce chapitre, nous introduisons la langue arabe. La section (4.1) est consacrée à son statut géographique, à ses diverses variantes et celle qui sera l'objet de l'étude. Dans la section (4.2) nous présentons les caractéristiques linguistiques et la classification des unités lexicales de l'arabe. Finalement, dans les sections (4.4), (4.5) nous aborderons les ressources linguistiques de l'arabe ainsi que les outils pour son traitement.

[1] Selon Mel'cuk [93], une unité lexicale est une entité trilatérale composée de (i): un sens, (ii): une forme phonique/graphique et (iii): un ensemble de traits de combinatoire (le syntactique)

4.1 la langue Arabe et ses variantes

L'arabe est une langue parlée par plus de 200 millions de personnes. Elle est langue officielle d'au moins 22 pays. C'est aussi la langue de référence pour plus d'un milliard de musulmans. Comme son nom l'indique, la langue arabe est la langue parlée à l'origine par le peuple arabe. C'est une langue sémitique (comme l'hébreu, l'araméen et le syriaque). Au sein de cet ensemble, elle appartient au sous groupe du sémitique méridional.
Le développement de la langue arabe a été associé à la naissance et la diffusion de l'islam. L'arabe s'est imposée, depuis l'époque arabo-musulmane, comme langue religieuse mais plus encore comme langue de l'administration, de la culture et de la pensée, des dictionnaires, des traités des sciences et des techniques. Ce développement s'est accompagné d'une rapide et profonde évolution (en particulier dans la syntaxe et l'enrichissement lexical).
L'arabe peut être considérée comme un terme générique rassemblant plusieurs variétés :

- l'arabe classique : la langue du Coran, parlée au VIIe siècle ;
- l'arabe standard moderne (l'ASM) : une forme un peu différenciée de l'arabe classique, et qui constitue la langue écrite de tous les pays arabophones. L'ASM reste le langage de la presse, de la littérature et de la correspondance formelle, alors que l'arabe classique appartient au domaine religieux et est pratiqué par les membres du clergé;
- les dialectes arabes : malgré l'existence d'une langue commune, chaque pays a développé son propre dialecte. Issus de l'arabe classique, leurs systèmes grammaticaux respectifs affichent de nettes divergences avec celui de l'ASM. On peut regrouper ces dialectes en quatre grands groupes :

    1. les dialectes arabes, parlés dans la Péninsule Arabique : dialectes du Golfe, dialecte du najd, yéménite ;
    2. les dialectes maghrébins : algérien, marocain, tunisien, hassaniya de Mauritanie ;
    3. les dialectes proche-orientaux : égyptien, soudanais, syro-libano-palestinien, irakien (nord et sud) ;
    4. la langue maltaise est également considérée comme un dialecte arabe.

L'arabe est un ensemble complexe dans lequel s'étendent des variétés écrites et orales répondant à un spectre très varié d'usages sociaux. Mais au delà de cette variété, les sociétés arabes ont une conscience aiguë d'appartenir à une communauté linguistique homogène, d'où l'importance de l'ASM qui forme un terrain commun pour cette large population. L'ASM est la langue des médias officiels, de la communication écrite et de tout type de communication non spontanée. Elle se distingue des dialectes arabes par son système grammatical partagé avec l'arabe classique. L'ASM, quoique qu'elle soit considéré comme le symbole le plus puissant de l'unité arabe, possède des variations régionales. Nous reconnaîssons un texte marocain vis-à-vis d'un texte égyptien ou d'un texte provenant des pays du Golfe. Cette variation est due aux différences qui ont lieu dans la formation de nouveaux vocabulaires. Mais elle est aussi la conséquence de l'histoire coloniale différente des régions impliquées. Les pays du Maghreb, par exemple, ont une tendance naturelle à regarder des exemples français, et le texte est largement influencé par la langue française même au niveau de la syntaxe et de la stylistique. Nous trouvons, par exemple الوزير الأول (de : le premier ministre français) au lieu du terme fréquent رءيس الوزرَاء (le président des ministres). Dans les pays arabes sans un passé colonial français, l'anglais remplace le français en tant que langue fournissant les modèles syntaxiques et stylistiques.

4.2 Grammaire et caractéristiques de l'arabe

La grammaire traditionnelle se divise en deux branches :

1. La morphologie, "اَلصَرف", qui comprend :
   (a) Morphologie dérivationnelle, qui étudie la construction des unités lexicales et leur transformation selon le sens voulu. Ainsi, la dérivation morphologique est décrite sur une base morpho-sémantique : d'une même racine, se dérivent différentes unités lexicales selon des schèmes qui sont des adjonctions et des manipulations de la racine. La racine [KTB] épouse plusieurs schèmes selon qu'on veut exprimer un procès accompli (c1 a c2 a c3 a) [kataba] ou inaccompli (y a c1 c2 u c3 u) [yaktubu], un nom d'agent (c1 a : c2 i c3 u n) [ka:tibun], un nom de patient (m a c1 c2 u : c3 u n) [maktu:bun], etc.
   (b) Morphologie flexionnelle concerne le marquage casuel pour le nom et l'adjectif ou la conjugaison du verbe, appelé "الَاعرَاب".

2. La syntaxe, "النحو", qui étudie la formation correcte des phrases garantit la grammaticalité de la phrase en analysant :
   (a) La position des unités lexicales les unes par rapport aux autres, déterminant ainsi l'ordre des unités lexicales.
   (b) Le marquage casuel des unités lexicales de la phrase. Ainsi, la fonction syntaxique de l'unité lexicale est déterminée en s'appuyant sur la morphophonologie.

Pour la reconnaissance des unités lexicales dans les textes, nous sommes confrontés à l'ambiguïté provoquée surtout par la voyellation partielle, l'agglutination et l'ordre relativement libre des unités lexicales dans la phrase [43].

Par exemple l'unité lexicale *ferme*, est hors contexte, un substantif, un adjectif ou un verbe. Alors que l'unité lexicale arabe RaLaKa غَلَقَ est un verbe à la 3ème personne masculin singulier de l'accompli actif, par contre sa forme non voyellée غلق (dans l'exemple donné ne sont représentées que les consonnes RLK) admet quatre catégories grammaticales :

- Substantif masculin singulier (RaLKun : une fermeture),
- Verbe à la 3è personne masculin singulier de l'accompli actif (RaLaKa : il a fermé ou RaLLaKa il a fait fermé),
- Verbe à la 3è personne masculin singulier de l'accompli passif (RuLiKa : il a été fermé),
- Verbe à l'impératif 2è personne masculin singulier (RaLLiK: fais fermer).

Une autre difficulté de l'arabe est l'agglutination par laquelle les composantes de l'unité lexicale sont liées les unes aux autres. Nous décrivons ci-dessous les propriétés linguistiques de la langue arabe, à savoir la voyellation, la flexion et l'agglutination.

4.2.1 Voyellation

La langue arabe s'écrit et se lit de droite à gauche, son alphabet compte 28 consonnes adoptant différentes graphies selon leur position (au début, au milieu ou à la fin d'une unité lexicale).
Une unité lexicale arabe s'écrit avec des consonnes et des voyelles. Les voyelles sont ajoutées au-dessus ou au-dessous des lettres. Elles sont nécessaires à la lecture et à la compréhension correcte d'un texte et elles permettent de différencier des unités lexicales ayant la même représentation.
Pour mieux comprendre prenons l'exemple de كتب du tableau 4.2.1. Le dictionnaire nous renvoie les voyellations lexicales suivantes:

- كَتَبَ, Il a écrit
- كُتِبَ, Il a été écrit
- كُتُبَ,Livres

Unité lexicale	1 ère intérprétation		2 ème intérprétation		3 ème intérprétation	
كتب	كَتَبَ	Il a écrit	كُتِبَ	Il a été écrit	كُتُب	Livres
مدرسة	مَدرسَة	Ecole	مُدَرِسَة	Enseignante	مُدَرَسَة	Enseignée

Table 4.1 – *Ambiguité causée par l'absence de voyelles pour les unités lexicales* كتب *et* مدرسة

4.2.2 Flexion

Une langue flexionnelle est une langue dans laquelle les unités lexicales varient en nombre et en flexion (soit le nombre des noms, soit le temps verbal) suivant les rapports grammaticaux qu'ils entretiennent avec les autres unités lexicales. L'ensemble des formes différentes d'une même unité lexicale fléchie constitue son paradigme [10]. D'aprés cette définition, l'arabe se classe comme une langue à morphologie extrêmement riche :
Le système flexionnel affiche un marquage varié. Par exemple, l'arabe contient trois cas [2] : le nominatif (NOM), qui est le cas par défaut, l'accusatif (ACC) pour les compléments verbaux [3] et le génitif (GEN) pour le dépendant d'une préposition. Les morphes sont divisés dans la translittération par le symbole "+" :

nAma	AlAwlAd+**u**	نَامَ الَاولَادُ
(V)PASSE	(N)+NOM	
a dormi	les enfants	
"Les enfants	ont dormi"	

qAbAlA	mohamad	AlAwlAd+**a**	قَابَلَ مُحَمَد الَاولَادَ
(V)PASSE	(N)	(N)+ACC	
a rencontré	mohamad	les enfants	
"mohamad	a rencontré	les enfants"	

salama	mohamad	3ly	AlAwlAd+**i**	سَلَّمَ مُحَمَد عَلي الَاولَادِ
(V)PASSE	(N)	(PREP)	(N)+GEN	
a salué	mohamad	sur	les enfants	
"mohamad	a salué	les enfants"		

La définitude est effectuée par un morphème préfixé (ال) ou suffixé (اً), et non par des unités lexicales autonomes. La marque du défini est réalisée par des lettres tandis que la marque de l'indéfini est réalisée

[2] Le cas d'une unité lexicale indique sa fonction, c'est-à-dire sa relation avec les autres unités lexicales de la phrase

[3] Rappelons que la tradition grammaticale ne fait pas de distinctions entre les différents compléments d'objet du verbe vu qu'ils portent tous la même marque casuelle, même les compléments prédicatifs tel l'attribut et le coprédicat.

par un signe diacritique fusionné au signe de la voyelle courte, c'est le signe appelé tanwiin : الَاولَادَ vs اولَادًا

AlAwlAd+a	vs	AwlAd+a+**n**
DEF+(N)+NOM	vs	(N)+NOM+INDEF
les enfants	des enfants	

4.2.3 Agglutination

L'arabe montre une forte tendance à l'agglutination : l'ensemble des morphèmes collés les unes aux autres et constituant une unité lexicale véhiculent plusieurs informations morpho-syntaxiques. Ces unités lexicales sont souvent traduisibles par l'équivalent d'une phrase en français. La structure d'une unité lexicale arabe est donc décomposable en cinq éléments : proclitique, préfixe, base, suffixe et enclitique. La base est une combinaison de lettres radicales (le plus souvent trois) et d'un schème. La base - avec préfixe et suffixe - forme le noyau lexical, éventuellement entouré d'extensions [36]. Comme le montre l'exemple suivant : وَلِيَضرِبُهَا. Les éléments clitiques sont séparés par le symbole "+" :

wa +	li +	ya + Dribu	+ haA
(COORD) +	(CONJONCTION) +	(V)SUBJONCTIF +	(PRO)
et	pour	frappent	elle
"et	pour	la	frapper"

Cet exemple révèle la complexité morphologique de l'arabe. Il s'agit du verbe يَضرِبُ employé au présent du subjonctif, 3ème personne du masculin pluriel, la base verbale est / ضَرَبَ / et la racine / ض ر ب/. Le pronom sujet n'est pas réalisé. En position proclitique, on utilise la conjonction de coordination "wa" و et la conjonction "li" ل. En position enclitique, on utilise le pronom complément d'objet 3ème personne du féminin singulier "haA" هَا "elle".

4.2.4 Pro-drop (= à sujet pronominal vide)

L'ASM néglige systématiquement la réalisation morphologique du pronom sujet. Cependant, le verbe s'accorde en personne, en genre et en nombre avec le pronom omis, comme l'affiche l'exemple suivant: / هُمُ / اكَلُ vs / هُنَّ / اكَلنَّ. Le pronom correspondant est mis entre / / :

Akalu /homo/	vs	Akalnna /honna/
(V)PASSE.3.MASC.PL	vs	(V)PASSE.3.FEM.PL
ont mangé /ils/	vs	ont mangé /elles/
"Ils ont mangé"	vs	"Elles ont mangé"

4.3 Les parties de discours en arabe

Les unités lexicales qui composent le discours sont regroupés par catégories selon les caractéristiques qu'ils ont en commun. Ces différentes catégories s'appellent les parties du discours [4].

[4] http://www.ebsi.umontreal.ca

Cette section donne une classification des unités lexicales de la langue arabe. Dans un premier temps, nous présentons la classification traditionnelle des unités lexicales (sous-section 4.3.1), ensuite des tentatives de classification plus récentes (sous-section 4.3.2).

4.3.1 Les parties de discours classiques

La grammaire traditionnelle compte trois classes [13]: le nom, le verbe et la lettre. La catégorie nominale rassemble toutes les unités lexicales n'ayant pas de sens lié au temps et regroupe les catégories du substantif et de l'adjectif. La catégorie verbale comprend toutes les unités lexicales référant à un état ou à une action au passé, au présent ou au futur. La classe lettre, quant à elle, se répartit d'une part, en lettres de l'alphabet, littéralement les lettres de construction, "حُروف المبَاني", qui s'unissent pour former des unités lexicales, et d'autre part, en lettres de signification, "حُروف المعَاني", dont le sens n'est complet que si elles sont utilisées avec un nom ou un verbe. La grammaire traditionnelle recense presque quatre-vingts particules, dont l'identification de la classe syntaxique exige d'étudier séparément les propriétés distributionnelles de chaque lettre.
Sur critères morphologiques, la classe du nom se répartit en deux groupes [13]:

1. Noms variables comprenant les deux propriétés suivantes :
   (a) Ils acceptent les changements morphologiques et comprennent des variantes numérales (singulier, duel et pluriel). Cette sous-catégorie contient les déverbaux (المصدَر tel le nom d'agent, le nom de patient, le nom de résultat, et le nom d'instrument).
   (b) Ils ont des formes dérivées adjectivales et diminutives. Ils se répartissent en noms dérivés du paradigme verbal et noms non dérivés. Ces derniers se subdivisent aussi, sur une base de distinction conceptuelle, en noms abstraits, relatifs à l'espace mental, et noms concrets, relatifs à l'espace physique.
2. Noms invariables regroupant des lexèmes tels que le pronom, le démonstratif, l'interrogatif, le relatif et certains numéraux. Ces noms sont dits invariables car la marque casuelle n'est pas identifiée phonologiquement. Cependant, ces lexèmes exercent les fonctions d'un nom.

4.3.2 Classification récentes des unités lexicales de l'arabe

A notre connaissance, les études qui ont cherché à classifier des unités lexicales en arabe selon les parties de discours sont très peu nombreuses [39]. Les démarches récentes de classification des unités lexicales se répartissent en deux approches. Certaines consistent en une classification identifiée pour les langues indo-européennes sans prendre en considération l'existence possible d'une classe n'existant pas dans ces langues, ou bien l'inverse. D'autres ont conservé la classification traditionnelle arabe tout en lui suggérant des raffinements.
Nous présentons une classification assez récente réalisée dans le cadre du développement d'un étiqueteur morpho-syntaxique [76] qui a servi de référence pour d'autres recherches comme [35]. [76] présente un étiquetage basé sur la classification traditionnelle et raffinée par les subdivisions proposées par [67]. Selon cette classification, les unités lexicales se répartissent en cinq classes : nom, verbe, particule, résiduel et ponctuation. Certaines sont raffinées en sous classes illustrées sur la figure suivante :

Nous avons présenté une description succinte de la grammaire arabe et avons décrit ses propriétés linguistiques :

- Une langue voyellée qui avec l'absence de voyellation entraîne une ambiguité à différencier des unités lexicales ayant la même représentation.

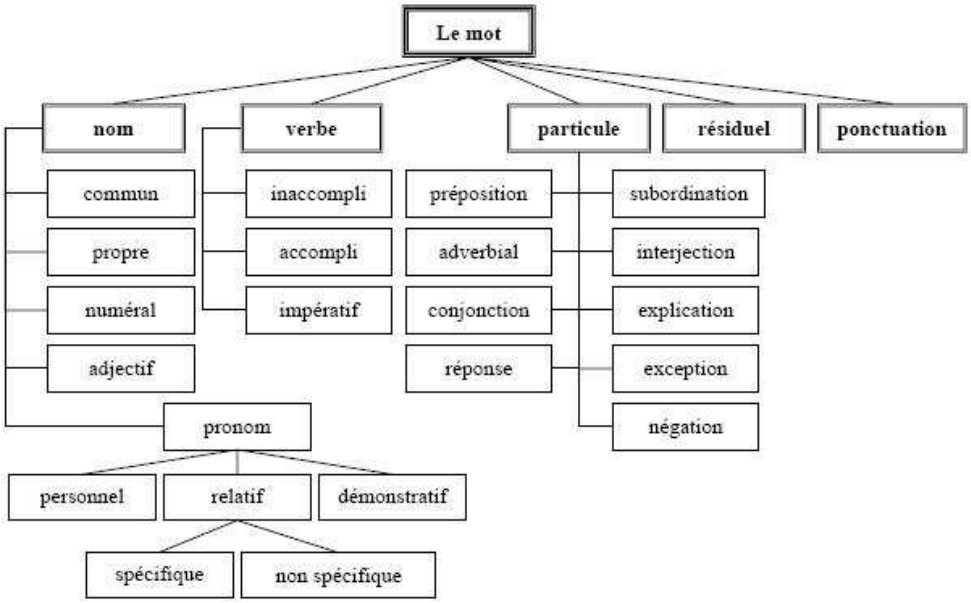

Figure 4.1 – Classification des unités lexicales proposée par [76]

- Une langue flexionnelle dans laquelle les unités lexicales varient en nombre et en flexion (soit le nombre des noms, soit le temps verbal), suivant les rapports grammaticaux qu'ils entretiennent avec les autres unités lexicales.
- une langue agglutinante où l'ensemble des morphèmes collées les unes aux autres et constituant une unité lexicale véhiculent plusieurs informations morpho-syntaxiques. Ces unités lexicales sont souvent traduisibles par l'équivalent d'une phrase en français.
- Une langue pro-drop où elle néglige systématiquement la réalisation morphologique du pronom sujet.

Nous avons ensuite présenté la classification traditionnelle tripartite -verbe, nom et particule-, puis nous avons décrit une classification structurale récente des unités lexicales en arabe, ainsi elles se répartissent en cinq classes : nom, verbe, particule, résiduel et ponctuation.

4.4 Ressources linguistiques : état des lieux

Les ressources linguistiques (RL) jouent un rôle essentiel dans les applications de la technologie des langues. Ainsi, d'une part les RL alimentent les différents processus des systèmes de TAL, d'autre part, elles sont de plus en plus exploitées pour accompagner le travail de modélisation linguistique par des méthodes statistiques [112].
Les RL à grande échelle connaissent une diffusion croissante, notamment grâce à des structures comme

le LDC [5] (Linguistic Data Consortium) aux Etats-Unis et l'ELRA [6] (European Language Resources Association) en Europe.
Nous donnons un aperçu des ressources linguistiques existantes pour l'arabe. Nous nous limitons à celles utiles pour l'analyse automatique des corpus textuels. Nous abordons successivement sont donc les lexiques (section 6.1.4), les corpus de textes monolingues bruts et annotés (section 4.4.2).

4.4.1 Lexiques

Un lexique se constitue d'une liste d'entrées lexicales auxquelles peuvent être associées des informations linguistiques relevant la morphologie, la syntaxe, ou la sémantique ainsi que sa fréquence d'usage, des exemples d'emploi, etc.
Toutes ces informations peuvent être regroupées en deux groupes distincts, les informations intra-lexicales et inter-lexicales. Les informations intra-lexicales (constituant la micro-structure du lexique) tandis que, les informations inter-lexicales (constituant la macro-structure du lexique) sont celles qui lient les unités lexicales entre eux dans le lexique. Nous distinguons différents types de liens :

- les liens morphologiques permettent de lier l'unité lexicale à sa forme de base. Ils regroupent les informations flexionnelles et dérivationnelles (lien entre une forme fléchie et son lemme).
- les liens sémantiques lient l'entrée lexicale avec ses informations sémantiques.

Nous présentons quelques uns des lexiques electronique de l'arabe, en abordant d'une part les lexiques monolingues et d'autre part les lexiques multilingues.

Lexique monolingue

DIINAR.1 : DIctionnaire INformatisé de l'ARabe DIINAR.1 (DIctionnaire INformatisé de l'ARabe - version1) [37] est l'une des bases lexicales les plus importantes dans le domaine du TALN traitant la langue arabe. Elle a été conçue et réalisée à l'ENSSIB par [66] [56] [132]. Elle est en instance de diffusion via ELRA/ELDA. Elle comprend un nombre total de 119 693 lemmes, entièrement voyellés, répartis comme suit : 29 534 noms, 19 457 verbes, 70 702 déverbaux dont le bilan total est présenté table 4.2.

Chaque entrée de DIINAR.1 est associée à des informations morpho-syntaxiques.

	Entrée lexicale
Verbes	19 457
Noms	29 534
Déverbaux	70 702
Total	119 693

Table 4.2 – *Composition de la base lexicale DIINAR.1*

1. Déverbaux
   Un substantif peut dériver d'un verbe par dérivation affixale (suffixale), c'est un processus qui nous permet respectivement d'obtenir à partir du verbe « لعب » « jouer », par exemple, le substantif : « لعبة », « jouet ».

[5] http://www.ldc.upenn.edu/
[6] http://www.elra.org/

2. Données nominales
   Les noms communs représentent 29 534 entrées et suivent 11 modèles de déclinaisons. Chaque nom peut être associé à des suffixes susceptibles de former de nouvelles entrées lexicales de la base, comme l'ajout de "ي " ou le "ية"de relation".

Lexiques multilingues

Il existe deux types de lexiques multilingues, ceux qui s'intéressent à la mise en correspondance de deux langues, souvent dans un objectif précis (lexiques bilingues), et ceux dont l'objectif plus ambitieux est de développer un mécanisme générique pouvant permettre la mise en parallèle d'informations lexicales pour un nombre a priori arbitraire de langues.

DIINAR-MBC

DIINAR-MBC (DIctionnaire INformatisé de l'ARabe, Multilingue et Basé sur Corpus) [37] est l'extension multilingue de DIINAR1 vers anglais et français. Il conserve les informations morphologiques et syntaxiques antérieurement développées et stockées dans DIINAR.1 et ajoute les définitions et les équivalents en français et en anglais.

DixAF

Le DixAF (Dictionnaire bilingue français arabe, arabe français), copropriété CNRS/ENS lettres et sciences humaines développé par Fathi Debili, est constitué de près de 125 000 liens binaires établis entre 43 800 entrées françaises environ et près de 35 000 entrées arabes. Il est disponible sous format Access. Les mots arabes sont majoritairement voyellés. Un certain nombre de catégories grammaticales sont également indiquées (noms, adjectifs, verbes, adverbes, pronoms, prépositions, etc.). Ce dictionnaire peut être utilisé pour des applications d'indexation bilingue français-arabe, arabe-français, de traduction ou d'interrogation bilingue.

AGROVOC

AGROVOC est un vocabulaire plurilingue de spécialité (regroupant environ 36 000 termes uniques) conçu pour couvrir l'ensemble des domaines ayant trait à l'agriculture, aux forêts, à la pêche, à l'alimentation et à d'autres domaines apparentés (comme l'environnement). Il se compose de termes d'indexation qui consistent en un ou plusieurs mot(s) se référant toujours à un seul et unique concept. Pour chaque terme, un bloc-mot s'affiche et indique la relation hiérarchique et non hiérarchique à d'autres termes: BT (broader term, terme plus large), NT (narrower term, terme plus étroit), RT (related term, terme connexe), UF (non descripteur). L'objectif principal d'AGROVOC est de normaliser l'indexation pour simplifier la recherche d'informations et la rendre plus efficace afin que les utilisateurs accèdent aux ressources dont ils ont besoin. AGROVOC est disponible en 16 langues: les cinq langues officielles de la FAO (qui sont anglais, français, espagnol, chinois et arabe), coréen, portugais, japonais, thaï, slovaque, allemand, hongrois, polonais, le farsi, l'hindi et italien.

AGROVOC en arabe contient actuellement plus de 24 699 descripteurs et plus de 1247 non descripteurs (synonymes). Chaque descripteur a son équivalent dans d'autres langues. Les non descripteurs sont des termes destinés à aider l'utilisateur à rechercher le(s) descripteur(s) approprié(s). Ci-dessous un exemple de la structure d'AGROVOC en arabe du terme تلوث " pollution " et تلوث الهوَاء " pollution

atmosphérique ". Le fait de savoir qu'un terme plus large pour définir تلوث الهوَاء " pollution atmosphérique " est تلوث " pollution ", et que des termes associés à ce mot sont جو "atmosphère" et خَاز الفضلَات " effluent gazeux " définit la gamme de l'information représentée par ces termes.

UNBIS

La base multilingue UNBIS est élaboré, au sein du Département de l'information, par la Bibliothèque Dag Hammarskjöld. Il contient la terminologie utile pour l'analyse thématique des documents et autres publications relatifs aux programmes et activités de l'Organisation des Nations Unies. Il est utilisé comme le fichier d'autorité du Système d'information bibliographique de l'ONU (UNBIS) et a été incorporé en tant que liste de sujets du Système de diffusion électronique des documents de l'ONU (SEDOC). UNBIS est un outil multidisciplinaire, abordant l'ensemble des domaines d'action de l'Organisation. Les termes intégrés ont pour objectif de refléter de manière précise, claire et concise, et à un niveau de spécificité adéquat, tous les domaines d'importance et d'intérêt des Nations Unies. Il est disponible dans toutes les langues officielles des Nations Unies (anglais, arabe, chinois, espagnol, français et russe). Il propose les relations conceptuelles génériques usuelles : l'équivalence du terme, ses génériques et spécifiques (TG/TS), la relation de synonymie (EM/EP), les termes associés (TA) ou encore une note d'usage (ou d'application).

UNBIS en arabe contient actuellement 7002 descripteurs et 2183 non-descripteurs (synonymes). Chaque descripteur a son équivalent dans d'autres langues.

4.4.2 Corpus

Le corpus se définit de fait comme l'objet concret auquel s'applique le traitement, qu'il s'agisse d'une étude qualitative ou quantitative. Le corpus est défini par [84] comme « l'ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique ». Mais les données ont un nom trompeur : elles ne s'imposent pas, elles sont construites. Certes, il y a un existant, directement sous forme de textes électroniques par exemple, et donc l'analyste n'a pas une totale liberté d'inventer ses données, il part d'une réalité, mais il reste des décisions du type : faut-il considérer tout ce qui est disponible ou en extraire un sous-ensemble plus significatif et équilibré ; comment éventuellement l'adapter au traitement envisagé. Ainsi, selon [104] le corpus doit vérifier trois types de conditions : des conditions de signifiance, des conditions d'acceptabilité, et des conditions d'exploitabilité.

- Conditions de signifiance : un corpus est constitué en vue d'une étude déterminée, portant sur un objet particulier, une réalité telle qu'elle est perçue sous un certain angle de vue. Les documents retenus doivent être adéquats comme source d'information pour correspondre à l'objectif qui suscite l'analyse.
- Conditions d'acceptabilité : le corpus doit apporter une représentation fidèle, sans être parasité par des contraintes externes. Il doit avoir une ampleur et un niveau de détail adaptés au degré de finesse et à la richesse attendue en résultat de l'analyse.
- Conditions d'exploitabilité : les textes qui forment le corpus doivent être commensurables. Le corpus doit apporter suffisamment d'éléments pour pouvoir repérer des comportements significatifs (au sens statistique du terme).

Dans notre cas, nous distinguons deux grandes catégories de corpus : les corpus de spécialités tentent de refléter l'usage de la langue dans un domaine particulier (corpus techniques, médicaux), tandis que les corpus généralistes s'intéressent à l'ensemble d'une langue et rassemblent souvent des textes plus diversifiés, représentatifs de sa diversité.

Corpus général

Le corpus de langue générale est consacré à une langue naturelle. Il tend à représenter la diversité des usages de la langue choisie. A ce titre, il est constitué d'un ensemble de données dont les conditions de production et de réception sont représentatives d'une grande variété de situations de communication (orale : monologue, interview , écrite : lettre, roman...), et de types textuels (exposé scientifique, fiction narrative, reportage...). Il permet la constitution de sous corpus en registre [7] pour des analyses contrastives par exemple. En outre, le corpus de langue générale est souvent ouvert, c'est-à-dire que son contenu est sans cesse augmenté de nouvelles données, ce qui autorise à terme des analyses diachroniques (néologismes, emplois morphologiques privilégiés). Enfin, le corpus de langue générale est de grande taille, il dépasse aujourd'hui plusieurs millions d'occurrences. Al-Hayat Le corpus Al-Hayat est distribué par l'organisme ELRA, il a été développé dans le cadre d'un projet de recherche de l'Université d'Essex, en collaboration avec Open University. Ce corpus est constitué d'articles extraits du journal Al-Hayat, qui ont été utilisés dans les campagnes TREC.
Les données sont réparties dans sept rubriques, suivant les critères de répartition des sujets du journal Al-Hayat : rubrique Générale, rubrique Automobile, rubrique Informatique, rubrique Actualités, rubrique Economie, rubrique Sciences, et rubrique Sport.
Le balisage, les nombres, les caractères spéciaux et la ponctuation ont été supprimés. La taille totale du fichier est de 268 Mo. Il contient 18 639 264 unités lexicales, 42 591 articles. An-Nahar Le corpus de textes du quotidien libanais An-Nahar distribué par ELRA, est consitué d'articles en arabe standard de 1995 à 2000, stockés sous la forme de fichiers HTML sur CD-ROM. Chaque année contient 45 000 articles et 24 millions de mots. Chaque article contient des informations telles que le titre, le nom du quotidien, la date, le pays, le type, la page, etc. NEMLAR : Network for Euro-Mediterranean LAnguage Resources Ce corpus a été produit dans le cadre du projet NEMLAR [8]. Le corpus écrit NEMLAR est constitué de 500 000 unités lexicales regroupés en 13 catégories différentes, visant à obtenir un corpus bien équilibré qui offre une représentation de la variété de traits syntaxiques, sémantiques et pragmatiques de la langue arabe moderne. Les différentes catégories sont illustrées dans la table 4.3.

Le corpus est fourni sous la forme de 4 versions différentes:

- Texte brut
- Texte entièrement voyellé
- Texte comprenant une analyse lexicale de l'arabe
- Texte enrichi linguistiquement avec les parties du discours

Agence France Presse L'Agence France Presse (http://www.afp.com/arabic/home/) est l'un des plus gros diffuseurs européen de dépêches en langues Arabe. Le corpus est constitué de 383 872 documents. Il a été encodé en utilisant le SGML et a été transcodé à Unicode (UTF-8). Le corpus inclut des articles journalistiques du 13 mai 1994 au 20 décembre 2000 avec approximativement 76 millions d'unité lexicale. Les données sont réparties dans six rubriques, suivant les critères de répartition des sujets du journal Agence France Presse : rubrique Générale, rubrique Informatique, rubrique Actualités, rubrique Economie, rubrique Sciences, et rubrique Sport. Chaque article contient des informations telles que le titre, la date, le pays, la page, etc. Corpus arborés Un corpus arboré est un corpus annoté par des informations de nature interprétative [128]. Les différents type d'annotation dont parle J.Véronis sont : l'annotation grammaticale, sémantique, multilingue ainsi que l'annotation phonétique. Il existe deux types d'annotation grammaticale. Le premier consiste à effectuer un étiquetage des catégories grammaticales et des informations morpho-syntaxiques associées. Le deuxième est un marquage de structures syntaxiques,

[7] registre est employé au sens de Biber (1995), pour une conception élargie des genres
[8] http://www.nemlar.org

Domaine du corpus	Nombre d'unité lexicale
Articles politiques (63)	48 000
Débats politiques (22)	30 000
Textes islamiques (12)	29 000
Expressions (6)	8 500
Broadcast news (4)	5 500
Affaires (10)	20 000
Oeuvres litéraires (24)	30 000
Articles journalistiques	100 000
Interviews (18)	56 000
Articles scientifiques (51)	50 000
Articles sportifs (98)	50 000
Explication d'entrées dictionnaire (12)	52 000
Textes de droit	21 000
	Total 500 000 mots

Table 4.3 – *Composition du corpus NEMLAR*

complètes ou partielles. En ce qui concerne l'annotation sémantique, deux catégories de celle-ci peuvent être distinguées : l'étiquetage du sens des unités lexicales et l'étiquetage de phénomènes discursifs. L'annotation multilingue, quant à elle, consiste à aligner des textes avec leur traduction. Cet alignement peut s'effectuer au niveau de la phrase ou de l'unité lexicale. Enfin, en ce qui concerne l'annotation phonétique, elle est également de deux types. Le premier consiste en une transcription de l'oral à l'aide de l'écrit, le deuxième consiste en un marquage de phénomènes prosodiques. Les corpus arborés disponibles en arabe sont :

1. Le corpus arboré pour la langue arabe élaboré à l'université de Pennsylvanie [9]. Ce corpus est disponible pour les membres du "Linguistic Data Consortium" en ligne et par CD-Rom. Le Penn Treebank pour l'arabe est en fait un corpus dont l'annotation a été produite de façon semi-automatisée. L'étiquetage et le parenthésage sont automatiquement produits puis corrigés par des annotateurs humains. Une grammaire est ainsi développée dans une perspective d'analyse automatique comprenant des relations prédicat/argument sous forme d'une grammaire syntagmatique. Nous illustrons cette annotation sur la phrase exemple: "AnhAr Alswk" "the market crumbled" Dans cet

```
( (S
(VP
(VBD AnhAr) )
(NP-SBJ
(DT Al)
(NN swk) )
(. .) ))
```

exemple, tout d'abord "Al" est étiqueté par DT (determiner), "swk" par NN (noun), "AnhAr" par

[9] Penn Arabic Treebank : MAAMOURI Mohamed ; BIES Ann ; JIN Hubert ; BUCKWALTER Tim (2003). "Arabic Treebank: Part 1 v 2.0.", LDC catalogue numéro LDC2003T06, ISBN 1-58563-261-9. Adresse électronique du site : http://www.ircs.upenn.edu/arabic/

VBD (past tense) et "." par .(sentence final punctuation). DT et NN sont ensuite annotés par NP-SBJ (un nom qui joue le rôle du sujet), VBD par VP (verbal phrase). NP-SBJ, VP et . sont enfin annotés par S(simple declarative clause).

2. Le corpus arboré élaboré à l'université de Prague, PADT (de l'anglais Prague Arabic Dependency Treebank) basé sur une syntaxe de dépendance avec une recherche visant à transformer d'une manière automatique les arbres syntagmatiques du Penn Treebank en arbres de dépendances.

Corpus spécialisé

Le corpus spécialisé est généralement consacré à une situation de communication ou à un domaine de la connaissance [60]. L'hypothèse de recherche porte sur la langue de spécialité en usage dans le domaine, dont il doit être représentatif. Les critères de sélection textuelle sont donc basés sur une configuration domaine-genre (ex : articles de recherche en biochimie) [103]. Il tend également à être homogène, c'est à dire à diversifier les sources de données ; ce qui importe n'est pas tant le nombre de mots que la quantité et la variété de documents rassemblés. Enfin, il n'y a pas de consensus sur la taille souhaitable pour un corpus spécialisé, mais il est réduit comparé au corpus de langue générale, oscillant entre cinq cent mille [1] et un million d'unité lexicale [101]. Ainsi, on retient, pour le corpus spécialisé, tout regroupement de données langagières créé à des fins spécifiques et représentatives d'une situation de communication ou d'un domaine de pratique. OnuAR Le corpus « OnuAR » est constitué des publications de l'ONU en arabe (traités, conventions, constitutions, législations, etc.) ; il regroupe au total plus d'un million d'unités lexicales classées par domaine et par type de publication. Il donne un aperçu représentatif de la terminologie utilisée en particulier dans les institutions rattachées aux Nations Unies (OMPI, FAO, etc.). Science Le corpus « Science » (http://www.comp.leeds.ac.uk/eric/latifa/research.htm) developpé à l'université de Leeds, est constitué des publications de science et technologie en arabe; il regroupe au total 101 214 d'unité lexicale annoté en XML. Il donne un aperçu représentatif de la terminologie utilisée en particulier dans les magazines scientifiques.

De nombreuses ressources linguistiques ont été construites ces dernières années, fournissant les conditions d'un développement accéléré des activités de recherches dans les domaines concernés. C'est particulièrement le cas pour la langue anglaise, qui concentre l'attention d'une part très importante de la communauté scientifique. En ce qui concerne l'arabe, nous avons donné un aperçu des ressources linguistiques. Nous nous sommes limités aux ressources concernant l'analyse automatique des corpus textuels. Les sujets abordés sont les lexiques monolingues et multilingues, les corpus de langue général, les corpus de langue de spécialité et les corpus annotés.

4.5 Outils de traitement automatique de la langue arabe

Les outils de traitement automatique de la langue arabe sont l'ensemble des recherches et développements visant à modéliser et reproduire, à l'aide de machines, la capacité humaine à produire et à comprendre des énoncés linguistiques dans des buts de communication. Notre objectif dans cette section est de recenser les principaux outils de TAL en langue arabe. Les sujets abordés sont donc les analyseurs morphologiques (section 4.5.1), les concordanciers (section 4.5.2) et les racineurs (section 4.5.3).

4.5.1 Analyseurs morphologiques

Buckwalter

L'analyseur de buckwalter developpé par LDC (Linguistic Data Consortium) permet de segmenter chaque unité lexicale en une séquence du type préfixe-stem-suffixe. Le préfixe est une combinaison de 0-4 caractères, le suffixe est composé de 0 à 6 caractères et le stem comprend un à plusieurs caractères. Il est constitué principalement de trois lexiques : préfixes (548 entrées), suffixes (906 entrées), et stem (78839 entrées). Les lexiques sont complétés par trois tables de compatibilité utilisés pour couvrir toutes les possibilités de combinaisons préfixe-stem (2435 entrées), suffixe-stem (1612 entrées) et préfixe-suffixe (1138 entrées). Ainsi, l'analyseur donne en sortie l'unité lexicale, sa catégorie morphosyntaxique et sa traduction anglaise.

Aramorph

L'analyseur morphologique Aramorph [2] segmente les unités lexicales, repère les différents composants et atteste son appartenance à la langue. Pour cela, le système est assisté par le lexique DINAAR.1 pour éviter les analyses théoriquement possibles et inexistantes dans la langue. Par la suite, l'analyseur donne une liste des traits associés à l'unité lexicale en entrée. Il offre deux types d'options. Le premier vise les traits morphosyntaxiques, le second concerne l'analyse des préfixes et suffixes.
En plus des étiquettes morphosyntaxiques, il donne en sortie d'autres informations comme la base, l'unité lexicale minimale vocalisé ou non ainsi que la forme compléte supposée vocalisée ou non.
Analyser les préfixes revient à decrire ses découpages possibles et d'examiner les compositions des clitiques. Ceci amène le système à faire la distinction entre les clitiques ayant la même forme mais appartenant à des catégories syntaxiques différentes.
Exemple : "وَالتلوث" " et la pollution" l'analyseur découpera le proclitique و et dira que و est celui de la liaison.

Xerox

L'analyseur morphologique de Xerox [12] est basé sur l'approche de transducteur à états finis. La segmentation de la phrase en unités lexicales est réalisé par un transducteur à états finis. Ce transducteur découpe la chaîne d'entrée en une séquence d'unités lexicales qui peuvent correspondre à une forme fléchie, une marque de ponctuation, etc. La deuxième étape est l'analyse morphologique des unités lexicales produites par la segmentation de la phrase. Cette étape est aussi réalisée par un transducteur qui relie la forme fléchie à la forme lexicale (et vice-versa). La forme lexicale est une séquence comprenant la représentation canonique de l'unité lexicale (le lemme), un ensemble d'étiquettes représentant le comportement morphologique de l'unité lexicale, et sa catégorie syntaxique.

ASVM

L'analyseur de Diab (ASVM) est un logiciel libre, développé en Perl par l'équipe de Mona Diab [35] à la Leland Stanford Junior University en 2004. Il s'agit d'une adaptation à l'arabe du système anglais YamCha basé sur les Support Vector Machines. Les données probabilistes ont été acquises pendant une phase d'entrainement sur le corpus annoté Arabic TreeBank. Ci-dessous la description des fichiers d'entrée et de sortie de l'analyseur.

ENTREE
Le texte à analyser doit être encodé en Buckwalter, qui est une table de correspondance biunivoque entre les caractères arabes et l'ASCII. Voici un exemple:

ولم يحتسب الحكم المجري سَاندور بول ركلة جزَاء صحيحة اثر عرقلة دَاخل المنطقة من قبل اليسَاندرو

« wlm yHtsb AlHkm Almjry sAndwr bwl rklp jzA' SHyHp Avr Erqlp dAxl AlmnTqp mn qbl AlysAndrw. »
SORTIE
Dans le fichier de sortie, chaque unité lexicale étant suivi d'un slash et de sa catégorie. Les clitiques s'écrivant attachés à leur hôte comme les conjonctions de coordination (fa-) et (wa-), la préposition (bi-) etc. sont étiquetés indépendemment.
« w/CC lm/RP yHtsb/VBP Al/DT Hkm/NN Al/DT mjry/JJ sAndwr/NNP bwl/NNP rklp/NN jzA'/NN SHyHp/JJ Avr/IN Erqlp/NN dAxl/IN Al/DT mnTqp/NN mn/IN qbl/NN Al/DT ysAndrw/NNP ./PUNC »

4.5.2 Les concordanciers

La réalisation manuelle des concordances écrites était un travail de grande envergure envisageable uniquement pour le oeuvres perennes. Le traitement automatique a facilité la tâche et a étendu leurs champs d'application à de nombreuses disciplines scientifiques. Dans le cas de la langue arabe, l'aboutissement d'un concordancier électronique necessite un travail prélable faisant appel à des ressources lexicales et des outils d'étiquetage morpho-syntaxique. L'approche classique de réalisation des concordanciers, basée sur une reconnaisance graphique des items dans les textes KWIC (Key Word In Context), est inefficace dans le traitement de l'arabe, dont l'écriture est non-vocalisée, et dont les structures de l'unité lexicale peuvent être décrites comme agglutinantes et hautement flexionnelles. Ainsi, L'outil AraConc développé pour l'arabe par [2] a pour objectif de donner les contextes et fréquences, et permettre l'exploration du corpus selon les traits proposés par l'analyse morphologique et selon les informations graphiques qui se trouvent dans le texte. La concordance finale arabe tourne autours du trio : unité lexicale, position et analyse morphologique. L'outil prend en entrée un texte ou un ensemble de textes. Il permet :

- La construction de listes de fréquences d'items, de racines ou tout autre trait de l'analyse morpho-syntaxique, par ordre alphabétique ou par ordre fréquentiel.
- La construction d'une concordance, La consultation de la concordance peut se faire par item, par la racine, par la base ou par analyse morpho-syntaxique.

4.5.3 Racineurs

Les racineurs se veulent d'abord un outil utile au TAL, ce type d'analyse « simpliste », traite de façon identique affixes flexionnels et dérivationnels. Les algorithmes de racinisation en arabe les plus connus sont ceux de [85] et [77]. Ci-dessous une description succinte de ces racineurs.

Racineur de larkey

L'approche de [85] est une analyse morphologique assouplie. Elle consiste à essayer de déceler les préfixes et les suffixes ajoutés à l'unité lexicale : par exemple le duel (ان) dans (معلمَان, deux professeurs), le pluriel des noms masculins (ون ، ين) dans (معلمون, des professeurs) et féminins (ات) dans (مسلمَات,

musulmanes) ; la forme possessive (نَا ، كم، هم) dans (كتَابهم, ses livres) et les préfixes dans les articles définis (ال ، وَال ، بَال ، كَال، فَال). L'ensemble des prefixes et suffixes à supprimer sont présentés dans le tableau 4.5.3

Préfixes			Suffixes	
1- Caractère	2- Caractère	3- Caractère	1- Caractère	2- Caractère
ت	ال	وَال	ة	اة
ل	بت	فَال	ه	ان
ا	تّ	بَال	ي	تَا
ي	يت		ا	تك
م	لت			تي
	مت			ته
	وت			تم
	ست			هم
	نت			هن
	بم			هَا
	لم			كم
	وم			وَا
	كم			ون
	فم			وه
	لّ			ية
	وي			نَا
	لي			ين
	سي			يه
	في			
	وَا			
	فَا			
	لَا			
	بَا			

Table 4.4 – *Liste des préfixes et suffixes*

Racineur de Khoja

Le racineur de Shereen khoja [77] developpé au sein de l'université de Lancaster, a été utilisé dans le cadre d'un système de recherche d'information développé à l'Université du Massachusetts [85]. L'approche de Khoja [77] consiste à détecter la racine d'une unité lexicale, d'une part, il faut connaître le

schème par lequel elle a été dérivé et supprimer les éléments flexionnels (préfixes et suffixes) qui ont été ajoutés, d'autre part comparer la racine extraite avec une liste des racines préalablement conçue.

4.6 Conclusion

Le but de ce chapitre était de présenter la langue arabe, de décrire plus particulièrement ses propriétés linguistiques :

- Une langue voyellée qui avec l'absence de voyellation entraîne une ambiguité à différencier des unités lexicales ayant la même représentation.
- Une langue flexionnelle dans laquelle les unités lexicales varient en nombre et en flexion (soit le nombre des noms, soit le temps verbal), suivant les rapports grammaticaux qu'ils entretiennent avec les autres unités lexicales.
- une langue agglutinante où l'ensemble des morphèmes collées les unes aux autres et constituant une unité lexicale véhiculent plusieurs informations morpho-syntaxiques. Ces unités lexicales sont souvent traduisibles par l'équivalent d'une phrase en français.
- Une langue pro-drop où elle néglige systématiquement la réalisation morphologique du pronom sujet.

Nous avons ensuite présenté la classification traditionnele tripartite -verbe, nom et particule-, puis nous avons décrit une classification structurale récente des unités lexicales en arabe, ainsi elles se répartissent en cinq classes : nom, verbe, particule, résiduel et ponctuation.

Nous avons donné un aperçu sur les différentes ressources linguistiques disponibles en arabe, à savoir les lexiques monolingues et multilingues, et les corpus bruts et annotés. Finalement, nous avons présenté les outils de TAL utilisés en arabe soit les analyseurs morphologiques et les racineurs de l'arabe.

CHAPITRE 5

Identification des termes complexes

L'un de nos objectifs présentés dans ce mémoire est l'amélioration des performances des systèmes de recherche d'information en langue arabe par la prise en compte des unités lexicales complexes qui sont plus précises que les unités lexicales réduites à un seul terme. De plus, traiter le problème de leur variation permettra d'introduire de la flexibilité dans la procédure d'appariement. Différentes réalisations linguistiques seront regroupées et considérées équivalentes si elles portent le même contenu informationnel. Dans leur grande majorité, les systèmes actuels de recherche d'information se contentent d'exploiter les unités simples et n'effectuent que peu de traitements de nature linguistique [119]. Ces traitements se limitaient à la troncature des mots extraits du corpus et à l'utilisation d'anti-dictionnaire de la langue. Ces traitements, bien que rudimentaires, sont toujours appliqués dans les travaux sur l'indexation car ils sont aisés à mettre en oeuvre sur des grands corpus. Ces travaux ne tiennent pas compte des phénomènes linguistiques tels que la variation morphologique, lexicale, syntaxique ou sémantique [70].
L'objectif ici est donc de pouvoir améliorer les performances (tant le rappel que la précision) des SRI en langue arabe en apportant des réponses au double problème de l'ambiguïté des unités lexicales simples et de leurs variations. Pour ce faire, nous avons développé un système d'identification de termes complexes sur corpus aisément portable d'un corpus à un autre (cf. chapitre 6) et produisant des résultats de bonne qualité en terme de précision/rappel, en s'appuyant sur une approche mixte qui combine modèle statistique et données linguistiques pour obtenir des termes complexes pouvant servir à représenter des documents.
Nous débutons ce chapitre par une étude linguistique menée sur corpus afin d'établir les spécifications linguistiques nécessaires à l'acquisition des termes complexes. Nous caractérisons ensuite les différentes variations que peuvent subir ces termes. Pour la découverte des termes complexes, nous utilisons plusieurs types de méthodes : l'analyse partielle qui permet une formalisation linguistique des spécifications linguistiques que nous privilégions par rapport à une analyse par frontières. Nous décrivons les régles morphologiques et nous abordons la technique choisie pour filtrer les noms composés, les modèles statistiques, l'évaluation de ces modèles, et la justification du choix final de n'en retenir qu'un seul.

5.1 Spécifications linguistiques des termes complexes

Toutes les définitions du terme que propose la littérature [109] [17] [34] pour ne citer qu'elles, en font un objet à deux facettes : une forme linguistique et une fonction de référence à une notion ou à un concept. Il faudrait en ajouter une troisième, celle du domaine, puisqu'un terme n'existe qu'en relation avec un domaine de spécialité [100].
Les termes se répartissent dans deux catégories, celle des termes simples qui sont constitués d'un seul «

mot plein »[1] et qui ne peuvent être reconnus qu'en fonction de leur contexte [69]. Ils sont le plus souvent polysémiques, même à l'intérieur d'un domaine [74]., et celle des termes complexes qui contiennent au moins deux mots pleins, éventuellement reliés par des « mots grammaticaux »[2].

5.1.1 Termes complexes

Il est essentiel d'obtenir un consensus sur ce que nous entendons par terme complexe. De prime abord, nous pouvons penser que la littérature va nous fournir une définition unique et précise de ce concept. Pourtant, c'est loin d'être le cas. Dans ce domaine, les réfléxions sont multiples et varient beaucoup selon l'approche choisie.

Le terme complexe peut s'apparenter à un nom composé. La majorité des recherches porte sur les rapports internes et externes des noms composés. Par rapports externes, nous entendons le comportement du nom composé dans la phrase. Les rapports internes sont, pour leur part, des dépendances entre les constituants d'un nom composé. Afin de déterminer les rapports établis entre chaque composant d'un nom composé, des recherches fondées sur la syntaxe ont été menées sur un large éventail et permettent ainsi de mieux connaître leur différentes caractéristiques.

Dans [91], Martinet définit le nom composé comme « un signe linguistique que la commutation révèle comme résultant de la combinaison de plusieurs monèmes [3] mais qui se comporte vis-à-vis des autres monèmes de la chaîne comme un monème unique ». Toutefois, il y a des différences entre cette appellation créée par Martinet et celles d'autres auteurs. Selon [58], Martinet classe parmi les termes complexes les mots dérivés, structure qui relève pour la plupart des auteurs, de la dérivation et non pas de la composition. De plus, dans [58], Gross considère que la définition de Martinet ne prend pas en considération la question sémantique. En effet, pour Martinet un terme complexe se défini principalement par rapport à son comportement syntaxique.

Dans [11], Benveniste a proposé le nom synapsie pour désigner un type particulier de composition qui, d'après lui, se produit principalement dans le domaine technique. En relevant le caractère spécial de ce type de composition, il mentionne qu'il consiste en un groupe entier de lexèmes, reliés par divers procédés, et formant une désignation constante et spécifique. Pour ce qui est du comportement syntaxique interne des noms composés, il a démontré que la composition nominale est une micro-syntaxe. Pour sa part, [80] fait ressortir que les relations syntaxiques entre les composants des termes complexes obeissent à un ordre hiérarchique. Il démontre à l'aide de plusieurs exemples que la structure hiérarchique est établie par paires de deux mots et qu'elle est progressive.
Du point de vue de l'axe syntagmatique, cette hiérarchie peut donner lieu à des segments très longs et entraîner une perte de cohésion entre les éléments. En effet, plus longue est la description plus elle s'approche de la paraphrase.

Ainsi, même s'il n'existe pas un consensus général sur la définition de composition, il existe un au niveau de la structure syntaxique des unités complexes [11]. Nous adoptons la même approche que [11] pour l'acquisition des unités complexes en arabe, ce sont les séquences complexes de structures Nom

[1] Mot dont le rôle essentiel est de porter un contenu (nom, verbe, adjectif).
[2] Mots dont le rôle essentiel est d'apporter la cohésion grammaticale de la phrase
[3] unité minimale de sens

Préposition Nom, Nom Adjectif et Nom Nom.

La variation terminologique est un phénomène linguistique que les chercheurs ont tenté de quantifier comme indicateur de l'activité scientifique. Depuis des travaux fondateurs de [122] en recherche d'information, la variation terminologique est passée au statut de problématique linguistique à part entière [60]. La création d'outils linguistiques adéquats [26] a permis de mesurer son importance et d'en faire l'étude en corpus. Ces outils ont été intégrés dans des applications terminologiques ou de recherche d'information [7]. Ils permettent de traiter des phénomènes complexes de variations comme celles relevant de morphologie [134], de la morphosyntaxie ([28] ; [73] ; [131] et de la sémantique comme l'hyponymie [96], synonymie [64], diachronique [46] ou les variations fondées sur la socioterminologie [50] , c'est à dire une perspective d'analyse de la terminologie qui privilégie le fonctionnement de la langue dans son usage social. Analyser l'usage implique la prise en compte des variations, car l'usage n'est pas toujours stable. L'auteure présente une liste de variantes les plus connues dans les langues de spécialité à savoir les variantes terminologiques : géographiques, discursives, temporelles, interlinguistiques et cognitives.

Nous avons tiré profit des travaux existants sur ces différentes langues pour établir un classement des termes complexes de l'arabe selon leurs structures morphosyntaxiques.
Par ailleurs, nous nous posons quelques questions qui paraîssent pertinentes à notre égard et que nous essayerons de répondre à la section 5.1.1 et 5.1.2 : quelle est la typologie la plus représentative des termes complexes en langue arabe ? les variations affectant les termes complexes sont elles suffisament nombreuses pour nécessiter d'être prises en compte ?

Notre travail se présente en deux parties : la première consacrée à la typologie des structures élémentaires, la deuxième aux variations qui peuvent affecter les termes complexes. Le domaine étudié est celui de l'environnement, les motivations de ce choix sont présentées dans le chapitre 6. Cette étude linguistique a été menée sur un corpus présenté dans le chapitre 6.

5.1.2 Typologie, composition des termes complexes terminologiques du domaine de l'environnement

Dans cette partie, nous présentons une étude linguistique sur les termes complexes du domaine de l'environnement. À notre connaissance, il n'existe pas d'étude sur la typologie générale des termes complexes en langue arabe. Il nous a semblé important de préciser quels types élémentaires de termes complexes sont effectivement présents dans ce domaine technique. Ensuite, nous examinons les différentes structures morphosyntaxiques des groupes nominaux complexes arabes et nous établirons leur classement en fonction de leur structure morphosyntaxique.

Classification des structures élémentaires

La méthodologie que nous avons adoptée est la suivante : nous avons examiné le corpus et extrait des suites de mots susceptibles d'apparaître dans des positions syntaxiques variées et qui appartiennent à l'un des types élémentaires décrits par [26] . Les termes complexes épousent des structures morphosyntaxiques exprimées en partie de discours (cf. chapitre 6). Pour vérifier le caractère terminologique du candidat terme extrait, deux solutions sont envisagées : la première consiste à utiliser la base terminologique AGROVOC [4] pour attester le candidat terme extrait. Si ce dernier n'existe pas dans la base

[4] www.fao.org/agrovoc/

terminologique, une deuxième solution est envisagée qui consiste à chercher la traduction de ces suites de mots en exploitant la propriété de compositionalité des sens des termes complexes. Ainsi, nous avons tiré profit de leur traduction française pour vérifier leur statut terminologique dans la banque terminologique Eurodicautom [5]. La plupart des termes de structure N1 N2 en arabe se traduisent par des termes de structure N1 PREP N2 en français. Nom Adjectif (N ADJ) Ce type élémentaire de terme complexe est formé d'un nom et d'un adjectif. L'adjectif s'accorde en définitude (1) en ajoutant l'article ال, en genre (2) en ajoutant le suffixe ة dans le cas du féminin et en nombre (3) par l'ajout du suffixe ة.

Les règles d'accord sont diversifiées et complexes ; citons à titre d'exemple que si le nom est un pluriel « brisé » [6], l'adjectif est au féminin singulier même si le nom est masculin.

1. - التلوث الكميَاءيّ (Lit. **la** pollution **la** chimique) (pollution chimique) *(m)*
2. - كَارثة طبيعية (Lit. catastrophe normal**e**) (catastrophe naturelle) *(f)*
3. - مخلفَات صنَاعية (Lit. déchets industriels**s**) (déchets industriels) *(f)*

Nom1 Nom2 (N1 N2)

Ce type élémentaire est formé de deux unités lexicales pleines. Cette structure syntaxique est très commune (Table. 5.2) et beaucoup plus fréquente qu'en français[99].

Il est défini par une relation d'annexion (al-idâfa) entre ces deux unités lexicales. L'annexion (ou dépendance) est l'équivalent du nom + complément du nom. Le N1 est complété et défini par le N2, sans recours à une préposition. Le N1 de l'annexion est toujours un nom défini par le N2.

Ci dessous quelques structures que nous avons rencontrées dans notre corpus :

- تلوث المَاء (Lit. pollution **l'**eau) (pollution de l'eau)
- تلوث الهوَاء (Lit. pollution **l'**air) (pollution de l'air)

Nom1 Préposition Nom2 (N1 PREP N2) La structure de type élémentaire est formée de deux unités lexicales pleines et d'une préposition. Les prépositions utilisées sont ب، ل، من correspondant respectivement aux prépositions françaises au, a et de. Ci-dessous quelques exemples :

1. N1 PREP N2 (PREP = ب)
   - N1 ب N2
     - التلوث بَالرصَاص (Lit. la pollution au le plomb) (pollution au plomb)
2. N1 PREP N2 (PREP = ل)
   - N1 ل N2
     - التعرض ل الَامرَاض (Lit. l'exposition à les maladies) (exposition aux maladies) *(f)*
     - المعَالجة ل السموم (Lit le traitement de les poisons) (traitement des poisons) *(m)*
3. N1 PREP N2 (PREP = من)
   - N1 من N2
     - التخلص من النفَايَات (Lit. l'élimination de les déchets) (élimination des déchets) *(m)*

[5] http://ec.europa.eu/eurodicautom/Controller

[6] Il suit une diversité de règles complexes dépendante du nom. Le phénomène du pluriel irrégulier dans l'arabe pose un défi à la morphologie, non seulement à cause de sa nature non concaténative, mais aussi parce que son analyse dépend fortement de la structure [79]

Afin de vérifier la representativité de ces structures morphosyntaxiques par rapport aux autres structures, nous avons calculé leurs occurrences sur le corpus $[AR - ENV]$ (chapitre 6) et sur la base terminologique AGROVOC.

Structures morphosyntaxiques	Occurrence
N ADJ	13 303
N1 N2	24 851
N1 PREP N2	1073

Table 5.1 – *Nombre de candidats termes de la base terminologique AGROVOC*

Structures morphosyntaxiques	Occurrence
N ADJ	23 037
N1 N2	44 185
N1 PREP N2	18 342

Table 5.2 – *Nombre de candidats termes de* $[AR - ENV]$

D'après notre étude en corpus, nous avons remarqué l'existence d'autres structures qui sont moins représentées. Ce sont les types élémentaires formés de trois unités lexicales pleines dont voici quelques exemples :

1. N1 ADJ PREP N2
   - التلوث النفطي ل التربة « Lit. la pollution le pétrolier pour le sol » «la pollution pétrolière du sol » *(m)*
2. N1 PREP N2 ADJ
   - التعرض ل المجَالَات المغنَاطيسية « Lit. l'exposition à les champs magnétiques » «exposition aux champs magnétiques » *(m)*
3. N1 N2 ADJ
   - تلوث الميَاه الجوفية «Lit. pollution les eaux les souterraines » « pollution des eaux souterraines » *(m)*
4. N1 ADJ1 ADJ2
   - التلوث النفطي البحري «Lit. la pollution le pétrolier le marine » « pollution pétrolière marine » (m)

Ces structures peuvent être amalgamées aux structures de termes complexes modifiés ou coordonnés. La différence entre terme modifié ou coordonné se situe au niveau de la stabilisation ou non du concept terminologique. Cette stabilisation du concept est garantie par la présence d'une abréviation ou par la présence du terme dans une liste terminologique [29].
La notion de stabilisation du concept peut être entendue de deux manières. Au niveau morphosyntaxique, on considère comme stable une séquence de mots qui ne permet pas d'insertion d'autres éléments linguistiques. Au niveau sémantique, les mots qui constituent une lexie complexe n'ont pas d'autonomie

contextuelle, si bien que le parcours interprétatif attribue un sens à la lexie, mais non à ses composants. En outre, de nombreux travaux établissent la stabilité d'une notion à partir du nombre d'occurrences de l'unité qui la caractérise, ou de la distribution, dans le lexique des termes, des éléments qui composent cette unité.

Les termes complexes comprenant trois unités lexicales pleines ont été estimés par [98] pour l'anglais à 5 % de l'ensemble des termes du domaine, cette estimation nous donne une idée de la marginalité de ce type de construction qui semble être partagée par l'arabe.

5.1.3 Variation des termes complexes

La variation de la forme linguistique des termes complexes est un phénomène qui relève de la terminologie textuelle qui refuse le caractère fixe attribué a priori aux notions terminologiques. Elle envisage le terme comme un construit, c'est-à-dire comme le fruit d'une analyse faite par le terminologue qui prend en compte sa place dans un corpus. Ce phénomène, aux multiples conséquences pour tous les usagers professionnels de terminologies, apporte un nombre de difficultés dès qu'il est question de mettre en place un traitement automatique [70], mais s'avère riche en information car il peut révéler des tendances au sein de domaines scientifiques non encore stabilisés conceptuellement [113].

S'il existe un consensus, sur l'existence et l'importance du phénomène, il disparaît lorsque l'on cherche à préciser ce que chacun entend par «variation des termes». Les types de variations reconnues dépendent de l'application envisagée [29]. La variation peut être comprise comme un phénomène qui n'entraîne aucun déplacement de sens et qui fait de deux formes variantes de parfaits synonymes. Au contraire, elle peut désigner des phénomènes de modification de forme dans lesquels le terme variant subit un déplacement de sens tout en gardant le même référent. Il faudrait être capable de déterminer jusqu'à quelle modification on peut aller sans rompre le lien avec le terme référent.C'est ainsi que s'impose le constat de la variation terminologique : étant donné un domaine d'activité, il n'y a pas une terminologie qui représenterait le savoir sur le domaine, mais autant de terminologies que d'applications dans lesquelles ces terminologies ont été utilisées. Ces terminologies diffèrent quant aux unités retenues et à leur description selon l'application visée. En recherche d'information, [72] pointe le silence généré par la non-identification des variantes de termes pour l'accès à l'information. Si une variante d'un terme contrôlé n'est pas reconnue, les documents pertinent ne sont pas indexés. Dans les systèmes de questions-réponses appliquées aux textes techniques [40], les variantes de termes rendent difficile l'obtention de réponses précises à des questions spécifiques. Cependant, la variation terminologique s'est révélée féconde dans l'acquisition de nouveaux termes [71], la structuration terminologique [118], leur gestion [19] ou encore dans la construction de terminologie [28]. Dans les systèmes Questions/réponses [110], le repérage des variantes représente une aide précieuse pour apparier correctement une question à une phrase ou un segment de texte. Enfin en fouille de texte, les travaux de [20] utilisent le logiciel FASTR afin de représenter le contenu de leur corpus d'étude à partir d'un réseau de termes extraits des textes pour identifier, dans un second temps, des régles d'association.

Pour étayer notre propos, nous présenterons en détail les variations que nous avons rencontrées pour les structures de type élémentaire de l'arabe. Nous avons suivi la typologie proposée par [29]. Tous les exemples en arabe sont accompagnés d'une traduction litterale ainsi que de la traduction présente dans le dictionnaire.

Variations graphiques

Nous notons comme des variations graphiques, le remplacement de certaines lettre comme ة en fin de terme complexe par la lettre ه. Par exemple, la structure élémentaire N1 N2 apparaît sous deux graphies différentes comme « تلوث التربة» et « تلوث التربه» « pollution du sol ».

Variations flexionnelles

Ces variations regroupent les différentes formes fléchies possibles pour un terme complexe. Les flexions concernent plus particulièrement la mise au pluriel du deuxième nom dans la structure N1 N2 (1) et la définitude (2).

1. Nombre :
   - تلوث المحيط (Lit. pollution l'océan) (pollution de l'océan)
   - تلوث المحيطات (Lit. pollution les océans) (pollution des océans)
2. Définitude
   Rappelons que la définitude est réalisée, elle aussi, par un morphème, préfixé (ال).
   - التلوث الهوَاءي (Lit. la pollution la atmosphérique) (pollution atmosphérique)
   - تلوث هوَاءي (pollution atmosphérique)

Variations morphosyntaxiques

Les variations morphosyntaxiques affectent la structure interne du terme de base, et les mots le composant subissent des modifications relevant de la morphologie dérivationnelle.

1. Morphologie dérivationnelle : une variation conservant la synonymie est celle mettant en jeu un adjectif relationnel. Par exemple :
   « بءر نفطي, puit pétrolier »« بءر من نفط , puit de pétrole ». Ce type d'adjectif existe dans beaucoup de langues mais il est plus au moins fréquent. L'arabe, comme l'anglais, tend à utiliser les adjectifs de relation avec une facilité que le français n'a pas encore égalée bien qu'il semble vouloir s'engager dans cette voie [127]. Le terme d'« adjectif de relation »(ou « adjectif relationnel »)permet d'exprimer cette idée de « relation » habituellement exprimée par une préposition. Ainsi, les adjectifs relationnels (AdjR) comme les adjectifs qualificatifs s'accordent en genre, en nombre et en définitude. Ils se différencient adjectifs qualificatifs par des propriétés morphologiques, paraphrastiques qui s'appliquent soit à l'adjectif seul, soit au groupe nominal dans lequel il apparaît [26].

   (a) Propriétés morphologiques
   Les adjectifs de relation sont des adjectifs dénominaux (dérivés d'un nom au moyen d'un suffixe); il indique qu'il existe un rapport entre le nom qualifié et le nom dont l'adjectif dérive [44].
   Les adjectifs relationnels sont dérivés d'un nom grâce à un suffixe « ي » pour le masculin et « ية » pour le féminin.

(b) Propriétés paraphrastiques

Un adjectif relationnel est généralement paraphrasable par un groupe prépositionnel, mais la préposition employée, ainsi que la marque de définitude, dépendent du nom de tête au sein du terme complexe.

– « بءر نفطي , puit pétrolier »« بءر من النَفط (Lit. puit de le pétrole), puit de petrole »
– « تلوث هوَاءي (Lit.pollution air) pollution de l'air »,« تلوث في الهوَاء, pollution dans l'air »

Variations syntaxiques

Les variantes syntaxiques modifient la structure interne de la structure du type élémentaire sans affecter les catégories grammaticales des mots pleins qui restent identiques. Nous distinguons :

S-1 Les variations de modification interne : insertion d'un modifieur au sein d'une structure de terme de base.

Les modifieurs apparaîssent à l'intérieur ou après le terme de base. Les modifieurs qui apparaîssent après n'altèrent pas la structure interne, au contraire de ceux qui s'insèrent au sein de cette structure. Nous décrivons en premier les variations de modification qui prennent place à l'intérieur du terme de base, puis celles qui prennent place après le terme de base. Les modifieurs qui peuvent être insérés à l'intérieur d'un terme de base sont principalement les adjectifs.

(A) Insertion de modifieurs à l'intérieur

(a) *Adjectif*

L'adjectif s'insère à l'intérieur de la structure N1 PREP N2 juste après le N1. Concernant la structure de terme complexe N ADJ, nous n'avons pas pu trouver d'exemple qui accepte la possibilité d'insertion d'adjectif après le N, d'où la seule structure retenue pour l'insertion d'adjectif est :

– N1 PREP N2 ⇒ N1 ADJ PREP N2
  * التكوين المستمر ل لتربة(Lit. la composition la permanente du le sol) (composition permanente du sol)

(B) Postposition de modifieurs après un terme de base comprenant deux unités lexicales pleines

Les adjectifs qui se présentent après le terme complexe et qui modifient le terme entier s'accordent avec le terme complexe. Voici quelques exemples :

– N1 N2 ⇒ N1 N2 ADJ
  * درجة الحرَارة العَالية (Lit. degré la température la élevée) ([degré de température] élevé)

S-2 Les variations de coordination

La coordination unit des termes dont les schémas interprétatifs sont semblables tant par la sélection sémantique des morphèmes que par le type de lien qui les unit. Par exemple, « تلوث البحر و المحيطَات pollution de la mer et des océans » est une variante mettant en jeu les deux termes تلوث البحر « pollution de la mer » et تلوث المحيطَات « pollution des océans ». La présence de ces deux termes au sein d'une coordination dont la conjonction est « و » (et) dénote de leur proximité sémantique.

S-2a La coordination d'expansion

– تلوث الميَاه و التربة ⇒ تلوث التربة
  pollution du sol ⇒ pollution des eaux et du sol

S-2b La coordination de tête

– المخَاطر و الوقَاية من التلوث ⇒ المخَاطر من التلوث
  Risques de la pollution ⇒ Risques et prévention de la pollution

Variations paradigmatiques

Les variations paradigmatiques s'appuient sur le principe de substitution de la linguistique distributionnelle. Il s'agit de remplacer l'un ou les deux mots pleins du terme de base par l'un de leurs synonymes sans modification de la structure morphosyntaxique.
Ces synonymes n'attestent d'aucun lien syntaxique ou morphologique avec les mots pleins du terme de base. Ces substitutions simple comme : حرق الفضلَات ⟷ حرق النفَايَات, incinération des déchets ⟷ incinération des ordures, permettent d'obtenir des termes synonymes à un terme de base. Ces variantes ont été étudiées pour le français par [63].

Nous avons présenté quelles étaient les structures des termes complexes de type élémentaire et les différentes variations qu'ils subissent. Nous avons obtenu les variations suivantes :

- Variations graphiques et flexionnelles
  Sous l'entrée d'un couple se trouvent : les flexions et les graphies du terme complexe rencontrées dans le corpus. Pour le remplacement de la lettre y par la lettre A en fin de terme complexe de la structure N1 N2, nous avons considéré que les termes complexes qui admettent les deux graphies sont des variantes des termes complexes.
- Variations morphosyntaxiques
  Ce sont les variations qui affectent la structure interne du terme de base et les termes le composant subissent des modifications qui relèvent de la morphologie dérivationnelle
- Variations syntaxiques
  Elles modifient la structure interne du type élémentaire sans affecter les catégories grammaticales des termes.
- Variations paradigmatiques
  Il s'agit de remplacer l'un ou les deux mots pleins du terme de base par l'un de leurs synonymes sans modification de la structure morphosyntaxique.

5.2 Extraction automatique des termes complexes

L'acquisition de termes complexes est un domaine qui a fait l'objet de nombreux travaux de recherches ces vingt dernières années. Deux grandes directions ont été empruntées pour recenser automatiquement les termes : les modèles linguistiques et les modèles statistiques. De nouvelles recherches entreprises au cours de la dernière décennie tendent à tirer profit de ces deux grandes approches pour proposer des méthodologies qui ne sont ni purement linguistiques, ni purement statistiques (modèles hybrides). Avant d'aborder le principe de notre méthodologie d'acquisition de termes, nous présentons les différents travaux effectués dans ce domaine.

5.2.1 Les modèles linguistiques

Les systèmes présentés ici sont qualifiés de linguistiques puisqu'ils font appel à des techniques d'analyse reposant sur les connaissances actuelles de la langue et de sa structure. L'outil NOMINO est historiquement le premier logiciel à acquérir des termes d'un corpus à partir de patrons morpho-syntaxiques [32]. Le pré-traitement du corpus est ici réalisé par le logiciel lui-même en s'appuyant sur une base de données lexicale et sur des règles de désambiguïsation lexico-syntaxiques. Ces règles s'appuient sur la reconnaissance des noms et sur les expansions en syntagmes nominaux repérées à partir de ces noms. Elles sont associées aux validités probables de ces syntagmes trouvés par expansion sur la base de leur

caractérisation morpho-syntaxique. Les noms et syntagmes jugés valides dans le corpus sont alors proposés sous la forme de liste alphabétique ou par fréquence d'occurrence. LEXTER [15] extrait également des candidats termes à partir d'un corpus préalablement étiqueté et désambiguïsé en exploitant le concept de frontière de terme. [15] adopte une approche s'articulant autour d'une analyse syntaxique locale ayant pour but non pas de recenser les candidats termes à partir de matrices de formation syntagmatique des termes, mais plutôt à partir des frontières de termes. Contrairement à NOMINO qui recherche dans le corpus des formes susceptibles d'être des termes, LEXTER effectue une analyse en négatif du corpus. Cette phase repère dans le texte les éléments syntaxiques qui ne peuvent être des constituants d'un terme (verbe, conjonction, pronom, adverbe) afin de relever les syntagmes nominaux maximaux. Ces syntagmes sont alors décomposés en tête et expansion puis listés et proposés comme candidats termes. L'approche à l'origine de LEXTER, qui utilise une analyse syntaxique locale, permet d'obtenir de bons résultats. Ces derniers peuvent être obtenus plus rapidement qu'avec une approche qui repose sur une analyse syntaxique complète de la phrase. En effet, cette dernière ne peut être fiable que si le système possède une grammaire et des dictionnaires exhaustifs qui rendent possible une analyse sans faille.

5.2.2 Les modèles statistiques

Les modèles statistiques d'acquisition de termes sur corpus ont été largement utilisés depuis de nombreuses années et ils continuent de connaître de grands succés. Ils sont caractérisés par leur grande robustesse et par le fait que les documents informatisés sont de plus en plus disponibles, rendant en cela la constitution de corpus volumineux plus aisée. De nombreux travaux cherchent à associer les mots apparaissant ensemble dans un texte de manière statistiquement significative. Les travaux de [23] ont pour but de repérer automatiquement l'ensemble des collocations contenues dans un ensemble de données textuelles. Ils présentent une mesure théorique, l'information mutuelle, qui rend possible l'évaluation du score d'association entre deux formes contenues dans un corpus. Bien que les travaux de [23] se situent en marge des recherches en acquisition automatique des termes, nous avons jugé nécessaire de les présenter puisque les techniques qu'ils ont proposées sont devenues le point de départ de nombreuses recherches en acquisition automatique de la terminologie. Même si leurs travaux ont été effectués sur la langue anglaise, l'information mutuelle est purement statistique et elle est indépendante des langues.
La technique précedente a l'inconvénient de n'extraire que des candidats termes binaires. Pour contourner ce problème, l'approche des segments répétés, développée en premier lieu dans un contexte lexicométrique, peut être utilisée. Son fonctionnement consiste à identifier dans le texte toute suite d'unités textuelles reproduite sans variations à plusieurs endroits d'un corpus. Cette méthode permet de détecter des objets linguistiques très hétérogènes comme des morceaux de syntagmes nominaux plus au moins figés ou des fragments de textes récurrents moins intéressants. Les résultats sont bruités pour être utilisés directement dans un cadre d'acquisition de termes, mais peuvent fournir un point de départ pour d'autres techniques. Les résultats obtenus par les méthodes statistiques sont intimement reliés aux corpus utilisés et ne peuvent être interprétés en dehors de ce contexte. On doit aussi s'assurer que les corpus analysés possèdent une taille suffisamment grande pour que les résultats soient significatifs. On considère généralement que l'application de techniques statistiques à des corpus de taille inférieure à 100 000 occurrences ne conduit pas à l'obtention de résultats fiables et justifiables.

5.2.3 Les modèles hybrides

Les modèles hybrides sont une combinaison entre les modèles linguistiques et les modèles statistiques. L'approche présentée adopte un ordre de traitement qui varie. En effet, certains auteurs préfèrent

commencer le traitement des corpus par une analyse linguistique dont les résultats sont filtrés à l'aide de techniques statistiques, ou bien par un couplage plus intime.

Approche linguistique suivie d'une approche statistique

L'outil ACABIT [26] est un outil d'acquisition terminologique sur corpus qui se compose de deux deux étapes :

- Un repérage linguistique des termes à l'aide de régles simples appliquées par des transducteurs au corpus étiqueté,
- Un filtrage statistique des candidats termes retenus à l'étape précédente.

Plusieurs indices statistiques ont été testés pour cette deuxième phase et leurs performances ont été comparées et rapportées dans [26].
[49] adhère également dans l'utilisation d'une approche linguistique suivie d'une approche statistique. L'évaluation des candidats termes extraits se fait par le biais d'un indice, la C-value.

Imbrication complexe des approches linguistiques et statistiques

Le système CLARIT [133] développé dans un but d'indexation, cherche à obtenir des séquences de mots décrivant au mieux le contenu d'un document. Il extrait des termes complexes et présente de nombreux points communs avec des techniques dédiées à l'acquisition de terminologie. Le principe est le suivant :

1. Extraction de tous les syntagmes nominaux et étiquetage catégoriel des constituants,
2. Analyse en dépendances des syntagmes en s'appuyant sur les catégories des mots et sur les formes attestées trouvées dans le corpus,
3. Génération des termes possibles à partir de l'analyse des syntagmes.

Durant la phase 1, un processus itératif permet de détecter ce que les auteurs appellent des atomes lexicaux, en comparant les fréquences d'apparition de leurs constituants. L'analyse de dépendance dans la phase 2 a pour objectif de regrouper deux à deux les mots adjacents d'un syntagme pour trouver la configuration la plus informative et restrictive au vu du corpus. Enfin, la phase 3 génère les termes d'indexation répondant à des schémas jugés intéressants par les auteurs dans le cadre de leur recherche d'information. Les résultats obtenus sont de bonne qualité et montrent que les performances de leur système sont améliorés par l'emploi de ces termes d'indexation complexes. Dans la section suivante, nous présentons l'approche que nous avons adoptée pour extraire d'un corpus des termes complexes qui combine modèle statistique et données linguistiques. Nous prenons en compte les résultats de l'étude linguistique de la section 5.1, et exposons quels sont les structures des termes complexes prises en compte, ensuite nous décrivons la méthode d'extraction des co-occurrences basée sur l'analyse partielle par patrons. Enfin, nous proposons de tester les modèles statistiques pour l'identification des descripteurs de document en RI.

5.2.4 Principe de la méthodologie

L'objectif de notre travail consiste à définir une méthode d'acquisition de termes complexes à partir de corpus pouvant servir à représenter les documents en recherche d'informations.
Nous partons de l'idée qu'un texte n'est pas seulement un sac de mots, mais c'est un ensemble fortement structuré de termes qui permettent de communiquer des informations d'une grande précision. Les

mots simples ne peuvent pas être considérés comme un langage de représentation expressif et précis du contenu sémantique. Le but que nous nous fixons est d'extraire les termes complexes :

- En adoptant la même approche que dans [29]. Nous recherchons les termes qui épousent l'une des structures des termes complexes de l'arabe présentées en section 5.1.
- En utilisant les statistiques pour distinguer parmi ces candidats termes (CTs) lesquels sont effectivement des termes complexes. Dans ce cas, il s'agit de décider quelle mesure est la plus adaptée à l'arabe et à la recherche d'information.

La section suivante présente l'analyse linguistique où nous utilisons l'analyse partielle par patrons pour l'extraction de termes complexes ainsi que les régles morphologiques appliquées. Ensuite, nous proposons de tester les mesures statistiques permettant d'extraire pour l'arabe les patrons identifiés caractéristiques des termes complexes.

5.2.5 Analyse linguistique

Pour la découverte des termes complexes et ses variantes, nous privilégions l'utilisation d'une analyse partielle qui permet une formalisation des spécifications linguistiques, par rapport à une approche par frontières. Nous utilisons l'analyse morphologique pour permettre d'identifier certaines variantes des termes complexes relevant de la morphologie.

Analyse partielle

L'analyse partielle a pour but de reconnaître les segments non récursifs d'un énoncé. Ainsi, elle fournit une première analyse linguistique superficielle de la phrase [3]. Cette analyse est robuste car elle permet de traiter du texte et n'est pas gênée par des phrases trop longues, des structures syntaxiques inhabituelles ou une incomplétude des lexiques ou des grammaires. Sa mise en oeuvre est aisée, puisqu'elle s'exprime à l'aide d'expressions régulières ou à l'aide d'automates finis. La grammaire d'une analyse partielle comprend deux types de règles qui s'appliquent sur un texte préalablement étiqueté.

- Recherche des patrons typiques
  L'identification des patrons typiques s'effectue par la recherche de certains types de syntagmes nominaux en tenant pour acquis que ceux-ci se composent de séquences de parties de discours. Il s'agit donc de définir les patrons admissibles sous forme de règles et de localiser les séquences correspondantes. Les patrons utilisent des étiquettes prédefinies . Or les jeux d'étiquettes diffèrent d'un étiqueteur à l'autre autant en nombre que sur les catégories prises en compte et des distinctions faites à l'intérieur de ces catégories. A l'aide de l'analyseur de Diab [35], nous pouvons exprimer les différents patrons identifiés en section 5.1.2. Les patrons de base recherchés sont les suivants pour l'arabe :
  – un nom et un adjectif,
  – un nom et un autre nom,
  – un nom, une préposition et un autre nom
- Recherche des termes au moyen de frontières
  La seconde technique pratique différentes coupes dans le texte en s'appuyant sur des parties de discours qui ne contribuent pas à former des termes. Elle permet ainsi d'identifier des termes complexes.
  Cette technique consiste à identifier des frontières de termes au moyen d'une série de repères dont nous donnons un bref aperçu ci-dessous. Les premiers indices sont définis comme des repères non

ambigus. Il s'agit de :

- Un signe de ponctuation (?, :, !, ,);
- Un verbe conjugué;
- Une conjonction de subordination;
- Un pronom.

Ces premières règles de coupe, appliquées telles quelles à une partie du texte, produisent le découpage illustré ci dessous et mettent au jour une liste de candidats potentiels. Les coupes sont représentées au moyen du symbole #.

NN/ اسرَاع IN/ ل JJ/ كيمَاوية NN/ بودرة
PUNC/ ، CC/ و NN/ السرطَان vb/ سبّبت # NN/ المَاكلَات NN/ انضَاج
PUNC/ . JJ/ الكلوي NN/ الفشل
PRP/ هو # JJ/ البصري NN/ التلوث
NN/ التلوث NN/ انوَاع NN/ احد NN/ بسَاطة IN/ ب

1. « بودرة كيمَاوية ل اسرَاع انضَاج المَاكولَات »« poudre chimique pour mûrissement de nourriture »
2. « السرطَان و »« le cancer et »
3. « الفشل الكلوي »« insuffisance rénale »
4. « التلوث البصري »« pollution visuelle »
5. « ب بسَاطة احد انوَاع التلوث »« Simplement un des types de pollution »

Nous avons mené des expérimentations sur un échantillon de 100 documents de notre corpus $[AR-ENV]$ pour répondre à la question que nous nous sommes posée sur la pertinence d'une approche par patrons par rapport à une analyse par frontières, en calculant la précision sur les 100 premiers termes extraits.

Type d'analyse	P(%)
Patrons	60 %
Frontières	35 %

Table 5.3 – *Comparaison entre l'analyse par patrons et l'analyse par frontières*

Les résultats montrent une différence s'avérant statistiquement significative de 40 % en faveur de l'analyse par patrons par rapport à l'analyse par frontières, ce qui confirme notre décision d'adopter une approche d'extraction des termes complexes par patrons qui repose avant tout sur le postulat qu'une description méthodologique par patrons facilitera l'adaptation de nos travaux sur la langue arabe qui ne sont pas facilement décrits à l'aide d'une analyse par frontières. Vu les particularités de l'arabe à savoir l'absence de ponctuation dans les textes peut engendrer une ambiguité à détecter les frontières des termes. Ainsi les patrons pris en compte pour l'identification des variations syntaxiques sont :

- Nom ADJ PREP Nom
- Nom PREP Nom ADJ

- Nom1 Nom2 ADJ

L'analyse par patron paraît insuffisante pour l'identification de certaines variantes relevant de la morphologie, ce qui nous amène à utiliser l'analyse morphologique pour acquérir d'autres variations de termes complexes.

Analyse morphologique

L'analyse morphologique s'appuie sur des régles générales qui permettent à partir d'une forme dérivée potentielle de retrouver un lien morphologique avec une autre forme.

Morphologie flexionnelle

La morphologie flexionnelle concerne tout ce qui a trait à la conjugaison, au genre ou au nombre de mots. Pour l'identification des variations flexionnelles, nous avons procédé de la même manière que pour l'identification des variations morphosyntaxiques en créant des règles de désuffixation/recodage 5.4 qui nous permettent de regrouper sous le même paradigme des termes complexes qui diffèrent en nombre et en définitude.

Suffixe	Exemples de règle		Traduction
-ات	-ات/ة	دراسات/دراسة	études/étude
-ون	-ون/	معلمون/معلم	instituteurs/instituteur
-ال	-ال/	التلوث/تلوث	la pollution/pollution

Table 5.4 – *Exemple de règles des variations flexionnelles*

Morphologie dérivationnelle

La morphologie dérivationnelle est la branche qui s'occupe de la formation de mots nouveaux à partir de mots existants, particulièrement via l'ajout de suffixes et préfixes (dans « بركاني »« volcanique », le suffixe ي s'adjoint à la base nominale « بركان »« volcan »pour créer un adjectif). Pour identifier les variantes morphosyntaxiques, nous avons créé pour chaque suffixe des règles de désuftixation/recodage de manière à générer les formes les plus prédictibles de noms dérivés possibles. Ces règles sont établies manuellement à l'aide des descriptions linguistiques décrites par [59]. Le tableau ci dessous resume les suffixes rencontrés dans notre corpus.

Suffixe	Exemples de règle		Traduction
ي	/ي	بركان/بركاني	volcanique/volcan
اني	/اني	نفس/نفساني	psychologue/psychique
وي	ة/وي	كرة/كروي	sphérique/sphère

Table 5.5 – *Exemple de règles des variantes morphosyntaxiques*

Nous avons identifié certaines variations qui subissent ces termes complexes qui relèvent de la syntaxe en se basant sur une analyse partielle par patron. Pour detecter les variations flexionnelles et dérivationnelles, nous avons effectué une analyse morphologique tout en créant des règles de désuffixation/recodage. Par contre, nous n'avons pas pris en compte les variations paradigmatiques.

5.2.6 Analyse statistique

La stratégie adoptée est l'extraction des séquences morphosyntaxiques caractéristiques des types élémentaires comprenant deux unités lexicales. Ces séquences morphosyntaxiques constituent une liste de candidats termes potentiels, cette liste de candidats termes sera soumise à diverses mesures statistiques. Ces mesures permetteront de calculer le potentiel terminologique de la séquence rencontrée. [75] introduit deux notions, l'unithood et le termhood. L'unithood fait référence à la mesure dans laquelle une séquence de mots forme une unité linguistique. Il comprend les mesures du Loglike ratio, l'information mutuelle et la fréquence. Termhood fait réference à la mesure dans laquelle une unité linguistique est liée au domaine des concepts. Il comporte les mesures proposées par [49], [131] et [92] Chaque mesure statistique repose sur un classement conceptuel des couples. Ce classement peut bien mettre plus en avant des expressions figées que des termes du domaine. Notre objectif étant d'établir une liste des termes du domaine de l'environnement, il est essentiel de déterminer quelle mesure est la plus adaptée à l'extraction des termes. Nous avons décidé de comparer les valeurs obtenues pour chaque mesure à une liste de référence des termes du domaine. Cette liste de référence est présentée ci-dessous. Cette évaluation s'effectue sur les 100 premiers couples extraits du corpus $[AR - ENV]$. Si le couple apparaît dans la liste de référence, il est considéré comme un bon candidat, sinon nous cherchons sa traduction compositionnelle en utilisant cette fois ci la banque terminologique Eurodicautom [7].

Liste de référence

Une liste de référence des termes du domaine de l'environnement peut être constituée manuellement à partir de corpus. Une alternative à ce travail repose sur l'acquisition d'une banque terminologique déjà existante de notre domaine. Nous avons utilisé la banque terminologique AGROVOC. [8]

La section suivante décrit les mesures que nous avons retenues pour trier nos candidats termes et ainsi évaluer leur performance dans un système de recherche d'information.

Fréquences

L'utilité de la fréquence pour le dépistage des termes a souvent été mentionnée et exploitée au sein de logiciels. Les travaux de Daille [26] en sont de bons exemples. La présente section s'intéresse à l'effet d'un tri en ordre décroissant de fréquence sur la précision. Encore une fois, cette dernière est évaluée sur la première partie de la liste des candidats termes.

Ce tri de la liste des candidats termes accorde un très grand intérêt aux candidats termes les plus fréquents. La répartition inégale des fréquences, telles que rencontrées dans le corpus, aura pour conséquence de rassembler les candidats termes les plus fréquents en tête de liste.

[7] http://www.agris.be/fr/research/dico.html
[8] www.fao.org/agrovoc/

LLR (LogLike Ratio)

Pour mesurer le caractère non accidentel de la combinaison de deux formes, différents calculs statistiques sont mis en œuvre. Ces mesures se fondent sur le principe d'association forte voulant que l'association récurrente de certains mots ne soit pas attribuable uniquement au hasard.

Pour illustrer le principe de l'association forte entre deux formes, prenons l'exemple « حمَاية البيئة »« Protection de l'environnement » trouvé dans notre corpus $[AR - ENV]$. Toutefois, un examen attentif de « البيئة »« l'environnement » revèle qu'il apparaît au total 744 fois. Dans 644 cas, il succède à « حمَاية »« Protection » et dans les 100 cas qui restent, il est utilisé dans une autre combinaison. De même, « حمَاية »« Protection » apparaît 307 fois et dans 200 cas, il précède « البيئة »« l'environnement »; il fait aussi partie de 107 autres combinaisons. Il est donc permis de penser que la combinaison est significative et se caractérise par une association forte.

Voyons maintenant comment mettre en œuvre un calcul statistique sur des paires de formes afin de valider l'intuition qu'on peut avoir sur le type d'association qui existe entre eux.

D'un point de vue statistique, les deux lemmes qui forment un couple sont considérés comme deux variables qualitatives dont il s'agit de tester la liaison. Les données se représentent sous la forme d'un tableau croisé, appelé tableau de contingence. Un tableau de contingence 2 x 2 utilisé en statistique pour mesurer l'association en deux variables qualitatives peut être utilisé pour mesurer l'association entre deux formes lexicales dans un ensemble C de contextes. À chaque couple de formes (F_i,F_j) est associé un tableau de contingence :

Les valeurs a,b,c et d résument les occurrences d'un couple :

	F_j	$F_{j'}$ avec j' $\#$ j
F_i	a	b
$F_{i'}$ avec i' $\#$ i	c	d

Table 5.6 – *Tableau de contingence*

a = le nombre d'occurrences dans C du couple (F_i,F_j),

b = le nombre d'occurrences dans C du couple (F_i,$F_{j'}$) $F_{j'} \# F_j$,

c = le nombre d'occurrences dans C du couple ($F_{i'}$,F_j) $F_{i'} \# F_i$,

d = le nombre d'occurrences dans C du couple ($F_{i'}$,$F_{j'}$) $F_{i'} \# F_i$ et $F_{j'} \# F_j$,

La somme $a + b + c + d$ notée N est le nombre total d'occurrences de tous les couples reconnus dans C. Le LLR, introduit par [42] est le test du rapport de vraisemblance appliqué à une loi binomiale. Il correspond à une information mutuelle généralisée et peut être exprimé avec nos indices de tableau de contingence.

$$\begin{aligned} LLR &= a\log a + b\log b + c\log c + d\log d - (a+b)\log(a+b) \\ &\quad -(a+c)\log(a+c) - (b+d)\log(b+d) \\ &\quad -(c+d)\log(c+d) + N\log N \end{aligned} \tag{5.1}$$

Dans notre application, l'ensemble C est représenté par l'ensemble des séquences textuelles relevées par nos patrons de termes et de leurs variations. De manière à augmenter la représentativité d'un couple, ces tables de contingence ont été calculées sur les formes lexicales. Une quinzaine de scores d'association a

été évalué dans [30] et a montré que le LLR se comportait mieux pour la détection de termes complexes dans un contexte monolingue.

LR/FLR (Left Right/ Frequency Left Right)

La méthode proposée par[97] repose sur l'idée que les mots qui forment une unité lexicale ont tendance à apparaître ensemble plus souvent que d'autres combinaisons de mots, et que la création d'un nouveau terme se fait le plus souvent par une combinaison grammaticalement correcte de termes simples et composés déjà existants. La méthode est fondée sur le calcul du nombre distinct de termes simples qui se trouve à gauche et à droite formant le terme complexe.

1. Pour un terme simple

$$LN(N) = \sum_{i=1}^{\#LDN(N)} \#L_i \tag{5.2}$$

$$RN(N) = \sum_{j=1}^{\#RDN(N)} \#R_j \tag{5.3}$$

Où $\#LDN(N)$ et $\#RDN(N)$ sont le nombre des termes simples distincts qui précèdent ou succèdent directement le N.
LN(N) et RN(N) sont les fréquences des termes qui précèdent ou succèdent le N.
$\#L_i$ et $\#Rj$ sont respectivement les fréquences du bigramme [LN_i N] et [N RN_j] dans le corpus.

2. Pour un terme complexe
Considérons CN = N_1 N_2 ...N_L où N_i (i=1...L) est un terme simple.
Ainsi la moyenne géométrique LR du terme complexe est définie comme suit :

$$LR(CN) = \prod_{i=1}^{L}((LN(N_i)+1)(RN(N_i)+1))^{\frac{1}{2L}} \tag{5.4}$$

3. Combinaison de LR et la fréquence du terme complexe
La méthode LR telle qu'elle est définie, ne reflète pas le nombre d'occurrence du terme complexe en entier. Par exemple, si l'unité lexicale « *natural language* » est plus fréquente que « *language natural* », la méthode LR donnera la même mesure sans se soucier de la fréquence du terme dans le corpus.
Ainsi pour rémedier au problème, [97] propose la méthode FLR qui prend en compte la fréquence ainsi que la LR de la manière suivante.

$$FLR(CN) = LR(CN) * f(CN) \tag{5.5}$$

C-Value

La C-Value proposée par [49], repose sur la prise en compte de certains mots (noms, verbes et adjectifs) avoisinant le terme complexe dans le corpus. Cette technique est semblable à celle mise avant par [97] qui prend en considération l'ensemble des unités lexicales simples utilisées au sein d'un candidat terme complexe. La méthode C-Value de [49] permet d'identifier des termes complexes (comprenant

plusieurs mots) mais aussi des termes enchâssés. La mesure C-value vérifie la stabilité des termes par rapport aux termes les plus longs.

- C-value
  La mesure est basée sur la fréquence du terme t dans le corpus f(t), la fréquence du terme comme un terme enchâssé n_t, le nombre de termes enchaînant le terme N_t, et la longueur du terme (nombre de mots) l :

$$C(t) = \begin{cases} \log_2 lf(t) & \text{si t n'est pas enchâssé} \\ \log_2 l(f(t) - \frac{n_t}{Nt}) \end{cases} \tag{5.6}$$

Nous avons mesuré la performance des mesures statistiques décrites précedemment en terme de précision. Le tableau 5.7 contient les données obtenues à l'aide de ces mesures. La précision est mesurée en fonction du nombre de termes identifiés par rapport au nombre de candidats termes non valides qui se trouve dans les 100 couples de la liste.

Type	Précision
Fréquence	87 %
LLR	85 %
FLR	60 %
C-Value	51 %

Table 5.7 – *Performance des mesures statistiques*

Ainsi, dans l'ensemble des documents, un tri à l'aide de la fréquence permet d'obtenir une bonne concentration des termes en tête de la liste de candidats termes. Les performances obtenues à l'aide de la fréquence nous conduisent à conclure qu'il s'agit d'une mesure permettant de bien cerner le potentiel terminologique de certains des candidats termes recensés.

Le tableau précédent récapitule les observations faites sur les documents du corpus $[AR - ENV]$ à la suite d'un tri selon la fréquence. La performance d'un tri aussi simple est surprenante si on compare les résultats à ceux obtenus à l'aide de la LLR. Cette comparaison est d'autant plus éminente lorsqu'on prend aussi en considération la complexité de la LLR alors que la fréquence est directement observable dans le corpus. La précision moyenne obtenue à l'aide de cet indice est supérieure à celle obtenue par LLR.

Malgré l'attrait d'un tri aussi simple, il est important de mentionner son absence de finesse lorsque l'on compare des tranches de fréquence très productives. En effet, cette approche ne permet pas d'opposer les candidats termes qui partagent la même fréquence et de déterminer lequel de ces candidats termes possède un potentiel terminologique plus important.

5.3 Conclusion

Cette étude linguistique de termes complexes de notre corpus montre que les termes complexes techniques ne sont pas des structures morphosyntaxiques figées et qu'ils subissent de nombreuses transformations. L'extraction des termes complexes se heurte au problème de la reconnaissance de ces unités lexicales complexes :

- Il serait impossible de statuer sur une séquence comprenant trois unités lexicales avant d'avoir déterminé les séquences comprenant deux unités lexicales

La section suivante présente les principales caractéristiques des séquences comprenant deux à trois unités lexicales et résume leur ambiguité structurelle. Notre tâche consiste à extraire d'un corpus les termes complexes du domaine, nous ne possédons pas la liste des termes complexes élémentaires mais nous nous appuyons sur leurs structures morphosyntaxiques. Il reste que ces structures sont ambigues comme le montre l'examen des séquences qui suit :

1. Séquences comprenant deux unités lexicales
   Les séquences extraites du corpus et appartenant à l'une des structures morphosyntaxiques de type élémentaire comprenant deux unités lexicales, représentent l'une des entités ci-dessous :
   - Un terme complexe de type élémentaire
   - Un groupe nominal qui n'a pas le statut de composé
2. Séquences comprenant trois unités lexicales
   Les séquences comprenant trois unités lexicales sont énumérées ci-dessous et annotées des différentes analyses qu'elles peuvent recevoir.
   - N1 N2 Adj
     – Terme complexe de type élémentaire comprenant trois unités lexicales;
     – Terme complexe de type élémentaire comprenant deux unités lexicales de structures N1 N2 modifié par un adjectif, N1 N2 devrait exister sans modifieur.
   - N1 ADJ PREP N2
     – Terme complexe de type élémentaire comprenant trois unités lexicales;
     – Terme complexe de type élémentaire comprenant deux unités lexicales de structures N1 PREP N2 modifié par un adjectif, N1 PREP N2 devrait exister sans modifieur.
   - N1 PREP N2 ADJ
     – Terme complexe de type élémentaire comprenant trois unités lexicales;
     – Terme complexe de type élémentaire comprenant deux unités lexicales de structures N1 PREP N2 modifié par un adjectif, N1 PREP N2 devrait exister sans modifieur.

Les ambiguités attachées aux séquences comprenant trois unités lexicales sont nombreuses. À ce problème s'ajoute celui des variantes des termes complexes.

Termes complexes et leurs variantes

Pour être complet, ce travail d'extraction doit indiquer pour chaque terme complexe ses variantes. En ce qui concerne les variantes graphiques, le lien est assez aisé : des procédures de normalisation devraient nous permettre de relier par exemple « التلوث الهوَاء» et « التلوث الهوَاءي» « pollution atmosphérique ». Pour les variantes morphosyntaxiques la tâche est moins triviale, nous avons effectué une analyse morphologique fondée sur des règles de dérivation générales. Pour la détection des variantes syntaxiques, Nous avons plebiscité l'analyse par patrons qui décrit exhaustivement les structures linguistiques recherchées par rapport à une analyse par frontière qui génère plus de bruit sans véritablement gagner en couverture. Les variantes paradigmatiques n'étaient pas prévues car il nous semble difficile de relier un terme complexe à ses variantes paradigmatiques dans le cadre des ressources monolingues.
La stratégie est l'extraction des séquences morphosyntaxiques comprenant deux unités lexicales, ces structures peuvent être modifiées ou coordonées. Cette liste de candidats est soumise à diverses mesures statistiques. Ces calculs permetteront de déterminer le statut de la séquence rencontrée. En conclusion de l'examen des mesures statistiques qui ont été presentées, nous remarquons que la fréquence d'un couple est un très bon révélateur de son caractère terminologique. Le problème est qu'elle ne permet pas d'isoler les termes complexes rares et que son classement introduit beaucoup de bruit. Nous avons choisi de ne

retenir que la LLR pour : sa nature de test statistique, sa tendance à prendre en considération la fréquence du couple, son bon comportement en dépit de la taille du corpus. Dans le chapitre suivant, nous abordons les méthodes informatiques qui ont été réalisées, les méthodes d'évaluation adoptées aini que les résultats obtenus dans un système de recherche d'information.

CHAPITRE 6
RI en langue arabe

Les mécanismes de RI classiques se heurtent surtout à des difficultés de nature linguistique. Les "sacs de mots" utilisés pour représenter les documents et les requêtes pendant le processus de recherche d'information ne véhiculent que relativement leur contenu sémantique. Ainsi, en ne prenant en compte ni les relations qu'entretiennent les termes les uns avec les autres, ni l'ordre des mots, ils produisent des pertes d'informations qui sont en majorité responsables des performances limitées des SRI. La mise en correspondance entre l'information recherchée par l'utilisateur et l'ensemble des documents disponibles ne peut pas non plus être réduite, comme l'offrent maintenant la majorité des modèles de RI, à une comparaison de chaînes de caractères. La richesse et la complexité de la langue arabe sollicitent un appariement plus fin, susceptible de considérer comme éventuellement pertinent un document qui ne comporterait aucun des termes utilisés dans la requête et, en revanche, de rejeter un document non intéressant même si ce dernier possède les chaînes contenues dans la requête.

Pour pallier à ces limites, une solution assez naturelle est de recourir aux techniques du TAL pour mesurer leur apport en RI. Ainsi, ce chapitre propose d'évaluer l'intérêt de combiner en RI des informations linguistiques appartenant à la fois au niveau morphologique, syntaxique de la langue, de sélectionner les schémas de pondération qui améliorent la performance de la méthode LSA pour la recherche d'information dans un corpus spécialisé en langue arabe et de comparer la performance du modèle vectoriel avec celle du modèle LSA. Le système que nous proposons pour l'évaluation de l'intégration de connaissances s'inspire du travail de [123], puisque nous ré-utilisons l'idée de cette architecture qui nous paraît pertinente pour représenter plusieurs informations linguistiques sous la forme de descripteurs.
Dans ce chapitre, nous présentons une description de la collection de test utilisée, les informations linguistiques que nous avons choisies d'exploiter et l'architecture de test mise en place pour leur intégration.

6.1 La collection de test : corpus en langue arabe standard dans un domaine de spécialité $[AR - ENV]$

Pour démontrer l'intérêt de représenter le contenu textuel par des unités lexicales complexes dans un processus de recherche d'information, nous devons disposer d'un corpus de langue de spécialité. A notre connaissance, il n'existe pas un corpus répondant à ces critères. Ainsi, nous avons décidé de construire un corpus à partir du web dans le domaine de l'environnement, restreint aux thématiques suivantes : la pollution, la purification de l'eau, la dégradation du sol, la préservation de la forêt, les catastrophes naturelles. Elles font l'objet d'une importante production langagière en arabe, comme l'atteste la présence de nombreux sites sur le web . L'élaboration du corpus s'est déroulée en deux étapes : le moissonage du web et la normalisation des textes. Les étapes ont été réalisées par des locuteurs natifs.

6.1.1 Moissonage du web

le moissonage de documents s'est déroulé entre 2004 à 2006, ainsi nous avons effectué :

1. une recherche sur le web à l'aide du moteur http://www.google.com/intl/ar/ pour l'arabe.
2. une recherche interne sur des portails, notamment « Al-Khat Alakhdar » [1] et « Akhbar Albiae » [2] , en utilisant le cas échéant le moteur de recherche propre au site.

La circonscription de la recherche aux thématiques choisies s'effectue par l'intermédiaire de l'utilisation de mots-clés qui doivent être précis. Deux stratégies de recherche sont possibles : la première en largeur qui examine la plupart des documents renvoyés par une seule requête, la seconde en profondeur qui n'examine que les premiers documents et en explore les liens internes. Nous avons mené une recherche en profondeur sur les vingt premiers résultats en utilisant par exemple les combinaisons des mots-clés "*environnement*", "*pollution*", "*bruit*". Pour garantir une bonne couverture de la thématique, nous avons élargi la recherche en utilisant des synonymes[3] et en nous appuyant sur des termes relevés dans les pages visitées comme "*dégradation*" ou "*pollution sonore*".

6.1.2 Normalisation

Pour chacun des documents sélectionnés, nous avons enregistré son url, et l'avons converti au format UNICODE. Ce nouveau standard de codage de caractères est loin d'être la norme sur le web. Pour l'arabe, 70 % des textes collectés sont encodés sous Arabe Windows. Les formats de documents rencontrés sont principalement du HTML. Les informations de mise en page du texte (gras, italique, retrait ...) ont été perdues lors de la conversion au format .txt.

6.1.3 Caractéristiques de la collection

Le tableau présente quelques indications des caractéristiques de la collection.

Caractéristiques	Collection $[AR-ENV]$
Nombre de documents	1062
Nombre total d'unité lexicale	475 148
Nombre total d'unité lexicale différentes	54 705

Table 6.1 – Quelques données sur la collection de test $[AR-ENV]$

6.1.4 Lexique et métriques

La métrique que nous utilisons pour mesurer l'adéquation d'un lexique à un corpus est définie comme des rapports entre les différents ensembles définis ci-dessous.
Nous posons les définitions suivantes :

- Le texte T du corpus est un ensemble ordonné d'unités lexicales. Les unités lexicales sont définies par leur forme graphique et repérées par leur position dans le texte.
- Le vocabulaire V du corpus est l'ensemble des vocables, i.e. l'ensemble des unités lexicales différentes du corpus. Les vocables sont des unités monolexicales.
- Le lexique L de la ressource est l'ensemble des lexies ou entrées lexicales de la ressource, qu'elles soient composées de un ou plusieurs unités lexicales.

[1] http://www.greenline.com.kw
[2] http://www.4eco.com
[3] Présents dans le dictionnaire Mounjid tolab

- La partie utile de la ressource PU est l'ensemble des lexies de la ressource qui apparaissent dans le corpus. C'est un sous-ensemble de L.

La couverture évoque l'idée d'un corpus tout ou partiellement « couvert » par le lexique. Elle est donc calculée relativement au corpus plutôt qu'à son vocabulaire. Ainsi, la couverture (Couv) est la proportion d'occurrences de mots correspondant à des vocables entrant dans les lexies de la partie utile du lexique. Dans la formule ci-dessous : $freq_i$ représente le nombre d'occurrences d'une lexie i de PU non incluses dans une occurrence d'une autre lexie plus large et $longueur_i$ est la longueur de la lexie en nombre de mots.

$$Couv = \sum_{i=1}^{PU} \frac{freq_i * longueur_i}{|T|} \tag{6.1}$$

Afin d'évaluer la diversité du corpus, nous avons testé cette métrique sur notre collection $[AR - ENV]$. Nous avons construit un lexique témoin (AGROVOC-LHI). Ce lexique de 67536 formes lexicales contient le lexique AGROVOC [4] et le lexique d'hydrologie de l'ingénieur . Ainsi, la Figure 6.1 montre

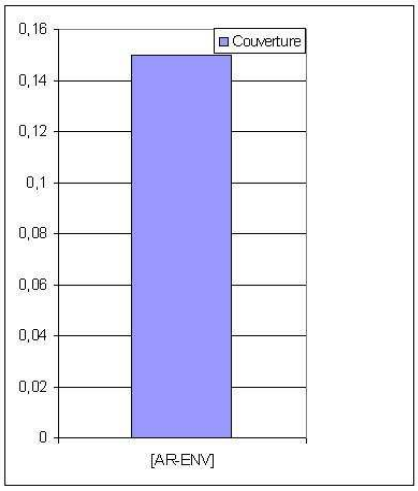

Figure 6.1 – Mesure de couverture sur le corpus $[AR - ENV]$

une couverture lexicale du corpus $[AR - ENV]$, atteignant 15 %.

[4] www.fao.org/agrovoc/

6.1.5 Distribution des catégories grammaticales

Dans le but d'étudier la distribution des catégories grammaticales (cf. Annexe A) pour en savoir la catégorie dominante sur un corpus de domaine de spécialité. Nous utilisons l'étiqueteur de Diab [35] qui fournit la catégorie grammaticale pour chaque unité lexicale. La figure 6.2 ci-dessous présente la distribution des parties du discours pour notre collection $[AR - ENV]$. Première remarque : notre corpus est riche en substantifs. Ainsi, il est employé avec une moyenne de 57,2 $\%$. Pour les trois autres catégories grammaticales, les valeurs associées sont plus faibles.

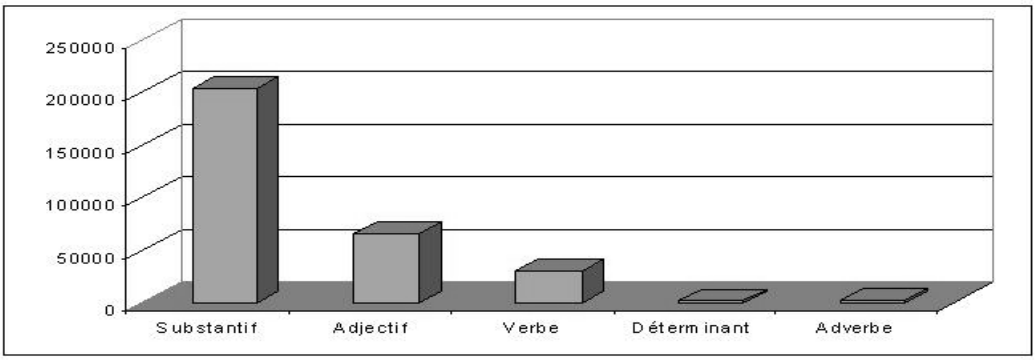

Figure 6.2 – Classement hiérarchique de cinq catégories grammaticales

6.1.6 Requêtes

Pour nos expérimentations, un jeu de 30 requêtes a été construit en s'inspirant des campagnes d'évaluation TREC. Elles comportent quatre champs : un titre nommant le thème, une description énonçant complètement l'objet de la recherche, un développement explicitant des critères de validité des rapprochements, des mots-clés fournissant le contexte terminologique et les concepts concernés. Cette forme apporte une information aussi complète et détaillée que possible, y compris des connaissances avancées sur le domaine grâce aux mots-clés.
La création des requêtes est plus ou moins calibrée et ajustée aux documents :

- les critères de validité que comprenaient les premières requêtes ont été inspirés par un premier balayage des documents du corpus;
- les requêtes qui apparaissent trop générales sont exclues;
- les requêtes perçues comme ambiguës sont également rejetées.

Un exemple de ces requêtes est présenté dans le tableau 6.1.6.

<title> حمَاية الغَابَات </title>
<desc> البحت عن النصوص التي تّحدث عن حمَاية الغَابَات </desc>
<narr> النصوص ذَات صلة بملَاحقة قَاطعي الَاشجَار طرق النهوض بَالتشجير
ونشر المسَاحَات الخضرَاء </narr>

<title> Préservation de la forêt </title>
<desc> Trouver les documents qui parlent de la préservation de la forêt. </desc>
<narr> Sont pertinents les documents qui évoquent les manières de favoriser le reboisement, la diffusion des espaces verts et la poursuite des destructeurs des forêts </narr>

Table 6.2 – Exemple de requête

Nous présentons le nombre moyen de documents pertinents par requête sur le tableau 6.1.6. Le jugement de pertinence est fourni par quatre locuteurs de la langue arabe sur l'ensemble des documents.

	$[AR-ENV]$
Nb. moyen de termes par requête	9,5
Nb. moyen de documents pertinents par requête	23,2

Table 6.3 – Nombre de documents pertinents

6.2 Architecture de connaissances linguistiques en RI

Comme nous l'avons déjà souligné, notre but est d'évaluer l'intérêt de combiner des informations linguistiques de nature variées en RI pour la langue arabe. Nous revenons plus précisement sur les connaissances d'ordre morphologique et syntaxique que nous avons retenues pour représenter les documents et les requêtes, puis décrivons l'architecture envisagée.

6.2.1 Connaissances linguistiques

Pour sélectionner les connaissances linguistiques qui enrichissent la représentation textuelle des documents et requêtes, nous nous sommes appuyées sur les nombreux travaux existants, présentés au chapitre précédent. Ces diverses études nous ont permis de recenser les connaissances de nature morphologique et syntaxique capables d'être insérées au sein d'un SRI. Nous présentons successivement les connaissances linguistiques de nature morphologique et syntaxique que nous avons choisies d'exploiter et les outils et méthodes utilisées pour leur acquisition. L'ensemble des traitements évoqués est appliqué sur les documents et requêtes issus de la collection ; seul un pré-traitement supprimant certaines diacritiques et caractères spéciaux caractéristiques de langue arabe est effectué. Tous les documents et requêtes sont pris en compte et considérés comme des termes d'indexation potentiels. Leur pondération et la suppression d'anti-dictionnaire sont effectuées ultérieurement lors de l'intégration des diverses représentations linguistiques au sein du SRI.

Connaissances de nature morphologique

Pour le niveau morphologique, nous avons choisi de considérer deux types de connaissances : connaissances morphologiques flexionnelle et dérivationnelle. Nous appliquons aux textes (et requêtes) une procédure de racinisation (stemming) qui permet d'extraire pour chaque unité lexicale sa pseudo-racine (stem). Nous nous sommes appuyées sur l'algorithme de Darwish [31] (cf. section 4.5.3), pour la normalisation des variantes morphologiques.
Le dernier type de connaissance utilisé est d'ordre morpho-syntaxique. Une analyse morpho-syntaxique des documents et requêtes est effectuée dont l'objectif est d'associer à chaque unité lexicale sa catégorie grammaticale (nom, verbe, adjectif...). Le principal intérêt de cet étiquetage est qu'il permet d'opérer un premier traitement de désambiguïsation des termes. L'étiqueteur utilisé ne peut ainsi associer qu'une seule étiquette à chaque unité lexicale ; il doit par conséquent choisir parmi toutes les catégories possibles d'une unité lexicale celle qui correspond au terme dans la phrase considérée, en s'appuyant sur son contexte d'apparition. Nous avons sélectionné l'analyseur morpho-syntaxique de [35] (cf. section 4.5.1) basé sur un apprentissage supervisé et qui donne en sortie une collection de textes étiquetés où toutes les unités lexicales ont été catégorisées.

Connaissances de nature syntaxique

Pour le niveau syntaxique, nous avons retenu deux principales structures permettant de rendre compte des relations et dépendances entre les mots. Nous prenons en compte tout d'abord les termes complexes (TCs). Leur reconnaissance est effectuée en utilisant la méthode décrite dans le chapitre 5, qui par le biais d'une analyse morpho-syntaxique des textes et de l'utilisation de méta-règles linguistiques, permet l'extraction automatique de termes complexes mais également la normalisation de leurs variantes. Nous proposons aussi d'exploiter les syntagmes nominaux présents dans les textes et les requêtes. Nous utilisons pour cela l'outil développé par Diab [35] qui, à l'aide d'une méthode combinant à la fois des techniques numériques et symboliques d'acquisition [5], identifie au sein des textes les syntagmes nominaux (SNs).

[5] La technique utilisée s'appuie plus précisément sur l'utilisation d'un algorithme d'apprentissage de règles de transformation à partir de corpus étiquetés morpho-syntaxiquement.

6.2.2 Architecture envisagée

De manière plus précise, l'architecture proposée peut être synthétisée de la façon suivante : les documents et requêtes passent dans un premier temps par un module de prétraitement, qui consiste en :

- Elimination des diacritiques
- Normalisation
  La normalisation transforme une copie du document original dans un format standard plus facilement manipulable. Cette étape est considérée nécessaire à cause des variations qui peuvent exister lors de l'écriture d'une même unité lexicale.
  Le document est normalisé comme suit :
  – Suppression des caractères spéciaux ;
  – Remplacement de آ avec ا;
  – Remplacement de la lettre finale إي ي;
  – Remplacement de la lettre finale ة avec ه.
- Tokenisation
  Le repérage des unités lexicales s'effectue par le découpage des textes en une suite d'unités lexicales à l'aide des caractères séparateurs d'unités lexicales (blanc, tabulation, ponctuation, etc.).

Après le module de pré-traitement, les documents et requêtes passent par un module de racinisation et d'élimination d'unités lexicales à l'aide d'un anti-dictionnaire pour étudier l'apport de ces derniers sur les performances d'un SRI en langue arabe, ensuite par un module de connaissances syntaxiques qui permet d'obtenir deux représentations différentes (SNs, TCs) d'un même document ou requête, comme illustré sur la Figure 6.3.

Cette architecture présente l'intérêt d'intégrer les différentes représentations linguistiques des documents et requêtes au sein du SRI. Elle nous permet d'obtenir, pour chacun des index, la liste ordonnée des documents retournés par le SRI à la suite de leur appariement. Ce sont sur ces listes que nous allons nous baser pour étudier l'apport respectif de chacune des connaissances linguistiques prises en compte. Cette étude fait l'objet de la section suivante.

6.3 Modèles de représentation

Dans le but de mieux représenter le contenu textuel, nous avons jugé intéressant d'une part d'étudier l'apport de l'analyse de la sémantique latente (modèle vectoriel étendu) par rapport au modèle vectoriel standard [115]. D'autre part, nous abordons l'influence des schémas des pondérations sur le choix de la dimension k du modèle de l'analyse sémantiqe latente ainsi que l'impact de la pondération sur les requêtes.

Après avoir représenté notre corpus en modèle vectoriel, nous avons calculé la précision moyenne sur l'ensemble des requêtes en faisant varier k, la dimension de la matrice réduite pour la LSA, entre 2 et le rang de la matrice correspondant aux documents. Les expériences menées ont prouvé que la valeur du facteur k donnant le meilleur résultat change avec la variation de la taille de la matrice initiale due au choix du type de pondération. Comme le montre [83], la valeur du facteur k est un paramètre qui influence sensiblement le résultat final.
La performance d'un système de recherche d'information est souvent évaluée en termes de deux paramètres la précision et le rappel. Nous avons choisi de calculer la précision sur 11 points de rappel. Comme ces valeurs du rappel peuvent ne pas être atteintes exactement, les valeurs de la précision seront donc interpolées. Ensuite nous avons calculé la précision moyenne qui consiste à moyenner les préci-

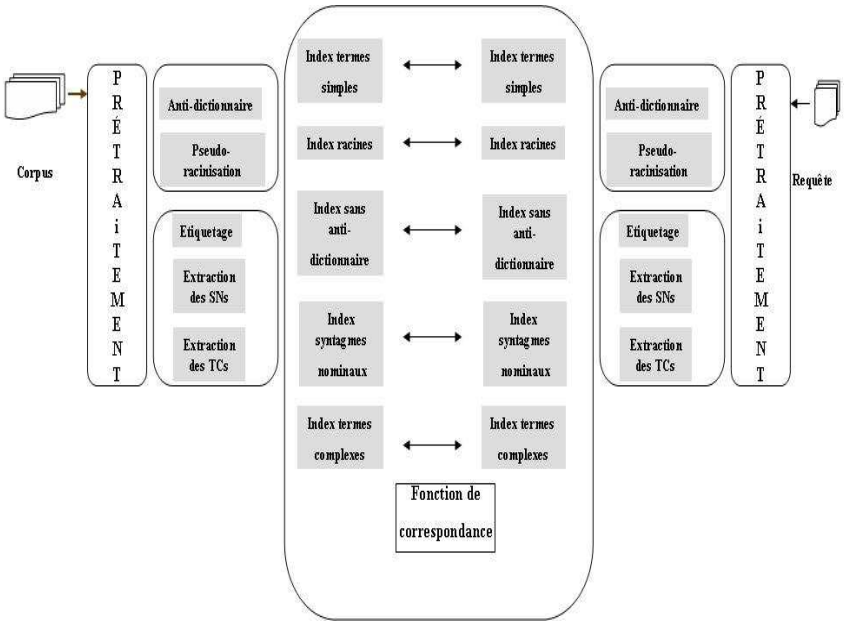

Figure 6.3 – Représentation des documents et des requêtes

sions interpolées pour l'ensemble des requêtes sur les 11 taux de rappel. Et pour le choix de la meilleure dimension k nous avons calculé pour chaque dimension une moyenne des précisions interpolées sur les 11 taux de rappel.

6.3.1 Influence des schémas de pondération

Dans le cas du modèle vectoriel standard étendu, nous avons exploré l'effet de cinq schémas de pondérations pour notre test de collection, dans deux d'études cas : requêtes courtes (comprenant le titre) FigureC.1 et requêtes longues (comprenant tous les champs) Figure 6.5 , en utilisant les schémas de pondérations les plus performants (log(tf + 1)xIdf, TfxIdf, Ltc et Tfc) trouvés dans [9] et l'Okapi BM-25. En comparant les courbes des quatre schémas de pondération log(tf + 1)xIdf, TfxIdf, Ltc et Tfc sur les figures C.1 et 6.5, nous remarquons que log(tf + 1)xIdf améliore le modèle ; alors que les trois derniers ont connu entre eux des améliorations et des dégradations sur l'ensemble des taux de rappel. Par ailleurs, le schéma de pondération Okapi BM-25 améliore encore la méthode LSA d'un gain atteignant respectivement pour les requêtes courtes et longues 6,17 % et 6,15 % par rapport au log(tf + 1)xIdf , et 16,75% et 33,33% lorsque aucune pondération n'est utilisée. Cela est dû au facteur de normalisation qui caractérise le schéma de pondération Okapi BM-25 par rapport aux autres schémas.

6.3.2 Apport du modèle LSA pour le modèle vectoriel

Nous avons étudié l'apport du modèle LSA par rapport au modèle vectoriel. Nous avons utilisé le schéma de pondération Okapi BM-25 qui, d'après les expérimentations précédentes, a été jugé comme

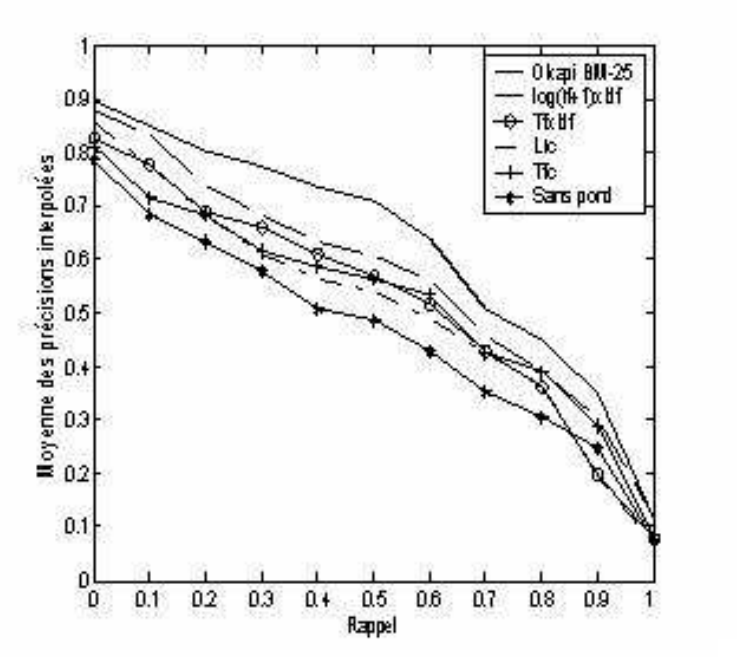

Figure 6.4 – Comparaison entre cinq schémas de pondération (requêtes courtes)

le plus performant, dans deux cas d'étude : requêtes courtes et longues. Les courbes de la figure 6.6, montrent une différence s'avérant statistiquement significative de 15,9 % pour le modèle LSA par rapport au modèle MVS, alors que les courbes de la figure 6.7 présentent un gain de 16,3 %.

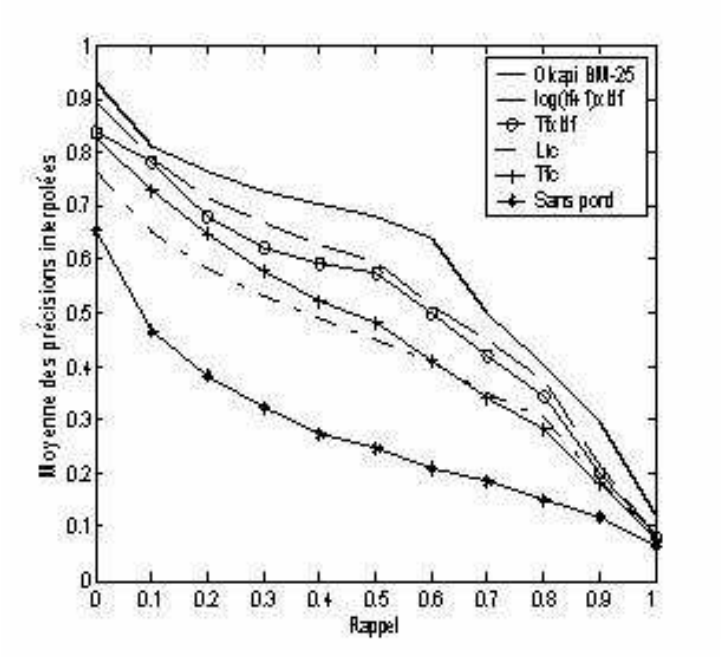

Figure 6.5 – Comparaison entre cinq schémas de pondération (requêtes longues)

6.3.3 Influence des schémas de pondération sur le choix de la dimension réduite k du modèle LSA

Nous mentionnons les paramètres influant sur le choix de la dimension réduite k du modèle LSA à savoir les types de requêtes utilisés et les schémas de pondération. Sur la base des résultats montrés dans le tableau 6.4, nous remarquons que la dimension réduite k est nettement plus intéressante dans le cas de la pondération Okapi BM-25 par rapport aux autres schémas de pondération.

6.3.4 Apport de pondération des requêtes

Nous précisons que nous avons utilisé des requêtes pondérées dans l'étude menée. Nous utilisons la pondération Okapi-BM 25 et le modèle de l'analyse sémantique latente. Dans cette sous-section, nous nous intéressons à étudier l'apport de la pondération sur les deux types des requêtes : courtes et longues.

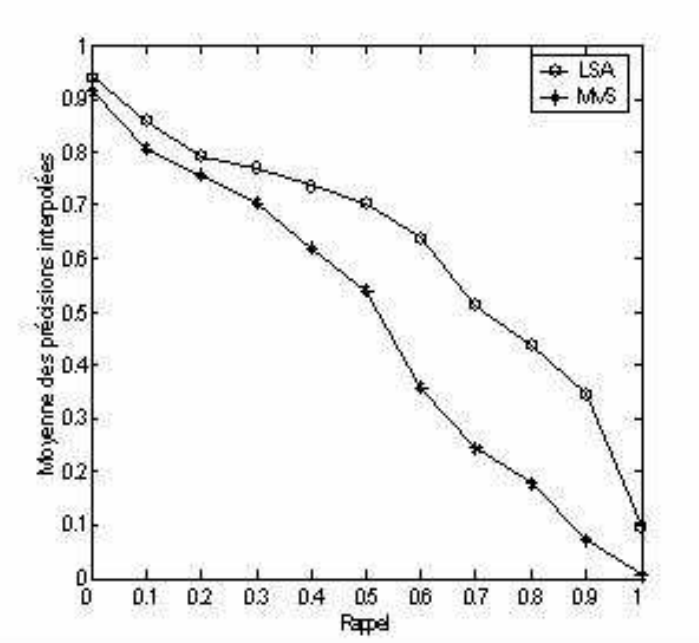

Figure 6.6 – Comparaison entre le modèle MVS et le LSA dans le cas des requêtes courtes

Pondération	Dimension k	
	Requêtes courtes	Requêtes longues
Sans pond	268	659
Okapi BM-25	122	182
Log(tf+1)xIdf	517	1019
Tfc	600	885
TfxIDF	1045	868
Ltc	1041	1016

Table 6.4 – L'influence des schémas de pondération sur le choix de la dimension k du modèle LSA

Nous avons constaté que la pondération des requêtes donne une amélioration de 2,70 % au profit des requêtes courtes par rapport aux longues ; présentée sur la figure 6.8. Le fait que cela n'est pas conforme à ce que nous trouvons dans la littérature [116], nous a poussé à chercher la cause de cette différence. Pour cela, nous avons effectué des tests pour des requêtes non pondérées où nous avons constaté que les requêtes longues présentent un gain par rapport aux requêtes courtes, atteignant 4,30 %. Alors que les requêtes courtes ont connu une dégradation de 3,60 %. Vu que les requêtes courtes reflètent mieux

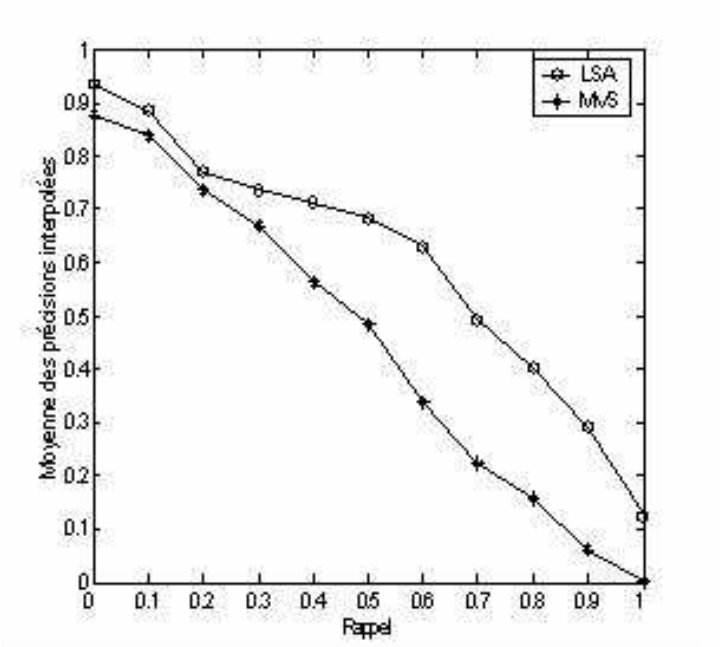

Figure 6.7 – Comparaison entre le modèle MVS et le LSA dans le cas des requêtes longues

la réalité du web, nous suggérons d'utiliser un système de recherche d'information où les requêtes sont pondérées.
Dans le but d'améliorer les performances de notre système de recherche d'information en langue arabe, il s'est avéré que l'utilisation d'un modèle vectoriel étendu permet de mieux présenter le contenu textuel des documents. En outre, la pondération des requêtes courtes permet une augmentation des performances de notre système.

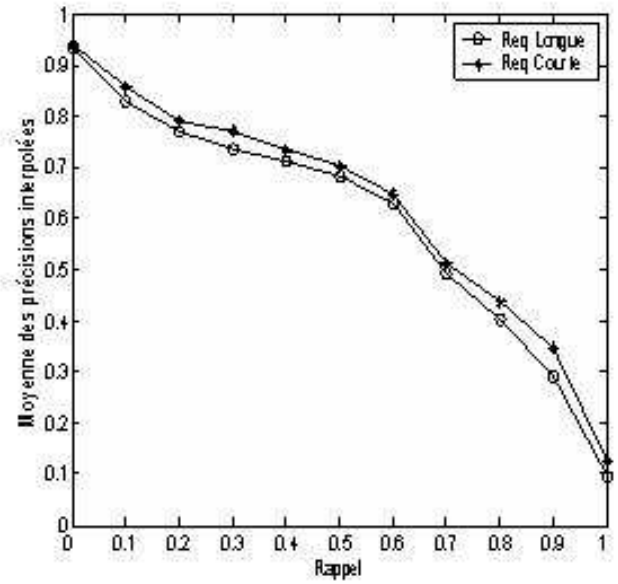

Figure 6.8 – Comparaison des requêtes pondérées longues et courtes

6.4 Impact respectif des connaissances linguistiques sur les performances des SRI

Pour évaluer l'apport individuel des connaissances morphologiques et syntaxiques, nous nous appuyons sur l'architecture de test présentée en section 6.2. Nous intégrons chacune des représentations linguistiques des documents et requêtes au sein du SRI. Ce dernier procède alors à l'appariement des index, évalue la pertinence de chaque document de la collection en fonction de la requête considérée et produit (pour chaque index) une liste de résultats qui correspond à l'ensemble des documents qu'il a retrouvé.
Nous abordons tout d'abord l'impact des connaissances morphologiques, ensuite celles des connaissances syntaxiques sur les performances de SRI en langue arabe.

6.4.1 Racinisation

Afin de pouvoir rassembler certains mots, nous avons raciné les mots de notre corpus. Le premier avantage de ce prétraitement est la réduction de la taille de la base d'indexes : la matrice relative à notre corpus possède 54 705 unités lexicales, après l'application du processus de pseudo-racinisation, nous obtenons une matrice réduite de 22 553 unités lexicales ; et avec l'utilisation de l'anti-dictionnaire [31] (voir Annexe), nous obtenons une matrice plus réduite de 22 491 unités lexicales seulement. Nous

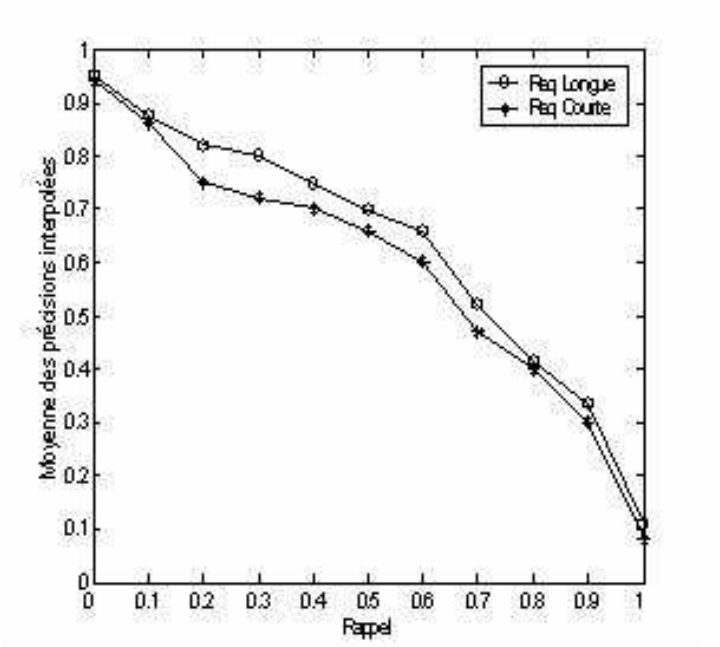

Figure 6.9 – Comparaison des requêtes non pondérées longues et courtes

avons effectué des expérimentations pour évaluer l'apport de ces prétraitements pour la recherche d'information, pour deux cas d'études : le premier où aucune pondération n'est appliquée, le deuxième où la pondération Okapi BM-25 est utilisée.
Sur la figure 6.10 (requêtes courtes, cas sans pondération), les résultats montrent que l'amélioration apportée par l'utilisation de l'anti-dictionnaire n'est pas tellement significative, par contre celle apportée par la combinaison de pseudo-racinisation et l'anti-dictionnaire est plus intéressante, atteignant un gain de 3,45 %. Dans le cas de requêtes longues (figure 6.11), nous constatons que l'utilisation de l'anti-dictionnaire présente un gain de 3,22 % par rapport à une approche ignorant tout prétraitement et 12,47 % par rapport à une approche de pseudo-racinisation, ceci montre l'importance d'appliquer une deuxième étape d'élimination des mots de l'anti-dictionnaire après la phase de pseudo-racinisation pour la langue arabe. Aussi nous remarquons que la combinaison de pseudo-racinisation et l'anti-dictionnaire dans ce cas est plus intéressante et donne un gain de 17,47 %. En comparant les résultats des requêtes courtes et longues, nous remarquons que l'effet de l'utilisation d'un anti-dictionnaire dépend essentiellement du type des requêtes.

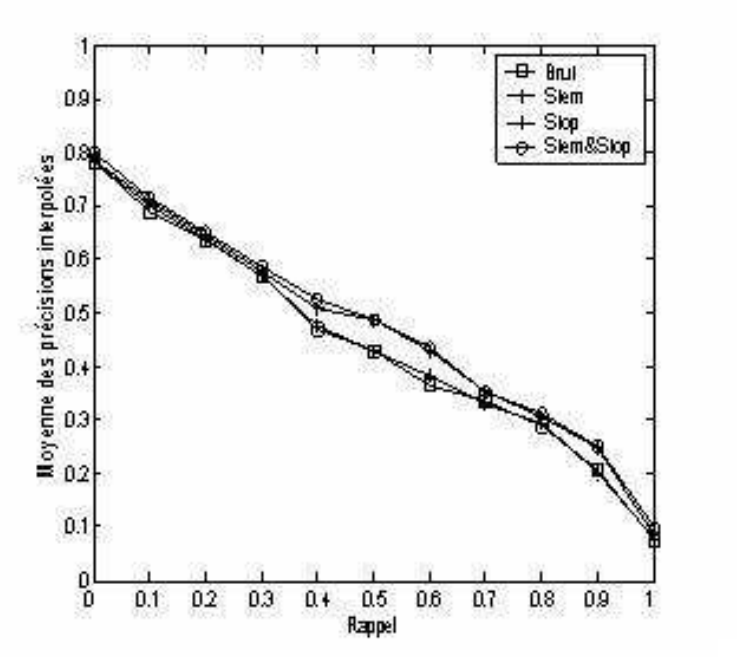

Figure 6.10 – Apport des traitements linguistiques (requêtes courtes, cas sans pondération)

Sur la figure 6.12 (requêtes courtes, cas du pondération Okapi BM-25), les résultats montrent que la performance apportée par la pseudo-racinisation atteint un gain de 4,8 % alors que dans le cas des requêtes longues (figure 6.13), elle présente un gain de 3,5 %. Par contre, l'utilisation d'anti-dictionnaire n'est pas significative pour les deux types de requêtes (courtes et longues).

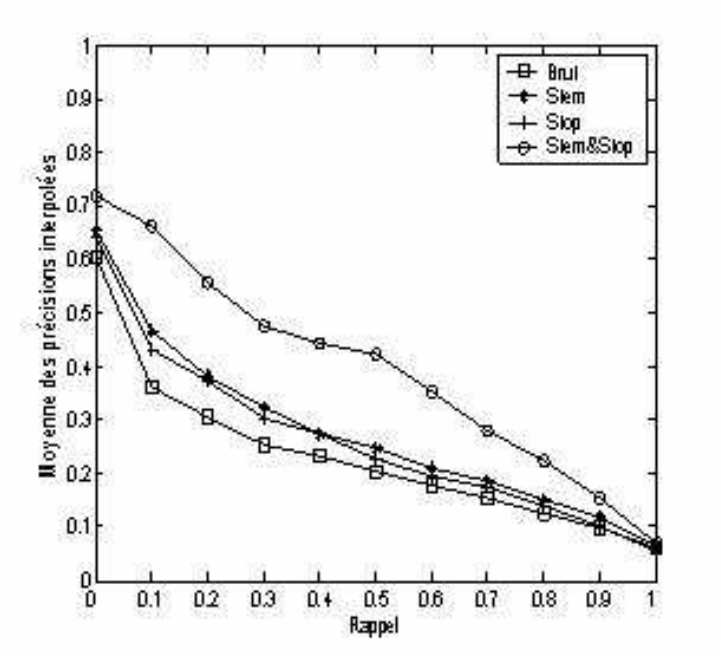

Figure 6.11 – Apport des traitements linguistiques (requêtes longues, cas sans pondération)

Nous remarquons que pour éviter le coût des tests récursifs réalisé pour chaque mot du corpus, afin d'éliminer ceux de l'anti-dictionnaire, nous pouvons utiliser des schémas de pondération comme l'Okapi BM-25, log(tf + 1)xIdf, TfxIdf, Ltc et Tfc, qui minimisent l'effet de ces mots en particulier et l'effet des hautes fréquences en général. Ce qui est confirmé par les résultats de [14] et [8].

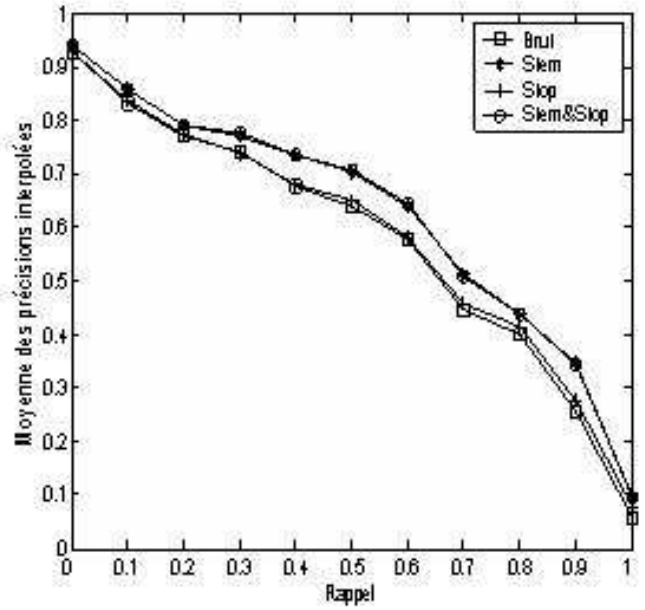

Figure 6.12 – Apport des traitements linguistiques (requêtes courtes, Okapi BM-25)

6.4.2 Syntagmes nominaux

Dans cette partie, nous reproduisons la même expérimentation effectuée dans [21] mais avec l'outil de [35] 4.5.1 d'extraction des SNs. Après l'analyse linguistique et l'extraction des SNs, les documents et les requêtes sont indexés.

- Stratégie 1 : indexer par des unitermes
  nous avons défini la base qui permet d'avoir les meilleures performances et qui consiste à utiliser une racinisation et une pondération Okapi BM-25. Ces performances sont comparées aux résultats obtenus avec l'intégration des syntagmes nominaux pour juger leur impact sur un SRI.
- Stratégie 2 : indexer par des SNs
- Stratégie 3 : indexer les unitermes et les syntagmes nominaux ensemble dans un même vecteur :
  Les syntagmes nominaux extraits sont ajoutés dans les documents et les requêtes comme étant des unitermes simples. Par exemple, si les deux unitermes "pollution" et "atmosphérique" forment un syntagme nominal alors ils sont remplacés par le syntagme nominal "pollution atmosphérique". Les unitermes et les syntagmes nominaux sont utilisés ensemble dans un index.

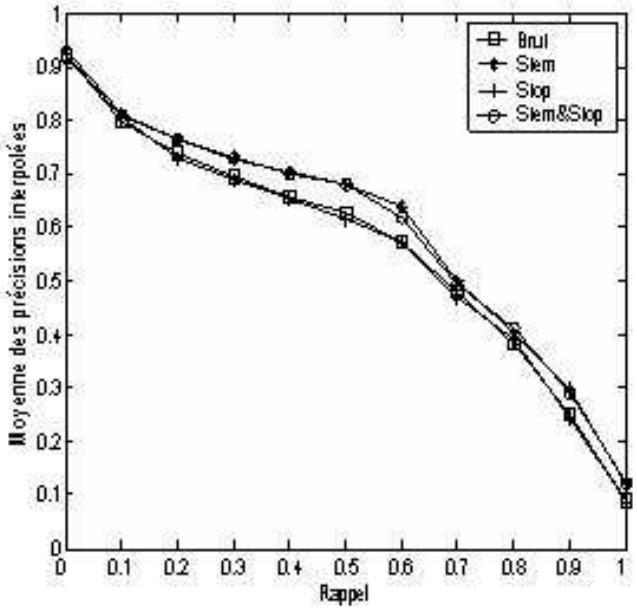

Figure 6.13 – Apport des traitements linguistiques (requêtes longues, Okapi BM-25)

Evaluation Rappel/Précision

En comparant les courbes de rappel et précision, présentées dans la Figure 6.14, nous pouvons constater que la prise en compte des syntagmes nominaux dans l'indexation donne des résultats décevants, certainement liés au taux d'erreurs d'extraction des SNs de l'outil utilisé. Cette degradation peut être expliquée par le manque de normalisation au niveau de la requête par exemple كَارثة تلوث الهوَاء "Lit.catastrophe de la pollution de l'air" et تلوث الهوَاء "pollution de l'air" qui devrait être normalisé sous le syntagme تلوث الهوَاء "pollution de l'air".

6.4.3 Termes complexes

Le besoin de représentation des documents avec des termes complexes s'est réveillé dès la prise de conscience des limites de la représentation avec un terme simple mais les tentatives ont été limitées. A la différence des expérimentations effectuées sur les SNs où l'extracteur utilisé permet d'extraire que la suite "Nom Nom", dans cette section nous étudions l'apport des TCs sur notre SRI en langue arabe, en adoptant notre approche d'extraction des termes complexes détaillée au chapitre 5 où nous avons mené une étude linguistique sur corpus et nous avons identifié trois patrons de base pour l'arabe :

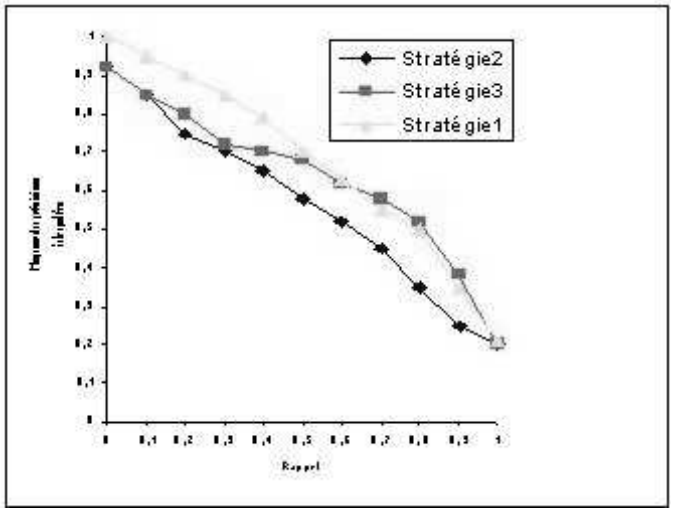

Figure 6.14 – Influence respective des SN(s) et des unitermes sur le SRI

- un nom et un adjectif,
- un nom et un autre nom,
- un nom, une préposition et un autre nom

.

Indexation avec des termes complexes et évaluation

Les évaluations que nous présentons ont été effectuées sur les données de notre collection. Outre l'indexation classique avec des unitermes (UT) où il s'agit de trouver les meilleurs résultats suivant les différents paramètres de pondération, nous avons testé les stratégies d'indexation suivantes :

- la stratégie (appelée Sm) qui permet d'indexer les unitermes et les termes complexes séparément. Pour chaque document ou requête un nouvel index est crée où les termes complexes extraits sont ajoutés. Ces termes complexes sont indexés indépendamment des unitermes. Cela crée deux sous-vecteurs : le premier correspond aux unitermes et le deuxième aux termes complexes.
- la stratégie (appelée Smp) qui emploie un index de termes complexes en pondérant par Okapi BM-25 les termes de la requête pour représenter la présence et l'importance du terme de la requête dans le document donné.

En comparant les taux de rappel et précision de la collection $[AR - ENV]$, nous pouvons constater qu'en intégrant des termes complexes dans l'indexation, nous obtenons de meilleures performances par rapport aux meilleurs résultats obtenus par l'utilisation des unitermes (les performances sont exprimées en terme de précision moyenne (Prec.moy) en 11 points de rappel et en pourcentage de variation par rapport aux performance de l'indexation avec des unitermes (Diff)).
En particulier, la stratégie Smp donne de meilleurs résultats que dans le cas où aucune pondération n'est utilisée. Cette amélioration est perceptible, où la stratégie Sm a permis d'augmenter les performances de 3,6 % alors que cette augmentation est de 5,8 % dans le cas de la stratégie Smp. En examinant les résultats de la Stratégie Sm et de la Stratégie Smp, nous constatons que le nombre de documents pertinents

retrouvés est presque identique pour les deux stratégies. Ce qui diffère est le classement des documents trouvés. En effet, la stratégie Smp permet de favoriser le classement de ces documents en les mettant en tête de la liste des documents trouvés. Ceci se reflète dans les résultats de la précision à faibles taux de rappel (Table 6.4.3) (précision à 5, 10 et 20 documents). Ces résultats, particulièrement la précision à 5 et 10 documents, confirme notre hypothèse que les termes complexes aident à augmenter la précision d'un SRI.

	Prec.moy	Diff
UT	26,1 %	
Sm	29,7 %	+3,6 %
Smp	31,9 %	+ 5,8 %

Table 6.5 – Précision moyenne

	5 doc		10 doc		20 doc	
	Prec.	Diff	Prec.	Diff	Prec.	Diff
UT	51,6 %		43,9 %		40,8 %	
Sm	53,2 %	+1,4 %	45,1 %	+1,2 %	42,1 %	+1,3 %
Smp	58,3 %	+6,7 %	45,8 %	+1,9 %	42,2 %	+1,4 %

Table 6.6 – Précision à 5, 10 and 20 documents

L'intégration des termes complexes dans le processus d'indexation a montré une influence réelle sur les performance d'un SRI. Particulièrement, l'indexation des termes complexes en pondérant les requêtes, a révélé de meilleurs résultats. Ces constatations confirment notre hypothèse avancée que l'utilisation des termes complexes constitue une représentation plus précise du contenu des documents que les unitermes et renforce notre approche d'indexation en considérant les termes complexes comme support des termes d'indexation.

6.5 Conclusion

L'objectif de ces expérimentations était de : mesurer l'apport en RI de multiples connaissances linguistiques, l'impact de l'analyse sémantique latente sur le SRI par rapport au modèle vectoriel, et évaluer la pertinence au sein d'un SRI. Les résultats de nos expérimentations montrent nettement l'influence de la méthode LSA par rapport au modèle vectoriel standard. À partir d'une plate-forme de test conçue pour intégrer au sein d'un SRI diverses représentations linguistiques de documents et requêtes, nous avons évalué et analysé l'apport individuel d'informations linguistiques de différents niveaux de langue pour retrouver des documents pertinents. Nos expérimentations conduisent à un certain nombre de remarques. Du point de vue de leur impact individuel, les résultats obtenus ont montré, l'impact positif et tranché de certaines connaissances linguistiques en particulier morphologiques (plus précisement les racines) et attestent de l'intérêt de recourir à ce type de connaissances en RI. Les résultats obtenus en exploitant des connaissances syntaxiques sont toutefois encourageants. Les expériences menées ont donné des résultats

décevants concernant l'indexation par des SNs, cette dégradation de la performance est certainement liés au taux d'erreurs d'extraction des SNs de l'outil utilisé. En outre, elle peut être expliquée par le manque de normalisation au niveau de la requête par exemple كَارثة تلوث الهوَاء"Lit.catastrophe de la pollution de l'air" et تلوث الهوَاء "pollution de l'air" qui devrait être normalisé sous le syntagme تلوث الهوَاء "pollution de l'air". Le fait que cela n'est pas conforme à ce que nous trouvons dans la littérature [62] [124] [126] qui perçoit une augmentation entre 5 % et 30 % pour les langues européennes, nous a poussé à developper un outil d'extraction de termes basé sur une approche mixte combinant une analyse linguistique qui consiste à définir des patrons syntaxiques et statitistique basée sur des mesures statitistiques (cf. chapitre 5). Ainsi, l'utilisation des termes complexes extraits dans l'indexation des documents a permis d'augmenter les performances du notre SRI. Cette augmentation se présente essentiellement sous la forme d'une augmentation de la précision dans les performances du SRI atteignant 5,8 %. En effet, les résultats de nos expérimentations ont montré que l'utilisation des termes complexes dans l'indexation permet de classer les documents pertinents trouvés dans les premiers rangs de la liste des documents trouvés. Ce taux de 5,8 % faible par rapport aux langues européennes peut être expliqué par la nature agglutinante de la langue arabe ce qui se répercute sur la segmentation des unités lexicales et par conséquent sur l'étiquetage de ces dernières. En outre, cette différence d'augmentation est dûe aussi à la nature des requêtes utilisées, la plupart des travaux utilisent des requêtes longues par contre dans nos travaux nous avons recours à des requêtes courtes fournissant que le titre.

CHAPITRE 7
Conclusion et perspectives

La Recherche d'Information a pour objectif de fournir à un utilisateur un accès facile à l'information qui l'intéresse, cette information étant située dans une masse de documents textuels. Afin d'atteindre cet objectif, un système de recherche d'information doit représenter, stocker et organiser l'information puis fournir à l'utilisateur les éléments correspondant au besoin d'information exprimé par sa requête. Notre thèse s'inscrit dans le cadre de la recherche d'information dans un domaine de spécialité en langue arabe. Ainsi un SRI en langue arabe doit prendre en considération ses caractéristiques singulières et proposer des outils et des techniques automatiques afin de permettre son traitement informatique. L'objectif de notre travail a été d'une part, d'identifier les termes complexes présents dans les requêtes et les documents. D'autre part, d'exploiter pleinement la richesse de la langue en combinant plusieurs connaissances linguistiques appartenant aux niveaux morphologique et syntaxique, et de montrer comment l'apport de connaissances morphologiques et syntaxiques permet d'améliorer l'accès à l'information.

7.1 Identification des termes complexes et ses variantes

Nous avons présenté une étude linguistique sur les termes complexes du domaine de l'environnement. Nous avons réalisé une typologie des termes complexes en langue arabe où nous avons précisé quels types élémentaires de termes complexes sont effectivement présents dans ce domaine technique. Ensuite, nous avons examiné les différentes structures morphosyntaxiques des groupes nominaux complexes arabes et nous avons établis leur classement en fonction de leur structure morphosyntaxique. L'identification des termes complexes se présente en deux parties : la première consacrée à la typologie des structures élémentaires, la deuxième aux variations qui peuvent affecter les termes complexes. La méthodologie que nous avons adoptée est la suivante : nous avons examiné le corpus et extrait des suites d'unités lexicales susceptibles d'apparaître dans des positions syntaxiques variées et qui appartiennent à l'un des types élémentaires décrits par [26] . Les termes complexes épousent des structures morphosyntaxiques exprimées en partie de discours. Pour vérifier le caractère terminologique du candidat terme extrait, deux solutions sont envisagées : la première consiste à utiliser la base terminologique AGROVOC [1] pour attester le candidat terme extrait. Si ce dernier n'existe pas dans la base terminologique, une deuxième solution est envisagée qui consiste à chercher la traduction de ces suites d'unités lexicales en exploitant la propriété de compositionalité des sens des termes complexes. Ainsi, nous avons tiré profit de leur traduction française pour vérifier leur statut terminologique dans la banque terminologique Eurodicautom.
Cette liste de candidats est soumise à diverses mesures statistiques. Ces calculs permetteront de déterminer le statut de la séquence rencontrée. En conclusion de l'examen des mesures statistiques qui ont été

[1] www.fao.org/agrovoc/

presentées, nous avons remarqué que la LLR d'un candidat terme est un très bon révélateur de son caractère terminologique et cela pour : sa nature de test statistique, sa tendance à prendre en considération la fréquence du candidat terme, son bon comportement en dépit de la taille du corpus.
Pour être complet, ce travail d'extraction doit indiquer pour chaque terme complexe ses variantes. En ce qui concerne les variantes graphiques, le lien est assez aisé : des procédures de normalisation devraient nous permettre de relier par exemple « التلوث الهوَاء» et « التلوث الهوَاءي» « pollution atmosphérique ».
Pour les variantes morphosyntaxiques la tâche est moins triviale, nous avons effectué une analyse morphologique fondée sur des règles de dérivation générales et en créant des règles de désuffixation/recodage (دراَسَات/دراَسة études/étude) en ajoutant le suffixe "ات" ou "ون" aux noms, pour désigner le pluriel régulier féminin respectivement le pluriel régulier masculin. Même si cela conduit à des bon résultats, il est dommage que nous ne puissions pas résoudre les problèmes liés au pluriel irrégulier où les noms suivent de nombreuses règles complexes. Pour la détection des variantes syntaxiques, Nous avons plebiscité l'analyse par patrons qui décrit exhaustivement les structures linguistiques recherchées par rapport à une analyse par frontière qui génère plus de bruit sans véritablement gagner en couverture. Les variantes paradigmatiques n'étaient pas prévues car il nous semble difficile de relier un terme complexe à ses variantes paradigmatiques, vu que ces dernières nécessitent le recours à des ressources lexicales monolingues qui sont en cours de développement pour l'arabe.

7.2 Evaluation des traitements linguistiques en recherche d'information

Afin de valider notre hypothèse à savoir que l'intégration des traitements linguistiques dans un processus de recherche d'information permet de produire une indexation de meilleure qualité. Ainsi, nous avons proposé une plate-forme intégrant plusieurs connaissances linguistiques appartenant aux niveaux morphologique et syntaxique. Pour le niveau morphologique, nous avons choisi de considérer deux types de connaissances en particulier. Nous prenons en compte des connaissances morphologiques flexionnelles et dérivationnelles. Nous avons appliqué aux documents (et requêtes) une procédure de racinisation (stemming) qui permet d'extraire pour chaque unité lexicale sa pseudo-racine (stem). Nous nous sommes appuyées sur l'algorithme de Darwish [31] pour la normalisation des variantes morphologiques.
Le dernier type de connaissance utilisé est d'ordre morpho-syntaxique. Une analyse morpho-syntaxique des documents et requêtes est effectuée dont l'objectif est d'associer à chaque unité lexicale sa catégorie grammaticale (nom, verbe, adjectif...). Le principal intérêt de cet étiquetage est qu'il permet d'opérer un premier traitement de désambiguïsation des termes. L'étiqueteur utilisé ne peut ainsi associer qu'une seule étiquette à chaque unité lexicale; il doit par conséquent choisir parmi toutes les catégories possibles d'une unité lexicale celle qui correspond au terme dans la phrase considérée, en s'appuyant sur son contexte d'apparition. Nous avons sélectionné l'analyseur morpho-syntaxique de [35] basé sur un apprentissage supervisé et qui donne en sortie une collection de textes étiquetés où toutes les unités lexicales ont été catégorisées.
Pour le niveau syntaxique, nous avons retenu deux principales structures permettant de rendre compte des relations et dépendances entre les unités lexicales. Nous prenons en compte tout d'abord les termes complexes (TCs). Nous avons proposé aussi d'exploiter les syntagmes nominaux présents dans les textes et les requêtes. Nous utilisons pour cela l'outil développé par Diab [35] qui, à l'aide d'une méthode combinant à la fois des techniques numériques et symboliques d'acquisition [2], identifie au sein des textes les syntagmes nominaux (SNs).

[2]La technique utilisée s'appuie plus précisément sur l'utilisation d'un algorithme d'apprentissage de règles de transformation à partir de corpus étiquetés morpho-syntaxiquement.

Nous avons évalué et analysé l'apport individuel d'informations linguistiques de différents niveaux de langue pour retrouver des documents pertinents. les résultats obtenus ont montré l'impact positif et tranché des connaissances morphologiques (plus précisement les racines) et attestent de l'intérêt de recourir à ce type de connaissances en RI. Les résultats obtenus en exploitant des connaissances syntaxiques sont encourageants. Les expériences menées ont donné des résultats décevants concernant l'indexation par des SNs, par contre les expériences présentées sur les TCs, nous ont permis de montrer l'efficacité de notre méthodologie d'extraction des TCs sur notre collection de test. L'analyse linguistique de surface de [35] s'est montrée suffisante pour notre besoin et efficace en termes de temps de traitement. L'utilisation des termes complexes extraits dans l'indexation des documents a permis d'augmenter les performances du SRI d'un gain de 5,8 % sur un jeu de requêtes de 30. Ainsi, les différentes expériences permettent de confirmer notre hypothèse que les termes complexes constituent une bonne représentation du contenu textuel, aussi le fait d'introduire des traitements linguistiques au niveau morphologique ou syntaxique permet une nette amélioration des performances du SRI en langue arabe. A savoir que sur un jeu de requêtes de 30 nous avons remarqué une augmentation des performances atteignant les 17,47 % en intégrant des connaissances linguistiques au niveau morphologique et 5,8 % au niveau syntaxique par rapport au cas où aucun traitement n'est appliqué. La thèse présente aussi des contributions secondaires. Nous avons montré que le modèle d'analyse sémantique latente (LSA) permet d'améliorer les performances d'un SRI en langue arabe par rapport à un modèle vectoriel. Les résultats montrent une différence s'avérant statistiquement significative de 15,9% pour le modèle LSA par rapport au modèle MVS, dans le cas des requêtes courtes alors qu'ils présentent un gain de 16,3% pour les requêtes longues.

7.3 Perspectives

Les différentes pistes explorées dans le cadre de cette thèse nous ont amenée à envisager de nombreuses perspectives. Nous présentons ici celles qui nous paraissent les plus prometteuses.

Emploi en extension de requêtes

Lorsque nous avons étudié l'intérêt de combiner des informations linguistiques en RI, nous avons observé des différences importantes entre les connaissances quant à leur efficacité respective à améliorer les performances des systèmes. Une perspective est de proposer de nouvelles expérimentations en exploitant des connaissances plus pertinentes d'un point de vue linguistique. En effet, ces expériences nous permettrait d'évaluer précisément si les résultats obtenus sont liés à la qualité des représentations proposées ou à la façon de les utiliser en RI. Parmi les informations susceptibles de mieux représenter le contenu textuel des documents et requêtes, nous pensons plus particulièrement à des informations sémantiques acquises en corpus telles que le recours aux variantes morphologiques d'ordre dérivationnel (comme par exemple pollution, polluants) qui pourraient être exploités lors de l'expansion de requêtes.

Couplage des connaissances linguistiques

L'exploitation d'une seule information de niveau morphologique ou syntaxique pour l'amélioration des performances des SRI ne permet d'exploiter la richesse de la langue que partiellement. Il serait intéressant d'évaluer l'intérêt en RI de coupler et d'intégrer ensemble au coeur d'un même système des connaissances appartenant à tous les niveaux de la langue. Selon cette approche, chaque document (ou requête) n'est plus associé à un seul type de descripteur mais à une combinaison de diverses représentations (chacune correspondant à un type d'information linguistique particulier).

Emploi aux corpus multilingues

L'emploi de notre approche dans le cadre de la recherche d'information multilingue nous semble prometteuse. Cet emploi consiste à interroger un corpus multilingue pour rechercher des documents écrits dans des langues différentes par le biais d'une unique requête. Cette perspective consiste à apparaître dans plusieurs langues des termes complexes dans le but que la recherche multilingue se fasse au niveau des termes d'indexation et non plus par traduction des requêtes.

Applications des modèles de langues

Dans l'ensemble des travaux, nous avons remarqué que si la convergence entre le TAL et la RI ne donnait pas des résultats plus tranchés, cela était dû principalement au fait que les traitements linguistiques n'étaient pas adaptés au domaine de la RI. Le problème serait de considérer que ce sont les mécanismes de représentation des contenus textuels et de mise en correspondance tels qu'ils sont proposés par la RI qui ne pas assez souples pour exploiter pleinement la richesse des informations linguistiques. En effet, comme nous l'avons remarqué les modèles vectoriels de RI sont, pour beaucoup, limités dans leur façon de représenter le contenu des documents et requêtes. Ces représentations (en "sac de mots" pour modèle vectoriel) ne prennent en compte ni les dépendances entre les unités lexicales ni leur ordre et sont ainsi peu adaptées à accueillir des informations plus riches que des unités lexicales simples. Une perspective prometteuse à notre travail, serait d'étudier l'apport de notre approche sur d'autres modèles de RI plus aptes à prendre en compte nos représentations enrichies. Nous pensons plus particulièrement aux modèles de langue, qui proposent des solutions prometteuses pour l'intégration de connaissances linguistiques.

Bibliographie

[1] R. ABBES. *Les réseaux collocationnels dans la construction et l'exploitation d'un corpus dans le cadre d'une communauté de discours scientifique*. Thèse de Doctorat, Université de Nantes, Nantes, France, 1999.

[2] R. ABBES. *La conception et la réalisation de concordancier électronique pour l'arabe*. Thèse de Doctorat, Institut national des sciences appliquées de Lyon, Lyon, France, 2004.

[3] S. ABNEY. Parsing by chunks. In Robert C. BERWICK, Steven P. ABNEY et Carol TENNY, réds., *Principle-Based Parsing: Computation and Psycholinguistics*, pages 257–278. Kluwer, Dordrecht, 1991.

[4] O. ALJLAYL, M.AND FRIEDER. On arabic search: Improving the retrieval effectiveness via a light stemming approach. In *11 the International Conference on Information and Knowledge Management (CIKM)*, pages 340–347, Virginia, USA, 2002.

[5] A. ARAMPATZIS, T. TSORIS et C. H. KOSTER. irena : Information retrieval engine based on natural language analysis. Rapport technique, Computing Science Institute, Nijmegen, Pays-Bas, 1996.

[6] A. ARAMPATZIS, T. Van der WEIDE, C.H. KOSTER et P. VAN BOMMEL. Linguistically motivated information retrieval. *Encyclopedia of Library and Information Science*, pages 201–222, 2000.

[7] A.R. ARONSON. Effective mapping of biomedical text to the umls metathesaurus : the metamap program. In *Proceedings of the American Medical Informatics Association Symposium (AMIA 2001)*, pages 17–21, Washington, DC, 2001.

[8] F. ATAA-ALLAH, S. BOULAKNADEL, D. ABOUTAJDINE et A. ELQADI. Evaluation de l'analyse sémantique latente et du modèle vectoriel standard appliqués à la langue arabe. *TSI, à apparaître*, 2007.

[9] F. ATAA ALLAH, S. BOULAKNADEL, A. EL QADI et D. ABOUTAJDINE. Amélioration de la performance de l'analyse sémantique latente pour des corpus de petite taille. *Revue des Nouvelles technologies de l'Information (RNTI)*, 1(1):317, 2005.

[10] S. BALOUL et P.B. de MAREUIL. Un modèle syntactico-prosodique pour la synthèse de la parole à partir du texte en arabe standard voyellé. In *7th Maghrebian Conference on Computer Sciences*, Annaba, 2002.

[11] E. BENVENISTE. Formes nouvelles de la composition nominale. In *Problèmes de linguistique générale*, pages 163–173. Gallimard, Paris, 1966.

[12] R. BESANÇON. *Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles des textes, Application au calcul de similarité sémantique dans le cadre du modèle DSIR*. Thèse de Doctorat, Ecole polytechnique fédérale de lausanne, Suisse, 2001.

[13] R. BLACHERE et M. GAUDEFROY-DEMOMBYNES. *Grammaire de l'arabe classique*. édition Maisonneuve et Larose, Paris, 1994.

[14] F. BOULAKNADEL, S.AND ATAA ALLAH. Recherche d'information en langue arabe : influence des paramètres linguistiques et de pondération de lsa. In *Actes des Rencontres des Etudiants*

Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL), pages 643–648, Dourdan, France, 2005.

[15] D. BOURIGAULT. Analyse syntaxique locale pour le repérage de termes complexes dans un texte. *Traitement Automatique des Langues (TAL)*, 34:105–117, 1993.

[16] C. BUCKLEY, G. SALTON et J. ALLAN. Automatic retrieval with locality information using smart. In *TREC*, pages 59–72, 1992.

[17] M.T. CABRÉ. *La terminologie, théorie, méthodes et applications*. Armand Colin, Paris, 1998.

[18] G. CARBONELL, J., Y. YANG, E. FREDERKING, R., D. BROWN, R., Y. GENG et D. LEE. Translingual information retrieval : A comparative evaluation. In *In Proceedings of the 15th International Joint Conference on Artificial Intelligence, IJCAI'97*, Nagoya, Japon, 1997.

[19] M. CARL, J. HALLER, C. HORSHMANN, D. MAAS et J. SHUTZ. The tetris terminology tools. *Traitement automatique des langues*, 43(1):73–102, 2002.

[20] H. CHERFI et Y. TOUSSAINT. Adéquation d'indices statistiques à l'interprétation de régles d'association. In *Actes des 6èmes journées internationales d'analyses statistiques des données textuelles (JADT 02)*, pages 233–244, San Malo, France, 2002.

[21] J. P CHEVALLET, M. GÉRY et H. HADDAD. Campagne de tests amaryllis expérimentations et résultats. In *Atelier final de la campagne Amaryllis II*, Paris, France, 2000.

[22] J.P. CHEVALLET. *Un modèle logique de recherche d'information appliqué au formalisme des graphes conceptuels. Le prototype elen et son expérimentation sur un corpus de composants logiciels*. Thèse de Doctorat, Université Joseph Fourier, Grenoble, France, 1992.

[23] K.W. CHURCH et P. HANKS. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th. Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, B.C., 1990. Association for Computational Linguistics.

[24] V. CLAVEAU. *Acquisition automatique de lexiques sémantiques pour la recherche d'information*. Thèse de Doctorat, Université de Rennes I, Rennes, France, 2003.

[25] V. CLAVEAU et P. SÉBILLOT. Extension de requêtes par lien sémantique nomverbe acquis sur corpus. In *In Proceedings of 11ème conférence annuelle sur le traitement automatique des langues naturelles (TALN)*, Fez, Maroc, 2004.

[26] B. DAILLE. *Approche mixte pour l'extraction de terminologie : statistiques lexicales et filtres linguistiques*. Thèse de Doctorat, Université de Paris 7, France, 1994.

[27] B. DAILLE. *Découvertes linguistiques en corpus, Mémoire d'Habilitation à Diriger des Recherches en Informatique*. Thèse de Doctorat, Université de Nantes, France, 2002.

[28] B. DAILLE. Conceptual structuring through term variations. In *Proceedings ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 9–16, Sapporo, Japan, 2003.

[29] B. DAILLE. Variations and application-oriented terminology engineering. *International journal of theoretical and applied issues in specialized communication*, 11(1):181–197, 2005.

[30] B. DAILLE, E. GAUSSIER et J.M. LANGÉ. An evaluation of statistical scores for word association. In *The Tblisi Symposium on Logic, Language and Computation*, pages 177–188. CSLI Publications, 1998.

[31] K. DARWISH. Building a shallow arabic morphological analyzer in one day. In *Proceedings of the workshop on Computational Approaches to Semitic Languages in the 40th Annual Meeting of the Association for Computational Linguistics*, pages 47–54, Philadelphia, USA, 2002.

[32] S. DAVID et P. PLANTE. De la nécessité d'une approche morpho-syntaxique en analyse de textes. *intelligence artificielle et sciences cognitives au Québec*, 2(3):140–155, 1990.

[33] S. DEERWESTER, S.T. DUMAIS, G.W. FURNAS, T.K. LANDAUER et R. HARSHMAN. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[34] A. DEPECKER. *Entre signe et concept*. Presses Sorbonne Nouvelle, France, 2002.

[35] M. DIAB, K. HACIOGLU et D. JURAFSKY. Automatic tagging of arabic text: From raw text to base phrase chunks. In *In Proceedings of NAACL-HLT*, pages 149–152, Boston, USA, 2004.

[36] J. DICHY. Pour une lexicomatique de l'arabe : l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot. *Meta*, XLII, 2:291–306, 1997.

[37] J. DICHY, A.F. BRAHAM et S. GHAZALI. La base de connaissances linguistiques dinaar1. In *Colloque international sur le traitement automatique de l'arabe*, pages 45–56, Manouba, Tunisie, 2002.

[38] G.M. DILLON et A.S. GRAY. fasit : A fully automatic syntactically based indexing system. *Journal of the American Society for Information Science*, 34(2):99–108, 1983.

[39] E. DITTERS. The description of modern standard arabic syntax in terms of functions and categories. *Langues et Littératures du Monde Arabe*, 2:115–151, 2001.

[40] J. DOWDALL, F. RINALDI, F. IBEKWE-SANJUAN et E. SANJUAN. Complex structuring of term variants for question answering. In *Proceedings ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 1–8, Sapporo, Japan, 2003.

[41] S.T. DUMAIS, T.K. LANDAUER et M. LITTMANN. Automatic cross linguistic information retrieval using latent semantic indexing. In *In Proceedings of the SIGIR Workshop on Cross Linguistic Information Retrieval*, Zurich, Suisse, 1996.

[42] T. DUNNING. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1994.

[43] D. EL KASSAS. *Etude contrastive de l'arabe et du français dans une perspective de génération multilingue*. Thèse de Doctorat, Université Denis Diderot - Paris VII, Paris, France, 2005.

[44] T. EL KOURY. Les emprunts médiatiques : De l'actualité à l'implantation. *Al Kimiya, annales de l'Institut de Langue et Traduction*, 10:15–35, 2004.

[45] J. FAGAN. *Experiments in Automatic Phrase Indexing for Document Retrieval : A Comparison of Syntactic and Non-Syntactic Methods*. Thèse de Doctorat, Université de Cornell, New-York, États-Unis, 1987.

[46] E. FAULSTICH. Principes formels et fonctionnels de la variation en terminologie. *Terminology*, 5(1):93–106, 1999.

[47] C. FELLBAUM. *wordnet : An Electronic Lexical Database*. The MIT Press, Cambridge, Massachussets, États-Unis, 1998.

[48] E.A. FOX. *Extending the Boolean and Vector Space Models of Information Retrieval with P-norm Queries and Multiple Concept Types*. Thèse de Doctorat, Université de Cornell, New-York, États-Unis, 1983.

[49] K. T. FRANTZI, S. ANANIADOU et H. MIMA. Automatic recognition of multi-word terms: the c-value/nc-value method. *Int. J. on Digital Libraries*, 3(2):115–130, 2000.

[50] J. FREIXA. Causes of denominative variation in terminology : A typology proposal. *International journal of theoretical and applied issues in specialized communication*, 12(1):51–77, 2006.

[51] M. FULLER et J. ZOBEL. Conflation-based comparison of stemming algorithms. In *In Proceedings of the 3th Australian Document Computing Symposium*, Sidney, Australie, 1998.

[52] E. GAUSSIER. Unsupervised learning of derivational morphology from inflectional corpora. In *Proceedings of Workshop on Unsupervised Methods in Natural Language Learning, 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, Maryland, États-Unis, 1999.

[53] E. GAUSSIER, G. GREFENSTETTE, D. HULL et R. ROUX. Recherche d'information en français et traitement automatique des langues. *Traitement automatique des langues*, 41(2):473–493, 2000.

[54] E. GAUSSIER, G. GREFENSTETTE et M. SCHULZE. Traitement du langage naturel et recherche d'informations : quelques expériences sur le français. In *Proceedings of 1ères journées scientifiques et techniques du réseau francophone de l'ingénierie de la langue de l'AUPELF-UREF*, pages 33–45, Avignon, France, 1997.

[55] F. GEY et D.W. OARD. The trec-2001 cross-language information retrieval track: Searching arabic using english, french or arabic queries. In *TREC*, pages 16–26, 2001.

[56] M. GHENIMA. *Système de voyellation de textes arabes*. Thèse de Doctorat, Université Lyon2, Lyon, France, 1998.

[57] G. GREFENSTETTE. sqlet : Short query linguistic expansion techniques : Palliating one-word queries by providing intermediate structure to text. In *Proceedings of 5ème conférence internationale sur la recherche d'informations assistée par ordinateur (RIAO)*, Montréal, Canada, 1997.

[58] G. GROSS. *Les expressions figées en français, noms composés et autres locutions*. Collection l'essentiel français, Ophrys, Paris, 1996.

[59] M. GUIDÈRE. *Arabe : grammaticalement correct! : grammaire alphabétique de l'arabe*. Ellipse, Paris, 2001.

[60] B. HABERT et Ch. JACQUEMIN. Noms composés, termes dénominations complexes : problématique linguistiques et traitements automatiques. *Traitement Automatique des Langues (TAL)*, 34(2):5–41, 1993.

[61] H. HADDAD. *Extraction et impact des connaissances sur les performances des systèmes de recherche d'information*. Thèse de Doctorat, Université Joseph Fourier, Grenoble, France, 2002.

[62] H. HADDAD. Utilisation des syntagmes nominaux dans un système de recherche d'information. In *Proceedings of 19èmes journées de bases de données avancées (BDA)*, Lyon, France, 2003.

[63] T. HAMON. *Variation sémantique en corpus spécialisé : Acquisition de relation de synonymie à partir de resources lexicales*. Thèse de Doctorat, Université Paris Nord, France, 2000.

[64] T. HAMON et A. NAZARENKO. Detection of synonymy link between terms: Experiment and results. In *Recent Advances in Computational Terminology*, pages 185–208. John Benjamins, 2001.

[65] D. HARMAN. How effective is suffixing ? *Journal of the American Society for Information Science*, 42(1):7–15, 1991.

[66] M. HASSOUN. *Conception d'un dictionnaire pour le traitement automatique de l'arabe dans différents contextes d'application*. Thèse de Doctorat, Université Lyon1, Lyon, France, 1987.

[67] J. A. HAYWOOD et H. M. NAHMAD. *A new Arabic grammar.* Percy Lund Humphries Publishers Ltd., London, 1962.

[68] D. HULL, G. GREFENSTETTE, B.M. SCHULZE, H. SCHÜTZE et J.O. PEDERSEN. Xerox trec-5 site report : Routing, filtering, nlp and spanish tracks. In *In Proceedings of the 5th Internation Conference on Text Retrieval (TREC)*, Gaithersburg, États-Unis, 1997.

[69] C. JACQUEMIN. État de l'art sur l'analyse des noms composés et des termes. Rapport technique 89, Institut de Recherche en Informatique de Nantes, Nantes, 1995.

[70] C. JACQUEMIN. Guessing morphology from terms and corpora. In *20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 156–165, Philadelphia, PA, 1997.

[71] C. JACQUEMIN. Recognition and acquisition : two inter- related activities in corpus-based term extraction. *Terminology*, 40(2):245–274, 1999.

[72] C. JACQUEMIN. *Spotting and Discovering Terms through Natural Language Processing Techniques*. MIT Press, Cambridge, 2001.

[73] C. JACQUEMIN et E. TZOUKERMANN. Nlp for term variant extraction: A synergy of morphology, lexicon, and syntax. In *Natural Language Information Retrieval,*, pages 25–74. Kluwer, 1999.

[74] M.P. JACQUES. *Approche en discours de la réduction des termes complexes dans les texts spécialisés*. Thèse de Doctorat, Université de Toulouse II, France, 2003.

[75] K. KAGEURA. Theories of terminology : A quest for a framework for the study of term formation. *Terminology*, 5(1):21–40, 1999.

[76] S. KHOJA. Apt: Arabic part-of-speech tagger. In *Proceedings of the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 81–86, Carnegie Mellon University, Pittsburgh, 2001.

[77] S. KHOJA et G. GARSUDE, R.AND KNOWLES. A tagset for the morphosyntactic tagging of arabic. In *Corpus Linguistics 2001 conference*, pages 1–13, Lancaster, UK, 2001.

[78] A. KILGARRIFF et M. PALMER. Special issue on senseval. *Computers and the Humanities*, 34(1/2), 2000.

[79] G. A. KIRAZ. Analysis of the arabic broken plural and diminutive. In *Proceedings of the 5th International Conference and Exhibition on Multilingual Computing (ICEMCO96)*, Cambridge, UK, 1996.

[80] A. KOCOUREK. *Lexical Phrases in Terminology*. Travaux de terminologie, Québec, 1979.

[81] W. KRAAIJ et R. POHLMANN. Viewing stemming as recall enhancement. In *Proceedings of the 19th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Zurich, Suisse, 1996.

[82] R. KROVETZ. Viewing morphology as an inference process. In *In Proceedings of the 16th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Pittsburgh, USA, 1993.

[83] T. K. LANDAUER, D. LAHAM, B. REHDER et M. E. SCHREINER. How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. In *M. G. Shafto and P. Langley (Eds.) Proceedings of the 19th annual meeting of the Cognitive Science Society*, pages 412–417, Mawhwah, Erlbaum, 1997.

[84] E. LAPORTE. Mot et niveau lexical. *Ingénierie des langues*, pages 25–46, 2000.

[85] L.S. LARKEY, L. BALLESTEROS et M.E. CONNELL. Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275 –282, Tampere, Finland, 2002.

[86] L. LEBART et A. SALEM. *Statistique textuelle*. Dunod, 1994.

[87] L. J. LEE. *Similarity Based Approaches to Natural Language Processing*. Thèse de Doctorat, Harvard University, USA, 1997.

[88] M. LENNON, D.S PIERCE, B.D. TARRY et P. WILLETT. An evaluation of some conflation algorithms for information retrieval. *Journal of Information Science*, 3(1):177–183, 1981.

[89] D.D. LEWIS. An evaluation of phrasal and clustered representations on a text categorization task. In *In Proceedings of the 15th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Copenhague, Danemark, 1992.

[90] D.D. LEWIS et W.B. CROFT. Term clustering of syntactic phrases. In *Proceedings of the 13th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Bruxelles, Belgique, 1990.

[91] R. MARTINET. *Eléments de linguistique générale*. Armand Colin, Paris, 1991.

[92] G. MAYNARD et S. ANANIADOU. Identifying terms by their family and friends. In *Proceedings of the 18th conference on Computational linguistics*, pages 530 – 536, Saarbrücken, Germany, 2000.

[93] I. MEL'CUK, A. CLAS et A. POLGUERE. *Introduction à la lexicologie explicative et combinatoire*. Duculot, Louvain-la-neuve, 1995.

[94] R. MIHALCEAN et D.I. MOLDOVAN. Semantic indexing using wordnet senses. In *Proceedings of Workshop on IR and NLP, 38th Annual Meeting of the Association for Computational Linguistics (ACL)*, Hong-Kong, Chine, 2000.

[95] F. MOREAU et P. SEBILLOT. Contributions des techniques du traitement automatique des langues à la recherche d'information. Rapport technique 1690, Rapport de Recherche IRISA, Rennes, 2005.

[96] E. MORIN. Des patrons lexico-syntaxiques pour aider au dépouillement terminologique. *Traitement Automatique des Langues*, 40(1):143–166, 1999.

[97] H. NAKAGAWA et T. MORI. Automatic term recognition based on statistics of compound nouns and their components. *Terminology*, 9(2):201–219, 2003.

[98] B. NKWENTI AZEH. Positional and combinational characteristics of terms. *International journal of theoretical and applied issues in specialized communication*, 1(1):61–95, 1994.

[99] M. NOALLY. *Le substantif épithète*. Presses de l'Université de France, Paris, 1990.

[100] G. OTMAN. Pourquoi parler de connaissances terminologiques et de bases de connaissances terminologiques. In *La Banque des Mots*, pages 5–27, Paris, France, 1994.

[101] J. PEARSON. *Terms in Context*. John Benjamins, Amsterdam, 1998.

[102] H.J. PEAT et P. WILLETT. The limitations of term cooccurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42(5):378–383, 1991.

[103] M.P PERY-WOODLEY. *Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle*. Thèse de Doctorat, Université Toulouse Le Mirail, Paris, France, 2000.

[104] B. PINCEMIN. Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative. In *Atelier Corpus et TAL : pour une réflexion méthodologique, Conférence TALN 99*, pages 26–36, 1999.

[105] B. PIWOWARSKY. *Techniques d'apprentissage pour le traitement d'informations structurées : Application à la recherche d'information*. Thèse de Doctorat, Université de Paris VI, France, 2003.

[106] Y. QIU et H.P. FREI. Improving the retrieval effectiveness by a similarity thesaurus. Rapport technique, Rapport interne, Department of Computer Science, ETH Zürich, Zürich, Suisse,, 1995.

[107] M. RAJMAN, R. BESANÇON et J.C. CHAPPELIER. Le modèle dsir : une approche à base de sémantique distributionnelle pour la recherche documentaire. *Traitement automatique des langues*, 41(2):549–578, 2000.

[108] M. RAJMAN et L. LEBART. Similarités pour données textuelles. In *Actes des 4es Journées internationales d'analyse des données textuelles, JADT'98*, Nice, France, 1998.

[109] A. REY. *La terminologie : noms et notions, Collection Que sais-je ?* 2e edn, Presses Universitaires de France, France, 1992.

[110] F. RINALDI, J. DOWDAL, M. HESS, K. KALJURAND, M. KOIT et N. KAHUSK. Terminology as knowledge in answer extraction. In *Proceedings of the 6th International Conference on Terminology and Knowledge Engineering (TKE- 2002)*, pages 107–112, Nancy, France, 2002.

[111] S.E. ROBERTSON. The probability ranking principle in information retrieval. *Journal of Documentation*, 33:294–304, 1977.

[112] L. ROMARY. Outils d'accés à des ressources linguistiques. *Ingénierie des langues*, pages 193–212, 2000.

[113] J. ROYAUTÉ. *Les groupes nominaux complexes et leurs propriétés : application à l'analyse de l'information*. Thèse de Doctorat, Université Henri Poincaré - Nancy 1, France, 1999.

[114] G. SALTON. The state of retrieval system evaluation. *Information Processing and Management*, 28(4), 1992.

[115] G. SALTON, A. WONG et C. S. YANG. A vector space model for automatic indexing. *Commun of the ACM*, 18(11):613–620, 1975.

[116] J. SAVOY. Morphologie et recherche d'information. Rapport technique, Institut interfacultaire d'informatique, Université de Neuchâtel, 2002.

[117] H. SCHMID. Probabilistic part-of-speech tagging using decision trees. *In Daniel Jones and Harold Somers, editors, New Methods in Language Processing,*, pages 154–164, 1997.

[118] A. SHMIDT-WIGGER. Building consistent terminologies. In *COLING-ACL 98*, Montréal, Canada, 1998.

[119] A. F. SMEATON. Using NLP or NLP resources for information retrieval tasks. In *Natural language information retrieval*, pages 99–111. Kluwer Academic Publishers, Dordrecht, NL, 1999.

[120] J.F. SOWA. *Conceptual Structures : Information Processing in Mind and Machine*. Addison-Wesley, USA, 1984.

[121] K. SPARCK JONES. What is the role of nlp in text retrieval? In *Natural language information retrieval*, pages 1–24. Kluwer Academic Publishers, Dordrecht, NL, 1999.

[122] K. SPARK-JONES et J.I. TAIT. Automatic search term variant generation. *Journal of Documentation*, 40(1):50–66, 1984.

[123] T. STRZALKOWSKI, F. LIN, J. WANG et J. PEREZ-CARBALLO. Evaluating natural language processing techniques in information retrieval. In *Natural Language Information Retrieval*, pages 113–145. Kluwer, Boston, MA, 1999.

[124] T. STRZALKOWSKI, F. LIN, J. WANG et J. PEREZ-CARBALLO. Evaluating natural language processing techniques in information retrieval. *In T. Strzalkowski, editor, Natural Language Information Retrieval*, pages 113–145, 1999.

[125] C. J. VAN RIJSBERGEN. *Information Retrieval*. Butterworths, USA, 1979.

[126] J. VILARES-FERRO, F.M. BARCALA et M.A. ALONSO. Using syntactic dependency-pairs conflation to improve retrieval performance in spanish. In *Proceedings of the 3th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, Mexico, Mexique, 2002.

[127] J.P. VINAY et J. DARBELNET. *Stylistique comparée du français et de l'anglais*. Didier, Paris, 1966.

[128] J. VÉRONIS. Annotation automatique de corpus : état de la technique. *Ingénierie des langues, Hermes*, 1(1):52–58, 2000.

[129] S. K.M. WONG, W. ZIARKO et P.C.N. WONG. Generalized vector space model in information retrieval. In *In Proceedings of the 8th Annual International ACM SIGIR Conference On Research and Development in Information Retrieval*, Montréal, Quebec, Canada, 1985.

[130] W.A. WOODS et J. AMBROZIAK. Natural language technology in precision content retrieval. In *In Proceedings of the International Conference on Natural Language Processing and Industrial Applications, (NLP+IA)*, Moncton, Canada, 1998.

[131] F. YOSHIKANE, F. TSUJI, K. KAGEURA et C. JACQUEMIN. Morpho-syntactic rules for detecting japanese term variation. *Natural Language Processing*, 10(4):3–32, 2003.

[132] R. ZAAFRANI. *Développement d'un environnement interactif d'apprentissage avec l'ordinateur de l'arabe langue étrangère*. Thèse de Doctorat, Université Lyon2, Lyon, France, 2001.

[133] C. ZHAI, X. TONG, N. MILIC-FRAYLING et D. EVANS. Evaluation of syntactic phrase indexing - clarit nlp track report. In *E. Voorhees and Donna K. Harman, editor, The Fifth Text REtrieval Conference (TREC-5)*, pages 347–358, Gaithersburg, Maryland, 1996.

[134] P. ZWEIGENBAUM et N. GRABAR. Liens morphologiques et structuration de terminologie. In *Ingénierie des connaissances*, pages 325–334. Harmattan, 2000.

[135] P. ZWEIGENBAUM, N. GRABAR et S. DARMONI. Apport de connaissances morphologiques pour la projection de requêtes sur une terminologie normalisée. In *Proceedings of 8ème conférence annuelle sur le traitement automatique des langues naturelles (TALN)*, pages 403–408, Tours, France, 2001.

Liste des tableaux

Table des figures

Annexes

ANNEXE A

Catégories grammaticales

Les catégories grammaticale de l'analyseur de diab [35] (voir section 4.5.1) :

- CC Conjunction de coordination
- CD Nombre cardinal
- DT Déterminant
- FW Mot etranger
- IN Préposition ou conjunction de subordination
- JJ Adjectif
- JJR Adjectif, comparatif
- JJS Adjectif, superlatif
- NN Nom
- NNS Nom pluriel
- NNP Nom propre singulier
- NNPS Nom propre pluriel
- PDT Predéterminant
- PRP Pronom personel
- PRP$ Pronom possessif
- RB Adverbe
- RBR Adverbe comparatif
- RBS Adverbe superlative
- SYM Symbole
- UH Interjection
- VB Verbe
- VBD Verbe, passé
- VBG Verbe, gerondif ou participe présent
- VBN Verbe, participe passé
- VBZ Verbe, 3ème personne du présent
- WP Wh-pronom
- WP$ Pronom possessif

ANNEXE B
Anti-dictionnaire

الي بين تحت علي اه ال ام ان او الَا تي في قد لقد لَا مَا مع هل ذَا
ASMA' AL-ISHARA
هذَا هذه هذَان هَاتين الَّاتي الَّاَاي الّوَاتي تلك
DAMA'IR (Pronoms)
انَا نحن انت انتمَا انتم انتن هو هي همَا هم هن
ASMA ASHART
مَا من اينمَا متي اين ايَان لمَا اذَا كلمَا مهمَا اذ حيث حيثمَا اني كيفمَا
ASMA AL ISTIFHAM
كيف هل من فيم مَا اين متي اني كم ايَان بم لم مّ لمَاذَا مَاذَا الَا
HOROUF Al-JARR (Préposition)
يمن علي الي في من عن كي و منذ حتي خلَا عدَا
AL-ATF (Coordination)
و ثم او ام بل لكن لَا حتي
NID'A
يَا ايَا هيَا اي
NAFY (Négation)
لن لم لمَا لَا مَا ان
TAWKEED
ان قد
AL-SHART
ان اذ لو لولَا امَا لمَا

ANNEXE C
Requêtes

```
Req1
<title> التلوث الهوائي < /title>
<desc> البحث على الصبوس التي تحدث عن تلوث الهواء. </desc>
<narr> تلوث الهواء والاضرار التي يمكن ان يسببها العوامل التي ادت إلى تلوث الهواء. </narr>

Req2
<title> تقنية الميلد < /title>
<desc> البحث على الصبوس التي تحدث عن تقنية الميلد. </desc>
<narr> الطرق والوسائل الحديثة لتقنية الميلد. </narr>

Req3
<title> تلوث البيئة </title>
<desc> البحث على الصبوس التي تحدث عن تلوث البيئة. </desc>
<narr> التلوث البصري التلوث الإشعاعي. التلوث الضوضائي. مصادر تلوث البيئة. تلوث الميلد التلوث الهوائي </narr>

Req4
<title> التلوث الإشعاعي < /title>
<desc> البحث على الصبوس التي تحدث عن التلوث الإشعاعي. </desc>
<narr> مصادر التلوث الإشعاعي وآثاره . </narr>

Req5
<title> معالجة الميلد < /title>
<desc> البحث على الصبوس التي تحدث عن معالجة الميلد. </desc>
<narr> طرق معالجة الميلد وتقنيتها . </narr>

Req6
<title> ملوثات الهواء < /title>
<desc> البحث على الصبوس التي تحدث عن ملوثات الهواء. </desc>
<narr> طرق تلوث الهواء. ثاني أوكسيد الكربون تكوين ملوثات الهواء. </narr>

Req7
<title> مخلفات الصناعة < /title>
<desc> البحث على الصبوس التي تحدث عن مخلفات الصناعة. </desc>
<narr> مخلفات الصناعة الصلبة وطرق التخلص فيها مخلفات المصانع مكونات مخلفات الصناعة طرق معالجة مخلفات الصناعة </narr>

Req8
<title> مرض السرطان < /title>
<desc> البحث على الصبوس التي تحدث عن مرض السرطان. </desc>
<narr> مسببات السرطان تأثير التلوث على صحة الإنسان. </narr>

Req9
<title> سرطان الرئة < /title>
<desc> العثور على الصبوس التي تحدث عن سرطان الرئة. </desc>
<narr> سرطان الرئة تأثير التلوث على صحة الإنسان مرض السرطان . </narr>

Req10
<title> الخلف الجوي < /title>
<desc> البحث على الصبوس التي تحدث عن الخلف الجوي. </desc>
<narr> آثار التلوث على الخلف الجوي. مكونات الخلف الجوي </narr>

Req11
<title> تغير درجة الحرارة < /title>
<desc> البحث على الصبوس التي تحدث عن تغير درجة الحرارة. </desc >
<narr> ظاهرة الإحتباس الحراري. تغير المناخ < /narr>

Req12
<title> الأشرار الكيماوية < /title>
<desc> البحث على الصبوس التي تحدث عن الأشرار الكيماوية. </desc>
<narr> الأشرار الكيماوية وطرق الحد منها. الأسمدة الكيماوية التلوث بالمبيدات السم بالمبيدات الحشرية </narr>

Req13
<title> حماية الغابات < /title>
<desc> البحث على الصبوس التي تحدث عن حماية الغابات. </desc>
<narr> منظمة قاطني الأشجار طرق التهوين بالشجر وتشجير المساحات الخضراء < /narr>

Req14
<title> مكافحة الصدر < /title>
<desc> البحث على الصبوس التي تحدث عن مكافحة الصدر. </desc>
<narr> الشجر لمكافحة الصدر الوقاية ومعالجة الصدر. < /narr>

Req15
<title> مخلفات الضجيج < /title>
<desc> البحث على الصبوس التي تحدث عن مخلفات الضجيج. </desc>
<narr> البيئة المزعجة تؤثر سلبيا على نفسية الأطفال الضوضاء تضعف المناعة. </narr>
```

Figure C.1 – Exemple des requêtes

ANNEXE D

Transcription de Buckwalter

ا	A	ر	r	غ	g
ب	b	ز	z	ف	f
ت	t	س	s	ق	q
ث	v	ش	$	ك	k
ج	j	ص	S	ل	l
ح	H	ض	D	م	m
خ	x	ط	T	ن	n
د	d	ظ	Z	ه	h
ذ	*	ع	E	و	w

Traitement Automatique des Langues et Recherche d'Information en langue arabe dans un domaine de spécialité : Apport des connaissances morphologiques et syntaxiques pour l'indexation

Siham BOULAKNADEL

Résumé

La Recherche d'Information a pour objectif de fournir à un utilisateur un accès facile à l'information qui l'intéresse, cette information étant située dans une masse de documents textuels. Afin d'atteindre cet objectif, un système de recherche d'information doit représenter, stocker et organiser l'information, puis fournir à l'utilisateur les éléments correspondant au besoin d'information exprimé par sa requête. La plupart des systèmes de recherche d'information (SRI) utilisent des termes simples pour indexer et retrouver des documents. Cependant, cette représentation n'est pas assez précise pour représenter le contenu des documents et des requêtes, du fait de l'ambiguïté des termes isolés de leur contexte. Une solution à ce problème consiste à utiliser des termes complexes à la place de termes simples isolés. Cette approche se fonde sur l'hypothèse qu'un terme complexe est moins ambigu qu'un terme simple isolé.

Notre thèse s'inscrit dans le cadre de la recherche d'information dans un domaine de spécialité en langue arabe. L'objectif de notre travail a été d'une part, d'identifier les termes complexes présents dans les requêtes et les documents. D'autre part, d'exploiter pleinement la richesse de la langue en combinant plusieurs connaissances linguistiques appartenant aux niveaux morphologique et syntaxique, et de montrer comment l'apport de connaissances morphologiques et syntaxiques permet d'améliorer l'accès à l'information. Ainsi, nous avons proposé une plate-forme intégrant divers composants dans le domaine public ; elle conduit à montrer l'apport significatif et tranché de plusieurs de ces composants. En outre, nous avons défini linguistiquement les termes complexes en langue arabe et nous avons développé un système d'identification de termes complexes sur corpus qui produit des résultats de bonne qualité en terme de précision, en s'appuyant sur une approche mixte qui combine modèle statistique et données linguistiques

Mots-clés : Langue Arabe, extraction terminologique, domaine de spécialité, traitement automatique du langage naturel, recherche d'information.

NLP and IR for Arabic language in specific domain: contribution of morphological and syntactical knowledge for indexing

Abstract

Information retrieval aims to provide to a user an easy access to information. To achieve this goal, an IRS must represent, store and organize information, then provide to the user the elements corresponding to the need for information expressed by his query. Most of information retrieval systems (IRS) use simple terms to index and retrieve documents. However, this representation is not precise enough to represent the contents of documents and queries, because of the ambiguity of terms isolated from their context. A solution to this problem is to use multi-word terms to replace simple term. This approach is based on the assumption that a multi-word term is less ambiguous than a simple term.

Our thesis is part of the information retrieval in Arabic specific domain. The objective of our work was on the one hand, identifying a multi-word terms present in queries and documents. On the other hand, exploiting the richness of language by combining several linguistic knowledge belonging at the morphological and syntax level, and showing how the contribution of syntactic and morphological knowledge helps to improve access to information. Thus, we proposed a platform integrating various components in the public domain; it leads to show significant contribution of these components. In addition, we have defined linguistically a multi-word term in Arabic and we developed a system of identification of multi-word terms which is based on a mixed approach combining statistical model and linguistic data.

Keywords: Arabic language, specific domain, nlp, IR, terminology extraction.

Discipline : INFORMATIQUE

Laboratoire : LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE.
CNRS FRE 2729. 2, rue de la Houssinière, BP 92 208 . 44 322 Nantes, CEDEX 3.