# Extraction de Connaissances a partir de Textes : M ethodes et Applications

Chiraz Latiri

**Université de LORRAINE** **École doctorale IAEM Lorraine**

# Extraction de Connaissances à partir de Textes : Méthodes et Applications

## Mémoire de Recherche

présenté et soutenu publiquement le 24 Juin 2013

en vue de l'obtention d'une

**Habilitation à Diriger les Recherches de l'Université de LORRAINE**

**(Spécialité Informatique)**

par

Chiraz Latiri Cherif

| | |
|---|---|
| *Président :* | M. Dominique Mery, Professeur (Université de Lorraine) |
| *Rapporteurs :* | Mme. Amel Bouzeghoub, Professeur (Institut Télécom SudParis)<br>M. Eric Gaussier, Professeur (Université Joseph Fourier, Grenoble I)<br>M. Pascal Poncelet, Professeur (Université de Montpellier 2) |
| *Examinateurs :* | M. Kamel Smaïli, Professeur (Université de Lorraine)<br>M. Yahya Slimani, Professeur (ISAMM, Université de la Manouba) |

**Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503**

Loria
Laboratoire lorrain de recherche en informatique et ses applications

Mis en page avec la classe thloria.

# Table des matières

# Remerciements

Je voudrais, à travers ces quelques lignes, remercier très sincèrement les membres du jury :
- M. Dominique Mery, Professeur à l'Université de Lorraine pour l'honneur qu'il m'a fait en acceptant de présider le jury de mon habilitation.
- Mme. Amel Bouzeghoub, Professeur à l'Institut Télécom SudParis, M. Eric Gaussier, Professeur à l'Université Joseph Fourier de Grenoble et M. Pascal Poncelet, Professeur à l'Université Montpellier 2, d'avoir bien voulu rapporter mon mémoire d'habilitation à diriger les recherches malgré leurs charges.
- M. Kamel Smaïli, Professeur à l'Université de Lorraine et M. Yahya Slimani, Professeur à l'Institut Supérieur des Arts Multimédia de la Manouba, pour l'intérêt qu'ils ont porté à mes travaux en acceptant de faire partie du jury en tant qu'examinateurs.

Il m'est difficile de tenir en quelques lignes tous les remerciements que j'aimerais adresser à ceux et à celles qui ont permis à ce modeste travail d'exister et de progresser au fil des années. C'est grâce à eux tous que ce mémoire a pu voir le jour.

Rédiger un mémoire d'Habilitation à Diriger les Recherches, c'est toujours faire un bilan de plusieurs années de recherches. Je tiens à dire combien les rencontres que j'ai eu la chance de faire avec d'autres chercheurs ont enrichi ma réflexion et ma maturité scientifique.

Au-delà de la formalité d'usage, c'est avec une grande reconnaissance que je remercie les membres de l'École Doctorale IAEM de Nancy-Université et les membres du jury pour le temps qu'ils ont consacré à l'évaluation de mon travail.

Mes remerciements et ma gratitude vont tout d'abord à Yahya Slimani, Professeur à l'Institut Supérieur des Arts Multimédias de la Manouba (Université de la Manouba), pour sa confiance et pour l'autonomie qu'il m'a accordée durant toutes ces années passées au sein du Laboratoire d'Informatique en Programmation Algorithmique et Heuristique (LIPAH, Faculté des Sciences de Tunis, Université Tunis El Manar). Je suis fière de l'avoir eu comme mentor et d'avoir appris à ses côtés la rigueur scientifique et les vraies valeurs universelles tant sur le plan humain que sur le plan scientifique. Je le remercie pour toutes les collaborations partagées durant les dernières années écoulées. Je n'oublierais jamais son soutien et ses précieux conseils dans les moments difficiles que j'ai affronté les dernières années. Merci cher professeur pour votre générosité et votre patience.

Je tiens ensuite à remercier Mohamed Ben Ahmed, Professeur émérite à l'École Nationale des Sciences de l'Informatique de l'université de la Manouba (Tunis), pour m'avoir adopté il y a déjà de nombreuses années et pour les conseils précieux qu'il m'a donné. Je lui addressee toute ma gratitude pour m'avoir confié la co-direction d'une thèse où j'ai eu l'occasion d'apprécier sa richesse et son exigence scientifiques.

Mes plus vifs remerciements s'adressent par la suite à Kamel Smaïli, Professeur à l'université de Lorraine. Je l'ai connu en 2008 sur les bancs d'une conférence et depuis ce jour, il me montre le chemin à suivre avec quelques années d'avance et il m'a soutenu avec une grande générosité pour rédiger cette habilitation. C'est aussi grâce à lui qu'une partie des mes contributions s'est articulée autour de la Traduction Automatique Statistique. Je le remercie du fond du cœur pour la confiance qu'il m'a accordée pour la co-direction de travaux de recherche et d'avoir accepté d'être mon parrain scientifique au sein de l'université de Lorraine. Je lui témoigne toute ma gratitude et ma reconnaissance pour ses conseils, ses encouragements, son sens critique, ses fortes convictions scientifiques, qui m'ont permis de progresser et de m'affirmer en tant que chercheur. Merci cher Kamel de te trouver toujours près de moi.

Une pensée amicale va à mes compagnons de toujours dans la recherche, Chiraz Trabelsi, Sadok Ben Yahia, Hatem Haddad et Tarak Hamrouni. Sans leur conseils patients, leurs encouragements, je me serais encore plus souvent détournée du chemin de la rédaction. Je les remercie pour la collaboration et les échanges scientifiques que nous partageons ensemble, sans oublier notre amitié indéfectible qui nous aide à surmonter les phases difficiles. J'exprime en particulier toute ma reconnaissance à Hatem et Tarak, qui mènent avec moi depuis six ans la lourde mission de dynamiser et pérenniser les activités de recherche du groupe "Fouille de données textuelles" au sein du Laboratoire d'Informatique en Programmation Algorithmique et Heuristique (LIPAH) de la Faculté des Sciences de Tunis.

J'exprime aussi tous mes remerciements à mes collègues de l'ISAMM où j'ai passé les six plus belles années de ma carrière académique en tant que directrice. Je garde le souvenir d'une équipe exceptionnelle tant sur le plan humain que sur le plan professionnel. Aujourd'hui, l'occasion est arrivée pour les saluer pour les moments de défis, de bonheur que j'ai partagé avec eux. Tout en servant l'école avec beaucoup d'engagement et d'abnégation, j'ai eu la chance et le plaisir de me ressourcer de leur dynamisme et leur amitié pour avancer dans mon projet d'habilitation. Une pensée particulière à mon amie Hajer Baazaoui, car nous nous sommes toujours encouragées mutuellement pour ne pas baisser les bras et progresser dans nos travaux de recherche.

Ma réflexion scientifique s'est également enrichie des nombreux échanges que j'ai eu avec les jeunes étudiants-chercheurs en master et en thèse. Je tiens à les remercier très sincèrement pour avoir contribué à mes recherches et d'avoir partagé avec moi leur vivacité de jeunesse. Je témoigne ici de leur mérite et je leur souhaite un avenir plein de succès. Une pensée particulière à mon étudiant Brahim Douar, dont la curiosité et la rigueur scientifique ont fait de nos réunions de travail un vrai plaisir partagé et ont donné lieu à d'excellentes contributions de recherche.

Ce mémoire est aussi le fruit de longues années de travail et de collaborations scientifiques avec des équipes de recherche françaises dans le cadre d'échanges scientifiques Tuniso-français et de projets CMCU. À ce titre, je remercie vivement les professeurs et chercheurs français et étrangers qui ont contribué de près ou de loin à mes travaux de recherche. Je commencerais par saluer tous les membres de l'équipe MRIM au sein du Laboratoire d'Informatique de Grenoble (LIG) qui m'ont beaucoup soutenu lors de l'élaboration de ma thèse de doctorat. Toute ma gratitude va à mon ami Engelbert Mephu Nguifo du LIMOS (Université de Clermont-Ferrand) qui m'a beaucoup apporté, que ce soit par ses conseils, son soutien ou par sa dynamique de recherche, à laquelle il m'a toujours associée. J'exprime aussi toute ma reconnaissance à Michel Liquière du LIRMM (Université de Montpellier 2) et Lynda Tamine du l'Institut de Recherche en Informatique de Toulouse (IRIT) avec qui j'ai eu le plaisir de travailler. Je les remercie pour la confiance qu'ils m'ont accordée pour co-diriger des travaux de recherche avec eux, et pour les échanges scientifiques fructueux partagés dans la convivialité.

Mes sincères remerciements s'adressent également à Caroline Lavecchia et David Langlois de l'équipe PAROLE du Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA) à Nancy, pour la collaboration fructueuse que nous avons mise en place ensemble durant les cinq dernières années.

Un grand Merci à Amina, Chahla et Sawssen, mes meilleures amies de toujours, tous les moments de détente et de bonheur passés à vos cotés, nos fous rires, nos pleurs, nos espoirs et nos rêves ont contribué à faire émerger ce travail. À l'avenir, ces moments seront plus fréquents.

Enfin, je ne peux clore ces remerciements sans faire une place spéciale à ma famille : tout d'abord ma mère et mon père, pour tout ce qu'ils m'ont apportés, leur présence, leur soutien moral constant et leurs encouragements, pour se soucier régulièrement de ma vie et de ma carrière ; mes sœurs et mon frère, pour leur attention, leur affection et les moments de bonheur partagés ensemble m'ont permis d'avancer.

À mon époux et mes deux très chers enfants Zeineb et Brahim : aucun mot ne peut faire oublier mes périodes d'absence (parfois même en étant physiquement près de vous). L'aboutissement de ce travail est aussi le fruit de ces instants. Et à ce titre, j'espère que vous ne me tiendrez pas rigueur. Merci pour votre grande patience et surtout votre amour sans égal.

Que ceux que je n'ai pas cité, et qui à leur manière m'ont apporté leur aide et leur soutien, m'excusent et soient remerciés du fond du cœur.

"*Aucun homme ne peut rien vous révéler sinon ce qui repose déjà à demi endormi dans l'aube de votre connaissance*"

**Gibran Khalil Gibran (Le Prophète, 1923)**

"*Le savoir acquis en exil est une patrie et l'ignorance en patrie est un exil.*"

**Ibn Rushd (Averroès, 1126-1198)**

À la mémoire de ma très chère amie Saoussen, qui accompagnera à vie mes pensées ... Paix à son âme.

Aux "Andaloussiates" qui m'ont accompagné durant mes longues soirées de travail ...

Merci.

## Première partie

# CV détaillé et Synthèse des Activités Académiques et Scientifiques

# CV de synthèse

## 1 État civil

**Chiraz Latiri *épouse* Cherif**

- Née le 24/03/1972 à Hammam Sousse, Tunisie.
- Nationalité : Tunisienne.
- Mariée et mère de deux enfants.
- **GSM :** (+216) 24 33 45 54
- **Tél professionnel** : (+216) 71 603 498
- **Fax** : (+216) 71 603 450
- **Email :** chiraz.latiri@gnet.tn
- **Adresse personnelle :** 45, Avenue Jugurtha, 1002, Mutuelleville, Tunis, Tunisie.
- **Adresse professionnelle :** Institut Supérieur des Arts Multimédias de la Manouba, Campus Universitaire de la Manouba, 2010 Tunis, Tunisie.

## 2 Titres académiques

| Année | Diplôme |
|---|---|
| **2004** | **Thèse de Doctorat en Informatique**<br>**Titre de la thèse :** *"Approche de découverte de règles d'association classiques et floues à partir de textes : Application à la Recherche d'Information."*<br>**Préparée à** l'École Nationale des Sciences de l'Informatique (Université de la Manouba, Tunis) et au sein de l'équipe MRIM (LIG, Grenoble).<br>**Date de la soutenance :** 10 avril 2004.<br>**Directeurs de thèse** :<br>Ali Jaoua : Professeur à la Faculté des Sciences de Tunis.<br>Marie-France Bruandet : Professeur à l'Université Joseph Fourrier, Grenoble. |
| **1997** | **Diplôme d'Études Approfondies en Modélisation et Informatique de Gestion**<br>**Titre du Mémoire :** *"Les Systèmes Experts Flous : Application au problème des lâchers d'eau au niveau des barrages du nord de la Tunisie."*<br>Préparé au sein du Groupe de Recherche et d'Aide à la Décision de l'Institut Supérieur de Gestion de Tunis.<br>**Date de Soutenance :** 7 avril 1997.<br>**Directeur de DEA** : Foued Ben Abdelaziz : Professeur à l'ISG de Tunis. |
| **1994** | **Maîtrise en Méthodes Informatiques Appliquées à la Gestion D'Entreprises** de l'Institut Supérieur de Gestion de Tunis. |
| **1990** | **Baccalauréat Tunisien,** section Math-Sciences, Lycée de Garçons de Sousse. |

# 3 Situation professionnelle

| | |
|---|---|
| **Depuis Septembre 2012** | Maitre-Assistante de l'Enseignement Supérieur à l'Institut Supérieur des Arts Multimédias (Département Informatique), Université de la Manouba. |
| **Septembre 2011 - Juin 2012** | En congé d'études en vue de préparer une Habilitation Universitaire en Informatique. |
| **Janvier 2006 - Juillet 2011** | Directrice de l'Institut Supérieur des Arts Multimédias de la Manouba, Université de la Manouba. |
| **Depuis Mai 2005 - Août 2012** | Maitre-Assistante de l'Enseignement Supérieur à l'École Supérieure de Commerce de Tunis, Université de la Manouba. |
| **Avril 2001 - Avril 2005** | Assistante de l'Enseignement Supérieur à l'École Supérieure de Commerce de Tunis, Université de la Manouba et responsable de la MIAGE à l'ESC. |
| **Septembre 1998 - Février 2001** | Assistante contractuelle à l'Institut Supérieur de Gestion de Tunis au sein du département MIAGE. |
| **Septembre 1995 - Septembre 1998** | Assistante vacataire à l'Institut Supérieur de Gestion de Tunis au sein du département MIAGE. |

# Responsabilités pédagogiques et administratives

L'ensemble des responsabilités que j'ai assurées sont présentées ci-dessous en les associant à mes différentes activités pédagogiques, administratives et associatives.

## 1 Coordination pédagogique de la maîtrise MIAGE à l'ESC de Tunis (Avril 2001 - Janvier 2006)

J'ai été chargée de la coordination de la section MIAGE à l'École Supérieure de Commerce de Tunis (ESC) de Avril 2001 à Janvier 2006. Cette coordination comprend la gestion de la planification pédagogique, le suivi du déroulement des cours et la gestion des projets de fin d'études. Durant cette période, j'ai assuré les tâches suivantes :

- Proposition de la répartition pédagogique des différents modules enseignés au niveau de la MIAGE en tenant compte des spécialisations des enseignants de l'équipe pédagogique.
- Suivi du déroulement des cours, des travaux dirigés ainsi que des travaux pratiques relatifs aux différents enseignements dispensés.
- Mise en place et animation de réunions de coordination pédagogiques.
- Planification des séminaires d'initiation à la recherche pour les étudiants en maîtrise de la MIAGE. Ces séminaires ont été animés par des professeurs invités, principalement des universités françaises.
- Gestion des stages en entreprise et des projets de fin d'études des étudiants en maîtrise, à partir de leur intégration jusqu'à la soutenance de leurs projets.
- Création d'un réseau d'entreprises nationales et internationales qui sont impliquées dans la formation MIAGE, à travers l'intervention d'experts professionnels dans l'encadrement des étudiants lors de l'élaboration de leurs projets de fin d'études.

## 2 Direction de l'Institut Supérieur des Arts Multimédias (Janvier 2006 - Juillet 2011)

Durant la période allant de Janvier 2006 à Juillet 2011, j'ai assuré deux mandats, en tant que directrice de l'Institut Supérieur des Arts Multimédias à l'Université de la Manouba (ISAMM). Cette institution comprend 2400 étudiants répartis sur trois départements : (*i*) le département *d'informatique* qui englobe un cycle d'ingénieurs en *Informatique et Multimédia* et une licence fondamentale en *Informatique* ; (*ii*) le département *Multimédia* qui offre une formation de licence appliquée en *Communication Multimédia*, une licence professionnelle (co-construite) en *Modélisation et Animation 3D* et deux mastères professionnels liés à *l'Image Numérique et à*

*l'Ingénierie des médias* ; et, (*iii*) un département de *cinéma* qui dispense une licence appliquée en *Cinéma et audiovisuel* et un mastère professionnel en *Production Audiovisuelle*.

Mes fonctions au niveau de la direction de l'ISAMM ont couvert la dimension administrative de la gestion de l'institution ainsi que les dimensions pédagogique et scientifique relatives aux trois départements de l'ISAMM.

Les missions les plus significatives que j'ai menées en tant que directrice de l'ISAMM sont résumées ci-dessous dans l'ordre chronologique de leur réalisation.

## 2.1 Missions administratives

- Restructuration de l'organigramme de l'ISAMM et création de la direction des études et des stages en 2007.
- Suivi du projet de construction des nouveaux bâtiments de l'ISAMM sur le campus universitaire de la Manouba depuis 2006 jusqu'au déménagement et à l'installation en septembre 2009.
- Création du services "production" et "post-production" rattachés au département *Cinéma* et la mise en place d'un studio de montage et de mixage en septembre 2007.
- Mise en place d'un studio de tournage dans les nouveaux locaux en septembre 2009.
- Mise en place d'un laboratoire de réalité virtuelle, destiné aux élèves ingénieurs en septembre 2010.
- Mise en place d'un ensemble de procédures pour la bonne application de la réforme LMD dans les trois départements de l'ISAMM.
- Elaboration du projet d'établissement de l'ISAMM pour la période 2010-2013. Ce projet a introduit plusieurs réflexions constructives qui sont sont liées à : (1) l'amélioration de la prestation de service d'enseignement, couvrant la stabilisation des effectifs et l'augmentation du taux d'encadrement ; (2) l'amélioration des méthodes de l'enseignement présentiel et à distance ; (3) la remise à niveau du corps administratif de l'ISAMM ; (4) la consolidation du cycle du mastère dans le cadre de la réforme LMD ; et, (5) la mise en place d'un mastère de recherche en Informatique, spécialité *Image numérique et intéraction*, rattaché à l'école doctorale en Informatique de l'Université de la Manouba.
- Insertion de l'ISAMM dans un réseau d'institutions de formation réputées ainsi que dans un grappe d'entreprises nationales et internationales issues des domaines de l'informatique, des médias numériques et du cinéma.
- Mise en place et signature de conventions cadre avec des institutions et universités étrangères francophones suivantes : l'école des Mines de Paris, l'ENIB de Brest, l'UFR ATI de l'Université Paris 8, l'École Ingémedia de l'Université de Toulon-Var, l'Université de Versailles Saint-Quentin-en-Yvelines (UVSQ), l'INSAS de Bruxelles, l'Université du Québec en Abitibi-Témiscamingue (UQAT), l'ESAV de Marrakech (Maroc) et l'Université de Mouloud Mammeri Tizi-Ouzou (Algérie).

## 2.2 Missions pédagogiques et scientifiques

- Participation active et régulière aux commissions pédagogiques de l'ISAMM pour l'élaboration des nouvelles habilitations des licences en Informatique et en Multimédia, et ce dans le cadre de l'application de la réforme LMD en Tunisie.
- Montage du projet de collaboration entre l'ISAMM et l'INSAS de Bruxelles en décembre 2007 pour la période 2008-2010, dans le cadre de la coopération mixte Tunisie-Wallonie/Bruxelles pour un appui aux formations en cinéma et audiovisuel dispensées à l'ISAMM.

- Mise en place, à l'ISAMM dans le cadre de la réforme LMD, du cycle d'ingénieurs en *Informatique et Multimedia* et des mastères professionnels en *Ingénierie des médias*, *Multimédia et Image numérique* et *Production et assistanat à la réalisation* en Septembre 2009. Ces formations ont induit des partenariats avec des écoles et des universités françaises, à savoir l'École des Mines de Paris, l'ENIB de Brest, l'UFR ATI de l'Université Paris 8, l'École Ingémedia de l'Université de Toulon-Var et l'Université de Versailles Saint-Quentin-en-Yvelines (UVSQ).
- Participation, dans la cadre de l'adhésion de l'ISAMM à l'Université Internationale de Multimédia, à la mise en place d'un mastère professionnel en *"Médias Numériques en Contexte Interculturel"* avec l'Université du Québec en Abitibi-Témiscamingue (UQAT), l'École Ingémédia de l'Université de Toulon-Var et L'Université de Versailles Saint-Quentin-en-Yvelines (UVSQ).
- Montage du projet Euromed Audiovisuel III intitulé "*Développement de l'industrie audiovisuelle Sud-Méditerranéenne par des formations d'excellence et des rencontres professionnelles, DIA Sud-Med*", avec L'École Supérieure des Arts Visuels de Marrakech (Maroc) et l'Académie Libanaise des Beaux Arts de Beirut (Projet validé par l'Union Européenne en Décembre 2010).
- Montage du projet de reconduction de la collaboration entre l'ISAMM et l'INSAS de Bruxelles en Décembre 2010 pour la période 2011-2013, dans le cadre de la coopération mixte Tunisie-Wallonie/Bruxelles pour un appui aux formations en cinéma et audiovisuel dispensées à l'ISAMM.
- Montage d'un projet de collaboration entre l'ISAMM et le centre de formation Technocité de Mons (Belgique) en Décembre 2010 pour la période 2011-2013, dans le cadre de la coopération mixte Tunisie-Wallonie/Bruxelles pour la mise en place à l'ISAMM d'un pôle d'excellence dans le domaine des médias numériques en Tunisie.
- Mise en place d'une licence "*Informatique et Multimédia*" au sein du département Informatique à la Faculté des Sciences de l'Université Mouloud Mammeri à Tizi-Ouzou (Algérie), en réponse à une forte demande du marché d'emploi algérien dans les secteurs des médias numériques. À ce titre, plusieurs rencontres et réunions entre l'équipe pédagogique de l'ISAMM et celle de l'Université Mouloud Mammeri ont été organisées durant l'année 2010 et ont donné lieu à la première habilitation de la licence "*Informatique et Multimédia*", dispensée en Algérie. La formation a démarré en septembre 2011.
- Participation à la mise en place d'un master de recherche en Informatique au sein de l'Institut Supérieur des Arts Multimédia de La Manouba (Démarrage prévu en septembre 2013).

## 2.3 Organisation de manifestations scientifiques et culturelles

- Organisation de l'école d'automne de l'Université Internationale du Multimedia (UIM) qui s'est tenue en Novembre 2008 à Hammamet.
- Organisation d'un séminaire sur *"La réalité virtuelle et les technologies du web : métiers d'avenir et défis de demain"*, les 12 et 13 Mars 2010 à l'Université de la Manouba avec l'école des Mines de Paris, l'ENIB de Brest et l'UFR ATI de l'Université Paris 8.
- Organisation d'un séminaire sur *"Le patrimoine Musical arabo-andalou au cœur des arts multimédias"*, les 7 et 8 Mai 2010 à l'Université de la Manouba avec la participation de plusieurs universitaires et professionnels maghrébins.
- Organisation de la *"Première Journée du Numérique"* avec la délégation Wallonie-Bruxelles en Tunisie le 24 Novembre 2010 à l'Université de la Manouba, dont le but est la mise en

réseau de sociétés belges dans le domaine de la production numérique avec les sociétés tunisiennes, partenaires de l'ISAMM et leur implication dans l'appui aux formations dispensées.

### 2.4 Autres participations

- Membre du comité du pilotage des programmes de certification C2I au sein du Ministère de l'Enseignement Supérieur et de la Recherche Scientifique de Mars 2006 à Juin 2009.
- Membre du jury du concours national de recrutement des assistants en Sciences et Techniques Audiovisuelles pour les sessions de Juillet 2006 et Juillet 2007.
- Rapporteur du symposium *"Education, Sciences et Développement Technologique"* pendant les *Assises de la Recherche Scientifique et de l'Innovation Technologique* tenues à Tunis les 19 et 20 Novembre 2007 et organisées par le Ministère de l'Enseignement Supérieur et de la Recherche Scientifique.
- Participation aux rencontres professionnelles *"État et perspectives du secteur audiovisuel au Maroc"* les 30, 31 octobre et $1^{er}$ novembre 2009 à l'ESAV de Marrakech (Maroc).
- Participation aux journées Audiovisuelles de Tunis, organisées par l'ambassade de France à Tunis, du 25 au 27 octobre 2010, dans la session *"Production et formation : comment mieux travailler ensemble ?"*.

## 3 Activités en milieu associatif

J'ai crée en Décembre 2008 l'**A**ssociation culturelle du **M**ultimédia et de l'**A**udio**VI**suel (AMAVI), dont je suis présidente jusqu'à ce jour. Cette association a pour objectif de promouvoir les métiers de demain liés aux nouveaux médias et à la création numérique. Elle a comme projets d'organiser des séminaires et des formations autour de ces thèmes. Dans ce contexte, l'association a soutenu une bonne partie des manifestations scientifiques et culturelles organisées par l'ISAMM.

La création de cette association est motivée par le constat que l'université tunisienne produit chaque année une centaine de travaux d'étudiants. Ces travaux sont sous forme de fictions et de documentaires, d'affiches, de spots publicitaires, de films d'animation 2D/3D, d'animatiques ou encore des sites en ligne. Il importe de signaler qu'une sélection de ces travaux a été, à plusieurs occasions, appréciée et primée à l'échelle nationale et internationale. L'association AMAVI vient ainsi encourager toute ces formes de réalisations numériques et audiovisuelles.

Au début de l'année 2011, l'association AMAVI s'est alliée avec l'**A**ssociation **T**unisienne des **L**ibertés **N**umériques (ATLN) et d'autres membres de la société civile pour la création et la mise en place d'un média citoyen dédié à l'information et à l'éveil socio-politique [1]. L'association AMAVI contribue principalement dans le volet de la création numérique. Le lien établit avec l'association ATLN a donné naissance à un partenariat très étroit avec Canal France International (CFI) dans le domaine de la promotion des nouveaux médias. À ce titre, j'ai organisé les 12 et 13 Janvier 2012, le colloque 4M Tunis qui s'inscrit dans la logique d'accompagnement de la transformation des médias traditionnels vers une logique "nouveaux médias" en Tunisie.

L'association AMAVI est également chargée, pour les deux années à venir, de développer une plate-forme Web "Marhaba Médias" pour structurer et mettre en réseau le partage d'information et la coopération médias et audiovisuelle dans les pays du Maghreb.

1. www.fhimt.com

De plus, je suis impliquée avec l'appui de CFI dans la mise en place d'un programme d'ateliers, de visites et de rencontres de haut niveau, destinés aux jeunes tunisiens et qui se tiendront entre Tunis et Paris à partir du mois de Septembre 2012 et 2013 (SAFIR'Lab). L'objectif principal de cette initiative est d'identifier les futurs élites ou leaders d'opinion issus de la société civile (réseaux sociaux, milieux associatifs) qui n'ont pas suivi de parcours classique scolaire ou universitaire et de leur offrir une formation ad hoc dans le domaine économico-politique et médiatique. À court terme, je suis ainsi chargée de suivre et accompagner leurs trajectoires professionnelles et animer ce réseau.

Je suis également membre de l'association ARIA, Association Francophone de Recherche d'Information et Applications et l'association internationale francophone d'Extraction et Gestion des Connaissances (EGC). Ces associations organisent chaque année, respectivement, les conférences francophones CORIA et EGC.

# Activités d'enseignement

Après la validation de ma première année de DEA, j'ai commencé mon cursus d'enseignement par un poste d'étudiante-contractuelle de 1995 à 1996 à l'ISG de Tunis au sein du département MIAGE. Depuis, mes activités d'enseignement se sont poursuivies comme décrit ci-dessous.

## 1 Chronologie et évolution

Les 17 années d'expériences d'enseignement et d'implication dans la vie pédagogique au sein de l'Université Tunis I et l'Université de la Manouba se répartissent comme suit : 2 années en tant que vacataire (Septembre 1996 - Septembre 1998), 3 années en tant que assistante contractuelle, 4 années en tant que assistante permanente de l'enseignement supérieur et 7 années en tant que maître assistante (depuis Mai 2005). L'évolution de mes enseignements au sein de l'université s'est faite en trois phases :

1. **Phase d'initiation** à l'enseignement correspondant à mes deux années de vacation durant lesquelles j'ai assuré les enseignements d'*algorithmique*, de *programmation ADA et C*, des *structures de données* et de *modélisation de l'information*, destinés aux étudiants du premier cycle de la maîtrise MIAGE à l'ISG de Tunis.

2. **Phase d'intégration** à l'équipe pédagogique de la MIAGE. Cette période de 7 années m'a permis de développer des cours dans le pôle "Ingénierie des SI" de la MIAGE de l'ESC de Tunis pour des futurs informaticiens et d'assurer en parallèle la coordination des chargés des TDs pour les cours de *Conception de Système d'Information*, *Génie logiciel* et *Fouille de données* dispensés dans la filière MIAGE.

3. **Phase de spécialisation** s'est opérée naturellement de 2001 à 2006 avec la coordination pédagogique de la filière MIAGE de l'ESC de Tunis. Les enseignements informatiques de la filière MIAGE sont généralement classés en deux catégories : les enseignements techniques et les enseignements liés à l'ingénierie des SI. J'ai été ainsi chargée, au sein de l'équipe pédagogique de la MIAGE, de dispenser des cours permettant d'aligner les enseignements techniques d'ingénierie du logiciel aux enseignements liés à l'ingénierie des SI. Dans ce cadre, j'ai été amenée à faire évoluer les enseignements que j'assure pour suivre l'évolution des technologies dans le domaine de l'ingénierie des logiciels, en intégrant les concepts qu'elles véhiculent, les nouvelles pratiques de développement qu'elles induisent et les impacts qu'elles peuvent avoir sur les phases de conception d'un SI.

   J'ai eu également l'occasion d'intervenir dans le cadre des mastères de recherche en informatique à l'ESC de Tunis et à la Faculté des Sciences de Tunis pour assurer un cours lié à mon domaine de recherche, intitulé "*Extraction de Connaissances à partir de Textes : Approches et Applications*".

# 2 Tableau récapitulatif des enseignements assurés durant la période 1995-2013

Le tableau ci-dessous résume les principaux cours et Tds que j'ai assurés depuis 1995.

| **Module enseigné et volume horaire semestriel** | **Public** | **Type** | **Années Universitaires** | **Établissement** |
|---|---|---|---|---|
| **Cycle d'ingénieurs en Informatique** | | | | |
| Interface Homme-Machine (42h) | $1^{ère}$ année du cycle d'ingénieurs en Informatique Multimédia | Cours intégré | à partir de Janvier 2013 | ISAMM |
| Gestion de projets Web (42h) | $2^{ème}$ année du cycle d'ingénieurs en Informatique Multimédia | Cours intégré | à partir de Janvier 2013 | ISAMM |
| **Niveaux Maîtrise et Licence** | | | | |
| Algorithmique et structures de données (42h) | $1^{ère}$ année de la maîtrise MIAGE | Cours et TDs | De 1995 à 2002 | ISG et ESC |
| Structures de données avancées (42h) | $2^{ème}$ année de la maîtrise MIAGE | Cours et TDs | De 1995 à 2002 | ISG et ESC |
| Langages de Programmation (ADA et C) (42h) | $1^{ère}$ année de la maîtrise MIAGE | Cours et TDs | De 1995 à 2002 | ISG et ESC |
| Logique (42h) | $1^{ère}$ année de la maîtrise MIAGE | Cours et TDs | De 2000 à 2002 | ESC |
| Conception des systèmes d'information (Merise et Merise 2) (42h) | $3^{ème}$ année de la maîtrise et L3 MIAGE | Cours et TDs | De 2002 à 2006 | ESC |
| Conception orientée objet (42h) | $3^{ème}$ année et L3 MIAGE | Cours et TDs | De 2002 à 2007 | ESC |
| Compilation (42h) | $3^{ème}$ année de la maîtrise MIAGE | Cours | De 2002 à 2003 | ISG et ESC |
| Processus unifié et UML (42h) | $4^{ème}$ année de la maîtrise MIAGE | Cours et TDs | De 2004 à 2006 | ESC |
| Génie logiciel et conduite de projet (63h) | $4^{ème}$ année de la maîtrise et L3 MIAGE | Cours et TDs | De 2004 à 2007 | ESC |
| Extraction des connaissances à partir des données (42h) | $4^{ère}$ année de la maîtrise MIAGE | Cours et TDs | De 2004 à 2006 | ESC |
| Conduite de projets multimédias (42h) | L3 de la licence appliquée en communication multimédia | Cours intégré | à partir de Septembre 2012 | ISAMM |
| **Niveaux Mastère professionnel et Mastère de recherche** | | | | |
| Bio-datamining (42h) | M1 Mastère professionnel *Bioinformatique* | Cours | 2005 et 2006 | ENSI |
| Conception orientée objet pour les applications multimédias (42h) | M1 Mastère professionnel *Multimédia et Image Numérique* | Cours | De 2008 à 2012 | ISAMM |
| Extraction des connaissances à partir de textes : Méthodes et Applications (21h) | Mastère de recherche en *Informatique* (M2) et Mastère de recherche *Optimisation des systèmes intelligents* (M2) | Cours | De 2011 à 2012 | FST et ESC |

# Activités de recherche

## 1 Contexte scientifique

L'extraction des connaissances à partir de textes (ECT) a constitué le noyau mes travaux de recherche depuis 2000. Elle représente un domaine scientifique pluridisciplinaire, fédérant des thématiques issues des sciences de l'information, de la linguistique, des statistiques et de l'intelligence artificielle. Selon Feldman *et al.* [Feldman *et al.*, 1998], l'ECT, appelée en anglais *Text Mining*, est définie comme "*une extension des approches traditionnelles de data mining aux données textuelles, tels que des documents semi-structurés, du texte intégral ou des corpus textuels*".

Pour faire face à l'augmentation sans cesse croissante du volume des données disponibles sous forme de corpus de textes ou de collections documentaires, l'un des principaux défis de la communauté de recherche en ECT est de proposer des méthodes et des techniques capables de traiter une telle masse de données textuelles pour extraire de la connaissance dans des délais raisonnables pour les utilisateurs. Très vite, l'ECT s'est trouvée au cœur de plusieurs domaines de recherche. L'usage globalisé des techniques de fouille de textes dans des applications réelles ne pourra se faire que par l'efficacité algorithmique des approches proposées. Ainsi, les défis propres à ce domaine amènent la communauté des chercheurs à innover autant du point de vue théorique et algorithmique, que de proposer des approches qui puissent être également déployées dans un cadre d'utilisation réel.

En effet, dès son apparition, l'ECT s'est trouvée au centre du domaine de la Recherche d'Information (RI). Ce croisement traite en grande partie des modèles, des techniques et des algorithmes permettant de sélectionner l'information pertinente en réponse à un besoin d'information, exprimé par un utilisateur à l'aide d'une requête. D'une manière générale, un processus de RI induit deux étapes fondamentales, à savoir : (*i*) l'étape de *l'indexation* permettant de produire, à partir d'un corpus textuel, des descripteurs canoniques qui identifient les granules d'information ; et, (*ii*) *l'étape de la sélection de l'information pertinente*, qui consiste à apparier les descripteurs issus de l'étape d'indexation avec les descripteurs de la requête utilisateur, dans le but d'identifier les informations qui répondent au mieux aux besoins couverts par la requête.

La revue de la littérature en RI a montré que des modèles tels que le modèle vectoriel ou le modèle probabiliste, font souvent appel à la technique d'expansion de requêtes ou de reformulation de requêtes afin de réduire le manque de correspondance entre la requête et les documents restitués. L'idée clé est d'étendre la requête par des termes additionnels, implicitement liés à ceux de la requête originelle [Latiri *et al.*, 2003c, Lin *et al.*, 2008]. Intuitivement, la finalité d'une telle technique ne se limite pas à l'amélioration de la mesure du rappel en récupérant des documents pertinents qui ne peuvent pas être trouvés par la requête utilisateur, mais également à améliorer la précision des documents restitués en les plaçant en haut de la liste des documents pertinents trouvés [Lin *et al.*, 2008].

La problématique de recherche qui nous intéresse dans le domaine de la RI, et particuliè-

rement dans le contexte de l'expansion de requêtes en RI, est la mise en œuvre d'une synergie entre les techniques classiques de RI et une technique d'ECT, à savoir l'extraction de règles d'association [Agrawal and Skirant, 1994]. Certains travaux de recherche ont déjà abordé cette problématique [Lin *et al.*, 2008]. L'idée clé est d'utiliser les connaissances additionnelles apportées par les règles d'association entre termes pour étendre les requêtes originelles, dans le but d'améliorer la pertinence système d'un SRI.

Par ailleurs, la RI ne cesse d'évoluer en tenant compte de nouvelles représentations et interprétations de connaissances offertes par l'ECT et par l'ingénierie de connaissances (IC). En effet, la majorité des systèmes de recherche d'information (SRIs) représentent les documents et les requêtes par des index souvent désignés par "sac de mots" [Baziz *et al.*, 2005]. Cette représentation stipule implicitement que les mots correspondent avec leurs sens. Plusieurs travaux de recherche ont mis en évidence les limites de tels modèles, qui sont étroitement liées aux ambiguïtés que peuvent véhiculer le manque d'expressivité des mots singuliers de l'index ainsi que l'imprécision des requêtes utilisateur. Pour pallier à ces limites, des travaux ont proposé d'utiliser des structures conceptuelles lors de l'indexation [Andreasen *et al.*, 2009]. Il importe de souligner que la majorité de ces travaux intègrent l'usage de ressources externes, telles que les ontologies et les hiérarchies des concepts dans le but d'assurer un gain de pertinence dans les SRIs, d'où l'apparition de *l'indexation sémantique* et de *l'indexation conceptuelle* en RI. Le document est ainsi représenté par un ensemble de concepts où un concept dénote un nœud dans une structure sémantique de type thésaurus ou ontologie, représentée par un ou plusieurs termes définis de manière non ambiguë. Ces structures peuvent être pré-existantes telles que WordNet ou MeSH [Díaz-Galiano *et al.*, 2008].

Dans le cadre de nos recherches, nous nous intéressons à coupler deux types de connaissances que nous pouvons atteindre par un processus d'ECT, à savoir les règles d'association entre termes qui représentent des connaissances implicites, et les ontologies qui traduisent plutôt des connaissances explicites relatives à un domaine. Le résultat de ce couplage est un réseau sémantique qui prend tout son intérêt dans une problématique d'indexation conceptuelle en RI [Ben Ghezaiel *et al.*, 2010].

Par ailleurs, l'ECT trouve aussi toute son importance dans le domaine de la Traduction Automatique Statistique (TAS). Divers travaux en TAS ont confirmé que les modèles basés sur *des séquences de mots* [Koehn *et al.*, 2003] permettent d'avoir des performances meilleures que ceux fondés sur les mots [Brown *et al.*, 1993]. Toutefois, dans le domaine de la TAS, il est indispensable d'utiliser des corpus d'apprentissage de très grande taille (de l'ordre de quelques centaines de milliers de phrases). Ce type de corpus représente un vrai challenge pour la communauté de l'ECT pour adapter les algorithmes de fouille de données à des contextes d'extraction textuels aussi volumineux et bruités.

De ce fait, l'ECT offre des techniques complémentaires pour contribuer à l'amélioration des modèles de TAS, à savoir : (*i*) l'exploration des séquences de mots par des méthodes de fouille de séquences [Dong and Pei, 2007] ; et, (*ii*) l'intégration de ces motifs séquentiels dans un modèle de traduction par le biais des règles d'association. Le croisement de la TAS avec le problème de l'extraction de motifs séquentiels est justifié par l'idée clé, propre à l'extraction de motifs séquentiels, permettant de distinguer à la fois, à l'intérieur des phrases du corpus, *un ordre d'apparition* des termes mais aussi de regrouper certains termes. Dans ce contexte, les règles d'association permettent l'extraction de règles *intra-phrases* alors que la recherche de motifs séquentiels permet l'extraction de règles *inter-phrases*. Ainsi, la notion de motifs séquentiels reste intuitivement applicable à la TAS, puisqu'il existe une relation d'ordre entre les termes dans les corpus parallèles, et par conséquent l'ordre d'apparition des termes dans une phrase peut être pris en compte.

En considérant l'ECT comme cadre fédérateur et les domaines connexes que nous avons cités, nous présentons, dans ce qui suit, un résumé des travaux menés dans le cadre de ma thèse de doctorat ainsi que le bilan des travaux de recherche post-thèse depuis 2005.

## 2 Résumé des travaux de recherche menés dans le cadre de la thèse de doctorat

Dans le cadre de ma thèse de doctorat [Latiri, 2004], je me suis intéressée à la technique de découverte de règles d'association (RA) à partir d'un contexte d'extraction textuel classique et flou.

En considérant l'Analyse Formelle de Concepts (AFC) comme fondement mathématique, j'ai proposé un algorithme, appelé Ice-Hasse, pour la construction du treillis de l'iceberg de Galois et un algorithme, appelé Gen-ra-re [Latiri *et al.*, 2003b], qui permet de générer les règles d'association non redondantes entre termes en explorant ce treillis. L'approche proposée a été validée par un ensemble d'expérimentions effectuées sur des collections textuelles de la campagne Amaryllis II [2].

Dans un cadre pratique, j'ai montré l'intérêt des règles d'association entre termes dans une problématique propre à la recherche d'information (RI), à savoir l'expansion de requêtes. Les tests l'approche d'expansion symbolique de requêtes moyennant les règles d'association entre termes, ont été menés sur deux collections Ofil et Inist de la campagne Amaryllis II. Les résultats expérimentaux ont montré une amélioration significative de la pertinence système d'un SRI expérimental [Latiri *et al.*, 2003c].

La dernière partie de la thèse a abordé l'extension de l'ensemble de mes propositions dans un contexte flou. Ceci m'a permis de m'orienter vers une nouvelle problématique de recherche, à savoir l'extraction de règles d'association floues entre termes et la définition d'un nouveau schéma de correspondance *requête-document* proposé pour la RI [Latiri *et al.*, 2002]. Ainsi, deux nouvelles extensions de la connexion de Galois floue ont été proposées [Latiri *et al.*, 2004]. Une approche d'expansion symbolique de requêtes avec les règles d'association floues entre termes a été également développée [Latiri *et al.*, 2003a].

## 3 Synthèse des travaux de recherche post-thèse (Depuis 2005)

Nos travaux de recherche se sont poursuivis durant les huit dernières années dans le domaine de l'ECT, avec un objectif bien précis, à savoir extraire d'autres motifs fréquents à partir de larges corpus textuels et montrer leur utilité dans le cadre d'applications réelles, telles que la RI ou la TAS.

Une telle démarche s'inscrit dans une double problématique : (*i*) définir les algorithmes adéquats pour la fouille de corpus de grandes tailles en prenant en compte le problème d'adaptation et d'optimisation du processus d'extraction de motifs intéressants ; et, (*ii*) le déploiement des connaissances découvertes dans des applications réelles manifestant des besoins et des défis différents.

2. Amaryllis est une Action de Recherche Concertée (ARC), organisée par l'Institut National français de l'Information Scientifique et Technique (INIST), avec le soutien de l'Agence Francophone pour l'Enseignement Supérieur et la Recherche (AUPELF-UREF) et le Ministère français de l'Education Nationale de la Recherche et de la Technologie (MERT). Deux cycles du projet ont déjà eu lieu, l'un en 1996-1997 et l'autre en 1998-1999. La méthodologie employée dans le projet Amaryllis est très proche de celle de Trec.

Sur l'ensemble de nos travaux de recherche, nous abordons et nous discutons la notion centrale de "*connaissance extraite à partir de textes*". Nous désignons par *connaissance* tout motif qui peut être découvert à partir d'un corpus textuel. Cette connaissance peut être déclinée en plusieurs motifs fréquents, tels qu'un ensemble de termes fréquents dans le corpus que nous appelons *termset*[3], *une séquence fréquente de termes* [Dong and Pei, 2007] ou encore *une règle d'association entre termes* appréciée par des mesures statistiques tels que le support et la confiance [Agrawal and Skirant, 1994]. En considérant une granularité textuelle variable au niveau de l'analyse du corpus, qui peut être, un document, une phrase ou un mot, nous nous intéressons ainsi aux relations existantes inter-granularités textuelles et intra-granularités textuelles, qui caractérisent un corpus et qui définissent une nouvelle représentation de ce dernier. Chaque forme de connaissance fait appel à une algorithmique dédiée à son extraction et trouve son usage et son intérêt dans des applications diverses liées à des domaines, qui présentent un intérêt pour de tels motifs textuels fréquents, comme par exemple la RI, l'IC ou encore la TAS.

Il importe de préciser, que dans le cadre de nos recherches, nous nous sommes intéressés principalement à l'application des fondements mathématiques de l'AFC [Wille, 1989] pour l'extraction de motifs fréquents à partir de textes. Ainsi, dans le contexte de la fouille de textes, l'AFC définit *un concept formel* par un ensemble d'objets, (*i.e.,* son *extension* est un ensemble de documents ou de phrases) auquel s'applique un ensemble d'attributs, (*i.e.,* son *intention* est un ensemble de termes ou de séquences de termes). Dans [Wille, 1989], Wille utilise la notion centrale de *treillis de concepts* ou *treillis de Galois* et l'applique tant à la découverte de concepts, qu'à l'acquisition de connaissances, et à la classification d'objets. De ce fait, dans le domaine de l'ECT, le treillis de Galois peut être vu comme un regroupement conceptuel et hiérarchique de documents (à travers les extensions du treillis), et interprété comme une représentation de toutes les implications entre les termes (à travers les intentions).

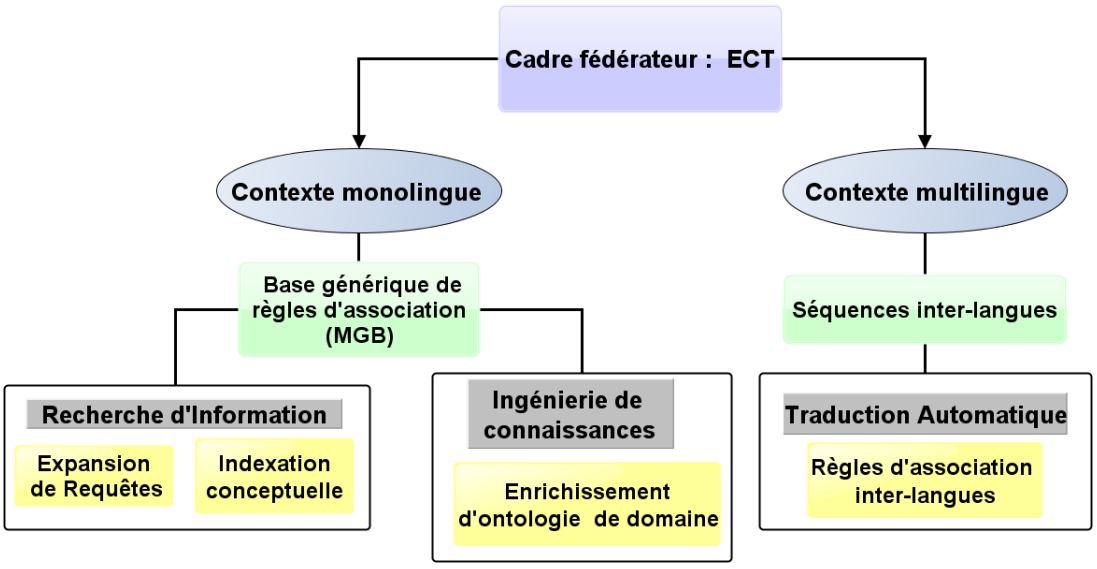


FIGURE 1 – Cadre de recherche et positionnement des contributions.

Comme le montre la FIGURE 1, nos Nos contributions de recherche s'articulent autour de deux

3. Par analogie à la terminologie *itemset* utilisée dans le domaine de data mining pour désigner un ensemble d'attributs.

problématiques. Tout d'abord, nous nous focalisons sur l'aspect algorithmique en utilisant les paradigmes de l'AFC [Wille, 1989] pour définir les motifs fréquents et les méthodes d'extraction à partir de textes. Nous proposons également une nouvelle base générique de règles d'association entre termes dédiée à l'ECT [Latiri *et al.*, 2012b]. Nous montrons, dans un deuxième temps, l'apport de l'utilisation de cette base de règles d'association dans deux applications liées à la RI, à savoir l'expansion de requêtes [Latiri *et al.*, 2012b] et l'indexation conceptuelle basée sur l'enrichissement d'une ontologie de domaine [Ben Ghezaiel *et al.*, 2010]. Nous étendons par la suite la définition de la base générique de règles d'association vers les règles d'association inter-langues (RAILs), où nous proposons d'extraire les séquences inter-langues fréquentes à partir d'un corpus parallèle de grande taille. Ces séquences sont ensuite utilisées pour définir un nouveau modèle de TAS à base de séquences [Latiri *et al.*, 2010b, Latiri *et al.*, 2011].

Pour chacune de nos propositions, nous nous attachons à définir les concepts associés et à développer les algorithmes permettant leur mise en œuvre. Tous ces travaux ont donné lieu à des évaluations sur des collections de test utilisées par la communauté RI ou celle de la TAS ou encore sur des bases de données synthétiques.

Nos différents travaux de recherche s'intègrent dans trois principaux axes complémentaires que nous allons présenter ci-dessous.

## 3.1 Axe 1 : Base générique de règles d'association entre termes (à partir de 2005)

*Les premières réflexions dans cet axe de recherche ont été abordées dans le cadre du mastère de recherche de Melle. Lamia Ben Ghezaiel (Soutenu en 2006), sous la direction du Pr. Mohamed Ben Ahmed (ENSI, Université de la Manouba - Tunisie) et moi même, et ensuite étendues en 2010 par de nouvelles réflexions personnelles.*

La découverte de motifs fréquents à partir de corpus de textes demeure le noyau central de nos recherches. Parmi ces motifs, nous distinguons les ensembles de termes fréquents, appelés *termsets fréquents*. Ainsi, une collection de documents peut être définie comme une famille de termsets fréquents, issus de l'ensemble des termes d'indexation. La découverte des termsets fréquents permet la génération de corrélations entre termsets, appelées *règles d'association*. Cependant, au delà d'une simple évaluation de corrélation entre termsets, une règle d'association lie fortement deux termsets distincts $T_i$ et $T_j$, qui constituent respectivement sa prémisse et sa conclusion. De ce fait, une règle traduit la probabilité d'avoir les termes de la conclusion dans un document, sachant que ceux de la prémisse y sont.

Toutefois, l'application des règles d'association dans le contexte de la RI ou de la TAS est loin d'être une tâche triviale, étant donné le nombre très important de règles potentiellement intéressantes qui peuvent être découvertes à partir d'une collection de documents. De plus, l'extraction des corrélations entre termes nécessite l'analyse de tous les textes d'une collection, qui est aussi une phase nécessitant des temps de calcul coûteux et des espaces mémoire assez conséquents. La taille des collections de documents représente ainsi un défi majeur pour les chercheurs du domaine de l'ECT.

Durant la dernière décennie, des techniques avancées, qui s'appuient sur les fermetures de la connexion de Galois, ont émergé pour pallier au problème de redondance des règles d'association. Ces techniques représentent une alternative permettant de réduire considérablement le coût de l'extraction des termsets fréquents et d'éliminer la redondance au sein de l'ensemble des règles d'association. Elles définissent des sous-ensembles réduits de l'ensemble des règles d'association

valides, appelés *représentations concises de règles d'association*, connues sous la désignation de *bases génériques* [Bastide *et al.*, 2000a]. En effet, ces techniques sont basées sur une partition de l'espace de recherche en classes d'équivalence disjointes dont les éléments partagent les mêmes caractéristiques. L'élément maximal dans une classe d'équivalence donnée est appelé *motif fermé*, tandis que les éléments minimaux sont appelés *générateurs minimaux* [Bastide *et al.*, 2000a]. La littérature récente a présenté des travaux qui ont donné des résultats pertinents quant à la compacité des bases génériques et dont l'impact sur la réduction du nombre des règles d'association découvertes a été prouvé [Ben Yahia *et al.*, 2009, Balcázar, 2010].

Étant donné que notre objectif principal est d'extraire des règles d'association entre termes à partir de corpus de textes volumineux, nous avons proposé la formalisation et l'extraction d'une nouvelle base générique minimale de règles d'association non-redondantes entre termes, appelée $\mathcal{MGB}$, permettant d'assurer un compromis entre l'*informativité* et la *compacité* de cette dernière [Latiri *et al.*, 2012b]. La spécificité de la base proposée est qu'elle est compacte, dans le sens où elle englobe un noyau minimal de règles d'association entre termes, qui sont approximatives et exactes sans aucune redondance. Ces règles sont dérivées à partir de la structure ordonnée du *treillis d'Iceberg de Galois augmenté*, *i.e.*, la partie supérieure du treillis de Galois ne conservant que les termsets fermés fréquents et qui sont à leur tour "décorés" par l'ensemble de leurs générateurs minimaux [Latiri *et al.*, 2006]. La caractéristique clé des règles dérivées est qu'elles ont des prémisses minimales, représentées par les générateurs minimaux, et des conclusions maximales. Dans ce contexte, les générateurs sont utilisés dans les prémisses des règles d'association découvertes, tandis que les termsets fermés fréquents permettent la dérivation des conclusions de ces règles. Il importe de souligner que la relation de précédence, définie dans le treillis d'Iceberg de Galois, aide à réduire le coût d'extraction des règles d'association, en évitant certaines combinaisons redondantes lors de la génération des candidats.

L'étude empirique que nous avons menée sur cinq collections de documents de la campagne d'évaluation Amaryllis II et la campagne CLEF 2003 [4] a fait ressortir que la base $\mathcal{MGB}$ apporte des gains substantiels en terme de compacité par rapport aux bases génériques extraites avec et sans perte d'information. Nous avons discuté dans [Latiri *et al.*, 2012b] les résultats des évaluations expérimentales par rapport à la nature du contexte d'extraction textuel, *i.e.,* dense et épars.

**Publications engendrées :**

- **C. Latiri**, H. Haddad and T. Hamrouni. Towards An Effective Automatic Query Expansion Process Using An Association Rule Mining Approach. *Journal of Intelligent Information Systems*, Volume 39, Issue 1, pages 209-247, August 2012, Springer.
- **C. Latiri**, L. Ben Ghezail and M. Ben Ahmed. Fast-MGB : Nouvelle base générique minimale pour l'extraction de règles d'association *Actes des Journées francophones EGC'2006, Revue des Nouvelles Technologies de l'Information, RNTI-E-6, pages 217-222, Lille, 17-20 Janvier 2006.*
- **C. Latiri**, W. Bellagha, and S. Ben Yahia. VIE-MGB : A Visual Interactive Exploration of Minimal Generic Basis of Association Rules. *In Proceedings of the* $3^{rd}$ *International Conference on Concept Lattices and their Applications, CLA'05, Olomouc, Czech Republic, pages 179-196, September, 7-9, 2005.*

4. Le "Cross-Language Evaluation Forum" (CLEF) offre des collections de données afin d'évaluer les systèmes de recherche d'information (*http ://www.clef-campaign.org/*).

## 3.2 Axe 2 : Règles d'association entre termes et ontologie au service de la RI (à partir de 2007)

Les contributions relatives à cet axe de recherche se situent dans le croisement de l'ECT et de la RI. L'étude proposée est centrée autour de l'utilisation de la base générique de règles d'association entre termes $\mathcal{MGB}$ pour l'amélioration des résultats de la recherche d'information.

Principalement, dans le domaine de la RI, la *pertinence* reste la notion centrale. Or, dans la pratique, les SRIs montrent des décalages importants entre la *pertinence utilisateur* et la *pertinence système*. Ces derniers sont liés essentiellement à deux problématiques fondamentales en RI : (*i*) l'inadéquation des documents restitués avec la requête originelle de l'utilisateur ; et, (*ii*) l'imperfection de l'indexation automatique des documents. Il va sans dire que ces deux problématiques sont corrélées.

Nos contributions dans cet axe de recherche abordent, d'une part la problématique d'interrogation en RI à travers l'utilisation de la base générique $\mathcal{MGB}$ dans une approche d'expansion automatique de requêtes [Latiri *et al.*, 2012b] et, d'autre part, la problématique d'indexation en RI par la proposition d'une nouvelle approche d'indexation conceptuelle basée sur une ontologie de domaine enrichie par la dite base [Ben Ghezaiel *et al.*, 2010, Ben Ghezaiel *et al.*, 2012, Latiri *et al.*, 2012a].

### Approche d'expansion automatique de requêtes par la base générique de règles d'association entre termes

*L'étude menée dans cet axe de recherche est le fruit d'un travail personnel, soutenu par quelques réflexions apportées par les co-auteurs de l'article [Latiri* et al.*, 2012b].*

Intuitivement, une règle d'association traduit la probabilité d'avoir les termes de la conclusion dans un document, sachant que ceux de la prémisse y sont. L'utilisation de telles dépendances, dans un processus d'expansion de requêtes, peut améliorer sensiblement la pertinence d'un SRI, car elles reflètent des corrélations fortes et implicites découvertes à partir de la collection de documents [Lin *et al.*, 2008].

Nous avons proposé un nouveau processus d'expansion automatique de requêtes moyennant la base générique $\mathcal{MGB}$ [Latiri *et al.*, 2012b]. Dans un premier temps, il s'agit de dériver la base générique $\mathcal{MGB}$ de règles d'association non-redondantes entre termes à partir d'une collection de documents, et de l'utiliser, dans un deuxième temps, pour étendre la requête originelle de l'utilisateur. Il importe de souligner que la base générique $\mathcal{MGB}$ est mieux adaptée au processus d'expansion automatique de requêtes, dans lesquelles l'ensemble des termes originels sera étendu par les conclusions des règles valides de la base générique $\mathcal{MGB}$, et ayant les termes de la requête initiale dans leurs prémisses respectives. Rappelons que les règles d'association entre termes de la base $\mathcal{MGB}$ sont non-redondantes et qu'elles sont dotées d'une prémisse minimale et d'une conclusion maximale, ce qui offre plus de termes candidats pour l'expansion.

Notre approche automatique d'expansion de requêtes peut donc être considérée comme une approche plus élaborée basée sur les co-occurrences de termes. Ceci est justifié par le fait qu'une règle d'association entre termes infère une relation globale entre les termes, qui ne dépend pas d'un document donné mais implique plutôt un ensemble de documents de la collection et caractérisant un ensemble de termes liés, *i.e.,* un termset. Dans ce cadre, contrairement à la technique d'analyse locale de co-occurrences de termes qui permet de dériver les corrélations entre les termes d'un document donné, *i.e.,* relations *intra-document*, les règles d'association entre termes

se dérivent par un processus de fouille globale de toute la collection de documents et offrent par conséquent des informations sur les corrélations de termes *inter-documents* ainsi que sur les relations *intra-document*. De ce fait, les règles d'association permettent d'expliciter des relations plus fines entre les termes que les approches classiques à base de co-occurrences de termes.

Pour valider l'approche proposée, nous avons mené des tests, moyennant le SRI expérimental LEMUR [5], et en utilisant trois schémas de pondération : $tf \times idf$, BM25$tf$ et OKAPI BM25. Nous avons réalisé nos différents tests sur les collections OFIL et INIST de la deuxième campagne d'évaluation AMARYLLIS, ainsi que sur les collections LE MONDE 94 et ATS 94 de la campagne d'évaluation CLEF 2003 (Collection 2001). La cinquième collection que nous avons également exploitée est composée conjointement des deux collections LE MONDE 94 et ATS 94.

**Nouvelle approche d'indexation conceptuelle**

*La problématique étudiée dans cet axe de recherche s'inscrit dans le cadre de la thèse de doctorat de Melle. Lamia Ben Ghezaiel (Thèse déposée auprès de l'école doctorale de l'ENSI le 14 décembre 2012), sous la direction du Pr. Mohamed Ben Ahmed (ENSI, Université de la Manouba - Tunisie) et moi même.*

Après avoir montré l'intérêt de l'utilisation de la base générique de règles d'association entre termes $\mathcal{MGB}$ dans l'expansion de requêtes en RI, nous avons proposé une deuxième contribution qui s'adresse plutôt à une problématique largement abordée en RI, à savoir l'imperfection de l'indexation automatique des documents. Nous avons introduit une nouvelle approche d'indexation conceptuelle en RI qui est fondée sur l'enrichissement d'une ontologie de domaine $\mathcal{O}$, à travers l'utilisation de règles d'association de la base générique $\mathcal{MGB}$ [Ben Ghezaiel *et al.*, 2010, Ben Ghezaiel *et al.*, 2012, Latiri *et al.*, 2012a]. Dans le cadre de nos recherches, nous abordons la question de l'utilisation des ontologies pour la RI du point de vue de l'ingénierie des connaissances, en particulier sur leur mode d'enrichissement et sur la nature de leur contenu en terme de connaissances représentées. Nous soutenons l'idée que les ontologies sont des représentations de connaissances pertinentes en RI du fait qu'elles comportent une dimension sémantique, et qu'elles peuvent être rattachées à d'autres motifs fréquents issus de l'ECT, telles que les règles d'association entre termes.

Le processus général d'enrichissement d'ontologie que nous avons proposé [Ben Ghezaiel *et al.*, 2010] considère en entrée une ontologie de domaine existante $\mathcal{O}$ et une collection de documents associée au même domaine. Le résultat généré par le processus d'enrichissement d'ontologie, à base de règles d'association, est exploré en tant que *réseau conceptuel proxémique* [Ben Ghezaiel *et al.*, 2011, Ben Ghezaiel *et al.*, 2012]. L'originalité de ce réseau réside dans sa généricité et son exhaustivité, qui sont dues à la combinaison des connaissances explicites de la structure ontologique d'une part, et des connaissances implicites issues de l'application de la technique d'extraction de règles d'association, d'autre part.

Du point de vue applicatif, nous avons proposé d'utiliser le réseau conceptuel proxémique en RI, par la proposition d'une nouvelle approche d'indexation conceptuelle. En effet, nous suggérons d'utiliser les relations entre les concepts du réseau proxémique, qui sont déjà pondérées par une nouvelle mesure de similarité sémantique que nous avons définie dans le processus de désambiguïsation des termes d'un document lors de phase d'indexation. Puis, nous suggérons de l'utiliser pour affecter un poids sémantique aux différents descripteurs des documents.

5. http ://www.lemurproject.org/

L'approche d'indexation conceptuelle de documents proposée comprend trois phases principales [Ben Ghezaiel *et al.*, 2011, Ben Ghezaiel *et al.*, 2012] : (*i*) identifier et pondérer les concepts représentatifs d'un document moyennant de nouvelles définitions de la représentativité statistique et de la représentativité sémantique ; (*ii*) sélectionner le meilleur sens relatif à un concept, et ce par la proposition d'une nouvelle approche de désambiguïsation ; et, (*iii*) construire le réseau sémantique propre à un document, noté par *Doc-*$\mathcal{O}_{\mathcal{MGB}}$.

Il importe de souligner que chaque nœud du réseau *Doc-*$\mathcal{O}_{\mathcal{MGB}}$ représente un champ de proximité sémantique et conceptuel, synthétisant trois sémantiques [Bachimont, 2000] : (*i*) une sémantique *référentielle* résultante du processus de désambiguïsation ; (*ii*) une sémantique *différentielle* induite par le calcul de voisinage du concept ; et, (*iii*) une sémantique *inférentielle* assurée grâce à l'enrichissement de l'ontologie par les règles d'association de la base $\mathcal{MGB}$.

Partant de cette structure, nous passons d'un d'index simple composé de mono-termes à un index conceptuel représenté par un réseau conceptuel proxémique tridimensionnel, traduisant le contenu sémantique d'un document [Ben Ghezaiel *et al.*, 2012].

Nous avons évalué notre approche d'indexation conceptuelle dans le domaine de la recherche d'information biomédicale. Pour ce faire, nous avons considéré comme ontologie, la structure poly-hiérarchique du thésaurus MeSH (Medical Subject Headings) [6] dans le processus d'enrichissement proposé et également pour désambiguïser les concepts dans les documents. Nous avons aussi utilisé la collection OHSUMED, proposée dans le cadre de la tâche TREC9-Filtering en 2000 et qui est constituée de titres et/ou résumés de 270 journaux médicaux publiés entre 1987-1991, extraits de la base MEDLINE [7]. L'objectif de notre évaluation expérimentale est double : d'une part, il s'agit d'étudier l'apport de l'enrichissement de l'ontologie par les nouveaux concepts issus des règles d'association de la base $\mathcal{MGB}$ ; et, de mesurer l'impact de l'approche de désambiguïsation proposée sur la performance de l'indexation, d'autre part [Ben Ghezaiel *et al.*, 2012, Latiri *et al.*, 2012a].

**Publications engendrées :**

- L. Ben Ghezaiel, **C. Latiri** and M. Ben Ahmed. Conceptual Indexing in IR based on Ontology Enrichment and Association Rules. *Submitted in Transactions on Large-Scale Data and Knowledge-Centered Systems, November 20, 2012, LNCS Journal Subline,LNCS Journal Subline Springer.*
- **C. Latiri**, L. Ben Ghezaiel and M. Ben Ahmed. Proxemic Conceptual Network based on Ontology Enrichment for Representing Documents in IR. *Proceedings of the* $18^{th}$ *International Conference on Knowledge Engineering and Knowledge Management (EKAW 2012), Galway City, Ireland, 8-12 October 2012*, Volume 7603 of LNAI, pages 72-86, Springer.
- L. Ben Ghezaiel, **C. Latiri** and M. Ben Ahmed. Ontology Enrichment based on Generic Basis of Association Rules for Conceptual Document Indexing. *In Proceedings of the* $4^{th}$ *the International Conference on Knowledge Engineering and Ontology Developpement (KEOD 2012), Barcelona, Spain, October, 4-7, 2012.*
- L. Ben Ghezaiel, **C. Latiri** and M. Ben Ahmed. Conceptual Indexing Documents in IR based on Ontology Enrichment. *Proceedings of the* $16^{th}$ *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, special session on Ontology-Based Information Retrieval (KES 2012), In Advances in Knowledge-Based and*

6. MeSH : *Medical Subject Heading* contient le vocabulaire contrôlé de la NLM (National Library of Medicine) utilisé pour indexer les articles et faire des recherches dans les bases de données indexées par MeSH dont MEDLINE.

7. http ://www.ncbi.nlm.nih.gov/pubmed/

*Intelligent Information and Engineering Systems M. Graña et al. (Eds.), pages 1920-1931, San Sebastian, Spain, September, 10-12, 2012*, IOS Press.

- L. Ben Ghezaiel, **C. Latiri**, M. Ben Ahmed and N. Khouja. Un réseau proxémique pour la recherche d'information : Application à la maladie des dystonies. *Actes de la première édition du Symposium sur l'Ingénierie de l'Information Médicale SIIM 2011, pages 25-40, Toulouse, France, 9-10 Juin 2011.*
- L. Ben Ghezaiel, **C. Latiri**, M. Ben Ahmed and N. Khouja. Enrichissement d'ontologie par une base générique minimale de règles d'association. *Actes de la Conférence en Recherche d'Information et Applications CORIA'2010, pages 289-300, 18-20 Mars 2010, Sousse, Tunisie.*
- **C. Latiri**, M. Mtir, and S. Ben Yahia. Méthode de construction d'ontologie de termes à partir du treillis de l'icerberg de Galois. *Actes des Journées francophones EGC'2005, Revue des Nouvelles Technologies de l'Information, RNTI-E-3, pages 365-376, Paris, 18-21 Janvier, 2005.*

### 3.3 Axe 3 : Règles d'association inter-langues pour la Traduction Automatique Statistique (à partir de 2008)

Notre intérêt pour le domaine de la Traduction Automatique Statistique (TAS) est justifié par le fait que des travaux pionniers en TAS ont confirmé que les modèles fondés sur *les séquences de mots* [Koehn *et al.*, 2003] permettent d'avoir des performances meilleures que ceux à base de mots [Brown *et al.*, 1993]. En utilisant des séquences de mots, les systèmes de traduction parviennent à préserver certaines contraintes locales sur l'ordre des mots.

Notre contribution émane ainsi de notre conviction qu'il est possible de proposer d'autres approches issues du domaine de l'ECT, capables de constituer une alternative aux méthodes d'IBM [Brown *et al.*, 1993] et d'améliorer la pertinence des modèles de traduction à base de séquences [Koehn *et al.*, 2003]. Notre objectif est d'étendre les approches d'extraction des motifs fréquents sur des corpus parallèles de très grande taille et d'extraire des connaissances utiles et pertinentes pour la TAS.

L'originalité de notre proposition repose ainsi sur le couplage de deux types de connaissances issues par un processus d'ECT, à savoir les règles d'association entre termes [Latiri *et al.*, 2012b] et les séquences fermées fréquentes de termes [Dong and Pei, 2007]. Ce couplage donne naissance à la définition d'un nouveau concept que nous appelons *Règles d'Association Inter-Langues* (RAILs) [Latiri *et al.*, 2010b]. Ce concept n'est autre qu'une extension de la notion de règle d'association telle qu'elle a été définie par Agrawal et *al.* [Agrawal and Skirant, 1994]. Intuitivement, l'interprétation d'une règle inter-langues dans le domaine de la TAS est que sa *conclusion*, représentant une unité linguistique dans une langue cible, est une traduction potentielle de sa *prémisse*, représentant une autre unité linguistique dans la langue source. Cette règle est appréciée par deux métriques, à savoir le *support* qui mesure sa fréquence d'apparition dans le corpus parallèle, et la *confiance* qui mesure sa validité.

Pour concrétiser l'ensemble de nos propositions, nous avons abordé la problématique d'extraction de séquences fréquentes dans le domaine l'ECT en prenant en compte l'ordre d'apparition des mots dans la phrase, et ce à partir de corpus de textes parallèles alignés au niveau de la phrase. Nous nous sommes intéressés particulièrement à la famille des approches dédiées à l'extraction *des motifs séquentiels fermés fréquents* [Chang, 2004], inspirées de la recherche d'itemsets fermés fréquents [Pasquier *et al.*, 1999]. Compte tenu de la densité des corpus parallèles et afin d'éviter de dériver un nombre très élevé de séquences, nous avons procédé à l'extraction d'un ensemble compact de séquences, *i.e.,* l'ensemble des séquences *fermées* fréquentes, tels que

leurs supports soient supérieurs ou égaux à un support minimal fixé. Généralement, dans les applications d'ECT, le seuil de support minimal *minsupp* est déterminé expérimentalement suite à l'étude de la distribution *zipfienne* des termes du corpus.

La deuxième contribution concerne la proposition d'un nouveau modèle de traduction faisant appel aux séquences fermées fréquentes et aux règles d'association inter-langues extraites à partir d'un corpus parallèle [Latiri *et al.*, 2011]. En effet, le principal avantage du modèle de traduction à base de règles d'association inter-langues est qu'il ne nécessite pas d'alignement, contrairement à ce qui se fait généralement dans la communauté de TAS. Il importe de souligner que l'alignement en TAS demeure une tâche fastidieuse faisant appel à des algorithmes complexes.

Ainsi, la mise en place d'un nouveau modèle de traduction à base de règles d'association inter-langues implique nécessairement la construction d'une table de traduction à partir des RAILs [Latiri *et al.*, 2010b]. Afin d'évaluer notre modèle de TAS, nous avons utilisé le décodeur PHARAOH [Koehn, 2004] pour mener une série de tests sur le corpus parallèle EUROPARL [Rama and Borin, 2011] en choisissant comme langues source et cible, respectivement, le français et l'anglais. Chaque stratégie d'évaluation est comparée à l'approche de référence de Koehn et *al.* [Koehn *et al.*, 2003] qui consiste à extraire la table de traduction de séquences à partir de l'alignement bi-directionnel de mots, et également à l'approche basée sur les triggers inter-langues [Lavecchia *et al.*, 2008], représentant la genèse de notre approche.

*Les évaluations expérimentales menées dans cet axe de recherche ont été élaborées dans le cadre du mastère de recherche de Melle. Cyrine Nasri (Soutenu en 2009), sous la direction de M. Yahya Slimani (Encadreur, ISAMM, Université de la Manouba) et moi même.*

**Publications engendrées :**

- **C. Latiri**, K. Smaïli, C. Lavecchia, D. Langlois and C. Nasri.Phrase-based machine translation based on text mining and statistical language modeling techniques. *In Proceedings of* $12^{th}$ *International Conference on Intelligent Text Processing and Computational Linguistics, CICLING'2011, Tokyo (Japan), February, 20-26, 2011.*, special issue of the *International Journal of Computational Linguistics and Applications, IJCLA journal*, Vol. 2, $N°$. 1-2, JAN-DEC 2011, pages 193-208, BAHRI Publications.
- **C. Latiri**, K. Smaïli, C. Lavecchia and D. Langlois. Mining Monolingual and Bilingual Corpora. *Intelligent Data Analysis Journal*, Volume 14(6) :663-682, November, 2010.
- **C. Latiri**, C. Nasri, K. Smaïli et Y. Slimani. Extraction des séquences fermées fréquentes à partir de corpus parallèles : Application à la traduction automatique. *Actes des Journées francophones EGC'2010, Revue des Nouvelles Technologies de l'Information, RNTI-E-19, pages 55-60, Hammamet, Tunisie, 26-29 Janvier 2010.*

## 3.4 Autres Contributions (à partir de 2008)

### Extraction des séquences et des triggers inter-langues pour la traduction automatique à base de l'information mutuelle conditionnelle multivaluée

*La problématique étudiée dans cet axe de recherche s'inscrit dans le cadre de la thèse de doctorat en co-tutelle de Melle. Cyrine Nasri, sous la direction de M. Kamel Smaïli (Université de Nancy 2, France), M. Yahya Slimani (Encadreur, ISAMM, Université de la Manouba, Tunis) et moi même. La soutenance de cette thèse est prévue pour fin 2014.*

Cet axe de recherche s'inscrit dans la continuité du modèle de traduction statistique à base des triggers inter-langues proposé dans [Lavecchia *et al.*, 2008]. Il s'agit d'une approche originale qui ne nécessite aucun alignement des mots au sein des corpus parallèles. Les triggers inter-langues permettent de mettre en évidence des unités fortement corrélées en se basant sur l'Information Mutuelle (IM). Les unités linguistiques considérées sont des séquences de mots sources et cibles. L'idée derrière le concept de triggers inter-langues est que si une séquence de mots sources est fortement corrélée à une séquence de mots cibles en terme d'IM, alors la présence de la première dans une phrase source déclenchera la présence de la seconde dans sa traduction et vice versa. L'utilisation des triggers inter-langues, sur les corpus parallèles, permet en effet de trouver les traductions possibles de séquences de mots et ainsi constituer une table de traduction [Lavecchia *et al.*, 2008].

Il importe de préciser que l'IM est une mesure de co-occurrence qui se calcule simplement en un seul passage sur le corpus parallèle. Pour sélectionner les triggers inter-langues, la méthode à base des triggers inter-langues suppose que deux séquences sources et cibles co-occurrent si elles apparaissent dans une même paire de phrases du corpus parallèle. De ce fait, elle ne requiert qu'un alignement au niveau des phrases et non au niveau des mots au sein du corpus parallèle, contrairement à l'approche pionnière de l'état de l'art [Koehn *et al.*, 2003].

D'une manière simplifiée, l'identification des séquences, dans le cadre de cette approche, opère par un procédé itératif pour identifier une suite de mots susceptible de présenter un intérêt selon l'IM. Ensuite, ces mots sont concaténés pour former une première séquence de taille 2 et le corpus est réécrit ainsi selon les nouvelles séquences. Les itérations suivantes permettent par conséquent de trouver des séquences de longueur 3, et ainsi de suite jusqu'à atteindre la condition d'arrêt.

Par ailleurs, pour éviter la propagation des erreurs dues à ces multiples étapes et afin d'améliorer les résultats présentés dans [Lavecchia *et al.*, 2008], nous proposons, dans le cadre de cet axe de recherche, d'utiliser l'Information Mutuelle Conditionnelle (IMC) pour extraire les séquences [Nasri *et al.*, 2011]. Nous suggérons également de généraliser l'information mutuelle conditionnelle à $n$ variables aléatoires pour la sélection des séquences inter-langues, permettant ainsi de ne pas s'arrêter à des séquences de longueur 3. Notre objectif étant d'aboutir à un modèle de traduction statistique basé sur l'information mutuelle conditionnelle et sur un modèle de langage à base de séquences au lieu d'un modèle $n$-grammes.

Les premières évaluations expérimentales ont été menées sur le corpus parallèle Europarl et moyennant le décodeur Moses [Koehn *et al.*, 2007]. Les scores BLEU trouvés sont encourageants et s'approchent du score de référence de l'état de l'art [Nasri *et al.*, 2011].

**Publication engendrée :**

- C. Nasri, K. Smaïli and **C. Latiri**. A new method for learning Phrase Based Machine Translation with Multivariate Mutual Information. *In Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE'12) , Hefei, China, September, 20-24, 2012.*
- C. Nasri, K. Smaïli and **C. Latiri**. Training Statistical Machine Translation with Multivariate Mutual Information. *In Proceedings of* $5^{th}$ *Language and Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics, LTC 2011, November, 25-27, 2011, Poznan, Poland.*

**Nouvelle approche de fouille de graphes**

*La problématique étudiée dans cet axe de recherche s'inscrit dans le cadre de la thèse de doctorat en co-tutelle de M. Brahim Douar, sous la direction de M. Michel Liquière (Université de Montpellier 2, France), M. Yahya Slimani (Encadreur, ISAMM, Université de la Manouba) et moi même. La Thèse a été soutenue le 27 Novembre 2012.*

Parallèlement à mes travaux de recherche liés à l'ECT, j'ai contribué dans une problématique dédiée à la fouille de données structurées et en particulier, la *fouille de graphes* [Kuramochi and Karypis, 2001]. Partant d'une représentation de sources de données sous forme de graphes, il s'agit de rechercher des sous-structures ou de règles, apparaissant dans ces graphes.

Avec la croissance importante du besoin d'analyser une grande masse de données structurées tels que les composés chimiques, les structures de protéines ou même les documents XML, la fouille de de sous-graphes fréquents est devenue un défi réel en matière de fouille de données. Ceci est étroitement lié à leur nombre exponentiel ainsi qu'à la NP-complétude du problème d'isomorphisme d'un sous-graphe général.

Face à cette complexité, notre objectif est de proposer une approche qui évite de recourir à l'isomorphisme de sous-graphes afin d'assurer une meilleure mise à l'échelle par rapport aux approches classiques.

Les travaux de l'état de l'art ont montré que les algorithmes de fouille de graphes utilisent comme opérateur de projection l'isomorphisme de sous-graphes [Nijssen and Kok, 2005]. Ces algorithmes se basent sur deux paradigmes de parcours de l'espace de recherche, à savoir le parcours en largeur et le parcours en profondeur. Leur but consiste à trouver les sous-graphes connectés ayant un nombre suffisant d'arêtes dans un seul graphe éparse non dirigé. La plupart de ces algorithmes utilisent des méthodes différentes pour énumérer et générer les motifs candidats. Une comparaison quantitative des approches de fouille de graphes est donnée dans [Wörlein *et al.*, 2005]. L'approche de fouille de graphes que nous avons proposée dans [Douar *et al.*, 2011a] se base sur une approche en largeur, inspirée de l'algorithme Apriori [Agrawal and Skirant, 1994] et nommée Fsg [Kuramochi and Karypis, 2001].

Notre approche, baptisée Fgmac acronyme de "Frequent subGraph Mining with Arc Consistency", introduite dans [Douar *et al.*, 2011b, Douar *et al.*, 2011a], tire son originalité du domaine de la programmation par contraintes (CSP) et plus précisément de sa technique de filtrage local qui est la technique de consistance d'arc. Nous avons ainsi introduit la notion de biais de projection afin de proposer un opérateur similaire à l'isomorphisme de sous-graphes, mais ayant une complexité polynomiale et des contraintes relaxées [Douar *et al.*, 2011c]. L'AC-projection, initialement introduite dans [Liquiere, 2007], propose un opérateur de projection basé sur l'algorithme de la consistance d'arc, issu du domaine de la programmation par contraintes. Cette méthode de projection possède des propriétés intéressantes, à savoir la polynomialité, la validation locale, la parallélisation et l'interprétation structurelle.

Nous avons ensuite utilisé cet opérateur dans un processus de fouille de graphes. En effet, comme première contribution dans cet axe de recherche, nous avons changé la fonction dédiée pour le calcul du support d'un motif, dans la mesure où l'AC-projection est désormais utilisée à la place de l'isomorphisme de sous-graphes pour vérifier si un graphe candidat apparaît dans une transaction donnée ou pas [Douar *et al.*, 2011c]. L'algorithme fgmac que nous avons proposé dans [Douar *et al.*, 2011a] commence par énumérer tous les graphes fréquents ayant une ou deux arêtes. Ensuite, un processus itératif est amorcé. Lors de chaque itération, il commence par générer les graphes candidats ayant une taille supérieure d'une seule arête par rapport aux fréquents de l'itération précédente. Il calcule par la suite la fréquence de chaque graphe

moyennant l'AC-projection comme opérateur de projection et non pas l'isomorphisme de sous-graphes. L'algorithme élague les sous-graphes qui ne satisfont pas la contrainte du support minimal. La particularité de l'algorithme FGMAC est de ne retourner que les graphes AC-réduits fréquents, qui ne représentent qu'un sous-ensemble des graphes fréquents à l'isomorphisme près [Douar *et al.*, 2011a].

La deuxième contribution dans cette problématique concerne la proposition d'un algorithme de fouille de graphes, appelé AC-MINER [Douar *et al.*, 2011b], qui utilise des opérations ensemblistes et des opérations sur le voisinage. Ceci nous permet d'avoir une approche de fouille de graphes AC-réduits fréquents, pouvant aisément être adaptée à un parcours en profondeur ou en largeur de l'espace de recherche. L'originalité de cet algorithme est que toutes les opérations utilisées sont directement issues des propriétés relatives à l'AC-projection. En le comparant à FGMAC, l'algorithme AC-MINER bénéficie d'un espace de recherche largement inférieur et d'un parcours en profondeur de cet espace. Cependant, l'algorithme FGMAC explore le même espace de recherche que celui qui est exploré par les approches classiques de fouille de graphes, tout en profitant de la polynomialité de l'opérateur de projection, *i.e.,* l'AC-projection, lors du calcul du support. Nous avons prouvé que la borne supérieure de cet espace serait égal à $2^{n^2}$, $n$ étant le nombre de sommets du graphe contenant l'ensemble des transactions.

Dans le but de montrer l'intérêt de recourir à l'AC-projection pour la fouille de graphes et souligner son apport, nous avons effectué une évaluation qualitative des motifs AC-réduits qui consiste à mesurer le pouvoir discriminant de ces motifs dans un processus de classification supervisée de graphes en terme de *pcc* et de *AUC* [Douar *et al.*, 2011a]. Pour nos évaluations expérimentales, nous avons sélectionné cinq bases de graphes largement citées dans la littérature, à savoir PTC-FM, PTC-FR, PTC-MM, PTC-MR et HIA [Smalter *et al.*, 2008]. Les évaluations expérimentales ont montré que le nombre de motifs fréquents découverts, par les algorithmes FGMAC et AC-MINER, est inférieur aux motifs à l'isomorphisme près découverts par l'algorithme FSG pour tous les supports. De même, les valeurs de *pcc* trouvées sont comparables, voire meilleures, pour certains scénarios que les valeurs de référence de l'état de l'art [Douar *et al.*, 2011a].

**Publications engendrées :**

- B. Douar, **C. Latiri**, M. Liquière and Y. Slimani. A projection bias in frequent subgraph mining can make a difference. ***Under revision*** *in International Journal on Artificial Intelligence Tools, 20 May 2012, World Scientific Publishing Company.*
- B. Douar, M. Liquière, **C. Latiri** and Y. Slimani. LC-MINE : a framework for frequent subgraph mining with local consistency techniques. *Knowledge and Information Systems Journal*, Springer (To appear in 2013).
- B. Douar, **C. Latiri**, M. Liquière, and Y. Slimani. B-FGMAC : Breadth-first Frequent sub-Graph Mining with Arc Consistency. *Journal of Artificial Intelligence and Soft Computing Research*, Volume 1, Number 4, pages 269-281, Polish Neural Network Society, 2011.
- B. Douar, M. Liquière, **C. Latiri** and Y. Slimani. Graph-based relational learning with a polynomial time projection algorithm. *In Proceedings of the* $21^{st}$ *International Conference on Inductive Logic Programming, ILP 2011,* volume 7207 of LNAI, pages 96-112, Springer.
- B. Douar, **C. Latiri**, M. Liquière and Y. Slimani. FGMAC : Frequent SubGraph Mining with Arc Consistency. *In Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2011, part of the IEEE Symposium Series on Computational Intelligence 2011, Paris (France), pages 112-119, April, 11-15, 2011.*
- B. Douar, M. Liquiere, **C. Latiri** and Y. Slimani. Nouvelle approche de fouille de graphes

AC-réduits fréquents. *Actes des Journées francophones EGC'2011, Revue des Nouvelles Technologies de l'Information, RNTI-E-20, pages 473-478, Brest (France), 25-28 Janvier 2011.*

– B. Douar, M. Liquière, **C. Latiri** and Y. Slimani. Un biais de projection peut faire la différence dans la fouille de graphes. *Première Journée de fouille de grands graphes, IRIT, Toulouse (France), 13 Octobre 2010.*

# 4 Publications scientifiques

Mon activité de publication scientifique après la thèse de doctorat englobe 25 articles et se résume comme suit :

– 5 articles publiés dans des revues internationales spécialisées dont 3 avec impact factor et référencés sur DBLP et/ou Scopus.
– 10 articles dans des conférences internationales avec comité de lecture dont 9 sont référencés sur DBLP et/ou Scopus.
– 10 articles dans des conférences francophones avec comité de lecture dont 5 sont référencés sur DBLP et/ou Scopus.
– Une présentation en tant que conférencier invitée (CORIA 2011).

## 4.1 Revues avec comité de lecture

1. B. Douar, M. Liquière, **C. Latiri** and Y. Slimani. LC-MINE : a framework for frequent subgraph mining with local consistency techniques. *Knowledge and Information Systems Journal*, Springer-Verlag (To appear in 2013) [Référencé sur DBLP].
2. **C. Latiri**, H. Haddad and T. Hamrouni. Towards An Effective Automatic Query Expansion Process Using An Association Rule Mining Approach. *Journal of Intelligent Information Systems*, Volume 39, Issue 1, pages 209-247, August 2012, Springer-Verlag [Référencé sur DBLP et Scopus].
3. **C. Latiri**, K. Smaïli, C. Lavecchia, D. Langlois and C. Nasri.Phrase-based machine translation based on text mining and statistical language modeling techniques. *International Journal of Computational Linguistics and Applications, IJCLA journal*, Vol. 2, $N°$. 1-2, JAN-DEC 2011, pages 193-208, BAHRI Publications.
4. B. Douar, **C. Latiri**, M. Liquière, and Y. Slimani. B-FGMAC : Breadth-first Frequent subGraph Mining with Arc Consistency. *Journal of Artificial Intelligence and Soft Computing Research*, Volume 1, Number 4, pages 269-281, Polish Neural Network Society, 2011.
5. **C. Latiri**, K. Smaïli, C. Lavecchia and D. Langlois. Mining Monolingual and Bilingual Corpora. *Intelligent Data Analysis Journal*, Volume 14, Issue 6, pages 663-682, November 2010, IOS Press [Référencé sur DBLP et Scopus].
6. **C. Latiri**, J. P. Chevallet, S. Elloumi, and A. Jaoua. Une Extension de la Connexion de Galois Floue pour la Recherche d'Information. *Revue I3 : Information - Intéraction - Intelligence*, Volume 3, $N°2$, pages 73-116, Janvier 2004, Cépaduès.

## 4.2 Conférences internationales avec comité de lecture et actes

1. Eya Znaidi, Lynda Tamine, Cecile Chouquet and **C. Latiri**. Characterizing Health-related Information Needs of Domain Experts. *In proceedings of the* $14^{th}$ *conference on Artificial*

*Intelligence in MEdicine (AIME 2013), pages 48-57, May 29 - June 1, 2013, Murcia, Spain.*

2. **C. Latiri**, L. Ben Ghezaiel and M. Ben Ahmed. Proxemic Conceptual Network based on Ontology Enrichment for Representing Documents in IR. *Proceedings of the* $18^{th}$ *International Conference on Knowledge Engineering and Knowledge Management (EKAW 2012), Galway City, Ireland, 8-12 October 2012*, Volume 7603 of LNAI, pages 72-86, Springer-Verlag [Référencé sur DBLP].
3. L. Ben Ghezaiel, **C. Latiri** and M. Ben Ahmed. Ontology Enrichment based on Generic Basis of Association Rules for Conceptual Document Indexing. *In Proceedings of the* $4^{th}$ *the International Conference on Knowledge Engineering and Ontology Developpement (KEOD 2012), Barcelona, Spain, October, 4-7, 2012*, SciTePress [Référencé sur DBLP].
4. C. Nasri, K. Smaïli and **C. Latiri**. A new method for learning Phrase Based Machine Translation with Multivariate Mutual Information. *In Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE'12) , Hefei, China, September, 20-24, 2012* [Référencé sur DBLP].
5. L. Ben Ghezaiel, **C. Latiri** and M. Ben Ahmed. Conceptual Indexing Documents in IR based on Ontology Enrichment. *Proceedings of the* $16^{th}$ *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, special session on Ontology-Based Information Retrieval (KES 2012), In Advances in Knowledge-Based and Intelligent Information and Engineering Systems M. Graña et al. (Eds.), pages 1920-1931, San Sebastian, Spain, September, 10-12, 2012*, IOS Press [Référencé sur DBLP].
6. C. Nasri, K. Smaïli and **C. Latiri**. Training Statistical Machine Translation with Multivariate Mutual Information. *In Proceedings of the* $5^{th}$ *Language and Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics, LTC 2011, Poznan (Poland), November, 25-27, 2011.*
7. B. Douar, M. Liquière, **C. Latiri** and Y. Slimani. Graph-based relational learning with a polynomial time projection algorithm. *In Proceedings of the* $21^{st}$ *International Conference on Inductive Logic Programming, ILP 2011,* volume 7207 of LNAI, pages 96-112, Springer-Verlag [Référencé sur DBLP et Scopus].
8. B. Douar, **C. Latiri**, M. Liquière and Y. Slimani. FGMAC : Frequent SubGraph Mining with Arc Consistency. *In Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2011, part of the IEEE Symposium Series on Computational Intelligence 2011, Paris (France), pages 112-119, April, 11-15, 2011* [Référencé sur DBLP et Scopus].
9. C. Trabelsi, **C. Latiri** and K. Ghedira. Generating closed frequent gensets under constraints based on FP-Trees. *In Proceedings of the IEEE Conference on Computational Engineering in Systems Applications, CESA'06, pages 1526-1531, Beijing (China), October, 4-6, 2006* [Référencé sur Scopus].
10. **C. Latiri**, W. Bellagha, and S. Ben Yahia. VIE-MGB : A Visual Interactive Exploration of Minimal Generic Basis of Association Rules. *In Proceedings of the* $3^{rd}$ *International Conference on Concept Lattices and their Applications, CLA'05, Olomouc (Czech Republic), pages 179-196, September, 7-9, 2005.*
11. **C. Latiri** and S. Ben Yahia. How can fuzzy association rules improve query expansion in IR ? In *Proceedings of the Second International Workshop on Advanced Computation for Engineering Applications, ACEA'03, Cairo (Egypt), December, 21-22, 2003.*

12. H. Haddad and **C. Latiri**. Query expansion using crisp association rules between terms. In *Proceedings of the Second International Workshop on Advanced Computation for Engineering Applications, ACEA'03, Cairo (Egypt), December, 21-22, 2003.*
13. **C. Latiri**, S. Ben Yahia, and A. Jaoua. Query expansion using fuzzy association rules. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Discrete Mathematics, Metz (France), pages 231-242, September, 3-9, 2003.*
14. **C. Latiri**, J. P. Chevallet, S. Elloumi and A. Jaoua. Extension of fuzzy Galois connection for information retrieval using a fuzzy quantifier. In *Proceedings of the ACS/IEEE International Conference on Computer Systems and Applications, AICCSA'03, pages 84-90, Tunis (Tunisia), July, 14-18, 2003* [Référencé sur DBLP].
15. **C. Latiri**, S. Ben Yahia and G. Mineau. Conceptual non-redundant association rules discovery : Application to query expansion. In *Proceedings of the First International Conference on Formal Concept Analysis : The State of the Art, ICFCA03, Damstadt (Germany), 28 February-1 March, 2003.*
16. **C. Latiri**, J. P. Chevallet and S. Elloumi. Fuzzy Galois Connection based Information Retrieval Model. In *Proceedings of the EUROFUSE Workshop on Information Systems, Villa Monastery, Varenna (Italy), September, 23-25, 2002.*
17. **C. Latiri**, S. Ben Yahia and S. Elloumi. Connexion de Galois Floue : Extension et application au textmining. In *Proceedings of the* $9^{th}$ *International Conference IPMU'2002, Annecy (France), volume 3, pages 1407-1413, July, 1-5, 2002.*
18. **C. Latiri** and S. Ben Yahia. Generating implicit association rules from textual data. In *Proceedings of the ACS/IEEE International Conference on Computer Systems and applications, AICCSA'01, Beirut (Lebanon), pages 137-143, June, 25-29, 2001* [Référencé sur DBLP].

## 4.3 Conférences francophones avec comité de lecture et actes

1. Eya Znaidi, Lynda Tamine, Cecile Chouquet and **C. Latiri**. Analyses exploratoires des requêtes d'experts médicaux : cas des campagnes TREC et CLEF. *Dans les actes de la* $2^{\text{ème}}$ *édition du Symposium sur l'Ingénierie de l'Information Médicale SIIM 2013, Lille, 1 Juillet 2013.*
2. A. Amri and M. Mbarek and C. Bechikh and **C. Latiri** and H. Haddad. Indexation à base des syntagmes nominaux. *Actes de JEP-TALN-RECITAL 2012, Atelier DEFT 2012 : DÉfi Fouille de Textes, pages 33-39, Grenoble, France, 4-8 Juin.*
3. L. Ben Ghezaiel, **C. Latiri**, M. Ben Ahmed and N. Khouja. Un réseau proxémique pour la recherche d'information : Application à la maladie des dystonies. *Actes de la première édition du Symposium sur l'Ingénierie de l'Information Médicale SIIM 2011, pages 25-40, Toulouse (France), 9-10 Juin 2011.*
4. B. Douar, M. Liquiere, **C. Latiri** and Y. Slimani. Nouvelle approche de fouille de graphes AC-réduits fréquents. *Actes des Journées francophones EGC'2011, Revue des Nouvelles Technologies de l'Information, RNTI-E-20, pages 473-478, Brest (France), 25-28 Janvier 2011* [Référencé sur DBLP].
5. B. Douar, M. Liquière, **C. Latiri** and Y. Slimani. Un biais de projection peut faire la différence dans la fouille de graphes. *Première Journée de fouille de grands graphes, IRIT, Toulouse (France), 13 Octobre 2010.*

6. L. Ben Ghezaiel, **C. Latiri**, M. Ben Ahmed and N. Khouja. Enrichissement d'ontologie par une base générique minimale de règles d'association. *Actes de la Conférence en Recherche d'Information et Applications CORIA'2010, Sousse (Tunisie), pages 289-300, 18-20 Mars 2010* [Référencé sur DBLP].
7. **C. Latiri**, C. Nasri, K. Smaïli et Y. Slimani. Extraction des séquences fermées fréquentes à partir de corpus parallèles : Application à la traduction automatique. *Actes des Journées francophones EGC'2010, Revue des Nouvelles Technologies de l'Information, RNTI-E-19, pages 55-60, Hammamet (Tunisie), 26-29 Janvier 2010* [Référencé sur DBLP].
8. B. Douar, **C. Latiri** and Y. Slimani. Approche hybride de classification supervisée à base de treillis de Galois : application à la reconnaissance de visages. *Actes des Journées francophones EGC'2008, Revue des Nouvelles Technologies de l'Information, RNTI-E-11, pages 309-320, INRIA, Sophia-Antipolis, Nice (France), 29 Janvier-* $1^{er}$ *Février 2008* [Référencé sur DBLP].
9. **C. Latiri**, L. Ben Ghezail et M. Ben Ahmed Fast-MGB : Nouvelle base générique minimale pour l'extraction de règles d'association *Actes des Journées francophones EGC'2006, Revue des Nouvelles Technologies de l'Information, RNTI-E-6, pages 217-222, Lille (France), 17-20 Janvier 2006* [Référencé sur DBLP].
10. **C. Latiri**, M. Mtir, and S. Ben Yahia. Méthode de construction d'ontologie de termes à partir du treillis de l'icerberg de Galois. *Actes des Journées francophones EGC'2005, Revue des Nouvelles Technologies de l'Information, RNTI-E-3, pages 365-376, Paris (France), 18-21 Janvier 2005* [Référencé sur DBLP].
11. S. Bsiri, M. H. Zargayouna, **C. Latiri**, and S. Ben Yahia. Découverte de règles d'association hiérarchiques entre termes. *Actes du* $21^{ème}$ *Congrés Informatique des Organisations et Systèmes d'Information et de Décision, INFORSID'03, Nancy (France), pages 333-347, 24-27 Mai 2003* [Référencé sur DBLP].
12. **C. Latiri**, S. Ben Yahia, G. Mineau, and A. Jaoua. Découverte des règles d'association non redondantes : Application aux corpus textuels. *Actes des Journées francophones EGC'2003, Lyon (France), Revue des Sciences et Technologies de l'Information - série RIA ECA 17(1), pages 131-144, 22-24 Janvier 2003* [Référencé sur DBLP].
13. **C. Latiri**, S. Ben Yahia, S. Elloumi, and A. Jaoua. Textmining : Extension de la connexion de Galois floue. *Actes des Journées francophones EGC'2002, Montpelier (France), Revue Extraction des Connaissances et Apprentissage 1(4), pages 375-386, 21-23 Janvier 2002* [Référencé sur DBLP].
14. **C. Latiri** and S. Ben Yahia. Textmining : Discovering explicit formal concepts from unstructured data. *Actes du* $XIX^{ème}$ *Congrès Informatique des Organisations et Systèmes d'Information et de Décision, INFORSID'01, Martigny (Suisse), pages 27-39, 29 Mai-* $1^{er}$ *Juin 2001* [Référencé sur DBLP].

### Distinction

*Prix AFIA (Association Française d'Intelligence Artificielle) pour la meilleure contribution scientifique de la conférence EGC 2003, Lyon (France), 22-24 Janvier 2003.*

## 4.4 Articles soumis et en révision

1. B. Douar, **C. Latiri**, M. Liquière and Y. Slimani. A projection bias in frequent subgraph mining can make a difference. ***Under revision*** *in International Journal on Artificial*

*Intelligence Tools, 20 May 2012, World Scientific Publishing Company.*

2. L. Ben Ghezaiel, **C. Latiri** and M. Ben Ahmed. Conceptual Indexing in IR based on Ontology Enrichment and Association Rules. *Submitted in Journal of Applied Ontology, April 15, 2013, IOS Press.*

### 4.5 Tableau récapitulatif du nombre de publications scientifiques durant la période 2001-2013

| Année | Revues | Conférences Internationales | Conférences Francophones | Total |
|---|---|---|---|---|
| *Publications post-thèse de doctorat* | | | | |
| 2013 | 1 | 1 | 1 | **3** |
| 2012 | 1 | 4 | 1 | **6** |
| 2011 | 2 | 3 | 2 | **7** |
| 2010 | 1 | - | 3 | **4** |
| 2009 | - | - | - | **0** |
| 2008 | - | - | 1 | **1** |
| 2007 | - | - | - | **0** |
| 2006 | - | 1 | 1 | **2** |
| 2005 | - | 1 | 1 | **2** |
| | **5** | **10** | **10** | **25** |
| *Publications liées au travaux de la thèse de doctorat* | | | | |
| 2004 | 1 | - | - | **1** |
| 2003 | - | 5 | 2 | **7** |
| 2002 | - | 2 | 1 | **3** |
| 2001 | - | 1 | 1 | **2** |
| | **1** | **8** | **4** | **13** |
| Total | **6** | **18** | **13** | **38** |

## 5 Activités d'encadrement et de co-encadrement

Je suis membre permanent du Laboratoire d'Informatique en Programmation Algorithmique et Heuristique (LIPAH) de la Faculté des Sciences de Tunis (FST) depuis 1999. Après l'obtention du doctorat en 2004, je me suis impliquée dans la co-direction de sujets de mastère de recherche et de thèses de doctorat à l'école doctorale de la FST et également à l'école doctorale de l'École Nationale des Sciences de l'Informatique (ENSI). Depuis 2007, j'anime un groupe de recherche au sein du laboratoire LIPAH dans le domaine de l'Extraction des Connaissances à partir de Textes. Les tableaux ci-dessous synthétisent mes activités d'encadrement et de co-encadrement.

Actuellement, je fais partie d'un groupe d'enseignants chercheurs, répartis sur plusieurs universités tunisiennes, qui travaillent sur la fouille de données et ses applications à différents domaines. Étant affiliée à l'ISAMM de l'université de la Manouba, mon objectif, à court terme, est de mettre en place une unité de recherche sur la RI multi-sources au sein de l'ISAMM. A travers cette unité de recherche, je me propose de réunir les différentes compétences qui activent dans les domaines de la fouille de données et la RI.

## 5.1 Encadrement de mastères de recherche

| **Année** | **Étudiant(e)** | **Sujet** | **Directeur** | **Résultat** |
|---|---|---|---|---|
| 2004-2005 | Lamia Ben Ghezail | Approche de génération d'une base générique minimale de règles d'association entre termes. | Pr. Mohammed Ben Ahmed (RIADI-GDL, ENSI) | Soutenu en 2006 |
| 2005-2006 | Chiraz Trabelsi | Approche binaire pour l'extraction sous contraintes de règles d'association entre gènes et facteurs de transcription en bioinformatique. | Pr. Khaled Ghedira (SOIE, ISG) | Soutenu en 2006 |
| 2006-2007 | Mehdi Mtir | Construction d'ontologies à partir de textes basée sur l'AFC. | Pr. Yahya Slimani (Encadreur, ISAMM, Université de la Manouba) | Soutenu en 2007 |
| 2006-2007 | Brahim Douar | Classification par les treillis de Galois : application à la biométrie. | Pr. Yahya Slimani (Encadreur, ISAMM, Université de la Manouba) | Soutenu en 2007 |
| 2007-2008 | Wissem Bellagha | Approche d'extraction de motifs séquentiels fréquents flous. | Pr. Yahya Slimani (Encadreur, ISAMM, Université de la Manouba) | Soutenu en 2008 |
| 2008-2009 | Cyrine Nasri | Approche d'extraction de séquences inter-langues : Application à la Traduction Automatique Statistique. | Pr. Yahya Slimani (Encadreur, ISAMM, Université de la Manouba) | Soutenu en 2009 |
| 2010-2011 | Mohamed Chebel | Extraction et déploiement de règles d'association inter-langues dans la Traduction Automatique Statistique. | M. Sadok Ben Yahia (LIPAH, FST) | Soutenu en 2011 |
| 2011-2012 | Soumeya Guesmi | Expansion de requêtes en RI par un réseau proxémique conceptuel. | Chiraz Latiri (LIPAH, FST) | Soutenu en 2012 |
| 2011-2012 | Malek Hajjem | Extraction de règles d'association inter-syntagmes nominaux pour l'enrichissement d'une ontologie multilingue. | Chiraz Latiri (LIPAH, FST) | Soutenu en 2012 |
| 2011-2012 | Sourour Bel Hadj Rhouma | Extraction de règles d'association inter-langues pour la traduction de requêtes en Recherche d'Information Multilingue. | Chiraz Latiri (LIPAH, FST) | Soutenu en 2012 |
| 2011-2012 | Asma Ben Achour | Extraction de règles d'association inter-langues pour l'expansion d'index en Recherche d'Information Multilingue. | Chiraz Latiri (LIPAH, FST) | Soutenu en 2012 |
| 2011-2012 | Nidhal Boubtane | Extraction des syntagmes inter-langues pour la traduction de requêtes en RIM. | Chiraz Latiri (LIPAH, FST) | Soutenu en 2013 |
| 2011-2012 | Marwa Mbarki et Amine Amri | Participation à la campagne d'évaluation DEFT 2012. | Chiraz Latiri (LIPAH, FST) | Soutenu en 2013 |

## 5.2 Co-encadrement de thèses de doctorat

| **Première inscription** | **Étudiant(e)** | **Sujet** | **Directeur** | **Résultat** |
|---|---|---|---|---|
| 2006-2007 | Lamia Ben Ghezail | Indexation conceptuelle en RI fondée sur l'enrichissement d'ontologie par une base générique de règles d'association | Pr. Mohammed Ben Ahmed (RIADI-GDL, ENSI) | Première inscription en Décembre 2007 (**Thèse déposée le 15 Mars 2013**) |
| 2007-2008 | Brahim Douar | Nouvelles Approches de Fouille de Graphes AC-réduits Fréquents | Pr. Yahya Slimani (Encadreur, ISAMM, Université de la Manouba) et M. Michel Liquière (LIRMM, Université de Montpellier 2) | **Thèse en cotutelle soutenue le 27 Novembre 2012, Composition du jury :** Pr. Yahya Slimani (Encadreur, ISAMM, Université de la Manouba), Pr. Jean Sallantin (Encdareur, LIRMM), Pr. Faiez Gargouri (Rapporteur, ISIMS, Université de Sfax), Pr. M. Mohsen Gammoudi (Rapporteur, ISAMM, Université de la Manouba) et Pascal Poncelet (Président, LIRMM) |
| 2009-2010 | Cyrine Nasri | Fouille de triggers interlangues pour la Traduction Automatique Statistique par l'IM conditionnelle multivaluée | Pr. Yahya Slimani (ISAMM, Université de la Manouba) et Pr. Kamel Smaïli (LORIA, Université de Lorraine) | Thèse en cotutelle, première inscription en Septembre 2009, date de soutenance prévue en 2014 |
| 2011-2012 | Mohamed Abdenadher | Formalisation et extraction des fermés de cliques maximales à partir de grands graphes : Application à l'analyse des réseaux sociaux | Pr. Yahya Slimani (ISAMM, Université de la Manouba) et M. Michel Liquière (LIRMM, Université de Montpellier 2) | Thèse en cotutelle, première inscription en Septembre 2011 |
| 2011-2012 | Eya Znaidi | Contribution à la définition de modèles de recherche d'information spécifiques au domaine médical | Pr. Yahya Slimani (ISAMM, Université de la Manouba) et Mme. Lynda Tamine (IRIT, Université Toulouse III - Paul Sabatier) | Thèse en cotutelle, première inscription en Septembre 2011 |
| 2012-2013 | Mohamed Chebel | Classification de documents multilingues à base de Treillis de Galois | M. Sadok Ben Yahia (FST, Université Tunis El Manar) et Pr. Engelbert Mephu (LIMOS, Université Blaise Pascal, Clairmont Ferrand) | Thèse en cotutelle, première inscription en Septembre 2012 |

# 6 Rayonnement et collaborations scientifiques

## 6.1 Participation à des campagnes d'évaluation

Nous avons participé au défi fouille de texte DEFT2012 [8]. Il s'agit d'un atelier d'évaluation francophone en fouille de textes qui existe depuis 2005 et qui propose, chaque année, des thématiques de recherche régulièrement renouvelées. Dans le cadre des l'édition 2011 du DEFT, les participants ont travaillé sur des articles scientifiques dans le domaine des Sciences Humaines et Sociales en effectuant des appariements entre résumés et articles scientifiques (à quel article scientifique correspond un résumé ?). Pour l'édition 2012, nous avons travaillé sur le même corpus d'articles scientifiques en focalisant la recherche autour de la problématique de l'indexation des articles scientifiques. Il s'agit d'identifier les mots clés choisis par les auteurs pour indexer leur article, à partir du résumé et de l'article scientifique complet. Dans ce contexte, nous avons proposé d'utiliser conjointement les syntagmes nominaux et les règles d'association entre termes. L'atelier de clôture s'est tenu lors de la conférence JEP/TALN2012 à Grenoble du 4 au 8 juin 2012.

## 6.2 Membre de comité de lecture de conférences scientifiques

J'ai assuré le rôle d'évaluateur d'articles scientifiques dans les conférences suivantes :

- The $7^{th}$ African Conference on Research in Computer Science (CARI'04), 22-25 Novembre 2004, Hammamet, Tunisie.
- Fourth International Conference On Concept Lattices and Their Applications (CLA'06), 30 Octobre au $1^{er}$ Novembre 2006, Hammamet, Tunisie.
- $1^{st}$ International Co-DEXA'07 co-located workshop on Advances in Conceptual knowledge Engineering (ACKE'2007), September 7, 2007, Regensburg, Germany.
- The Fourth International Workshop on Advanced Computation for Engineering Applications (ACEA08), 23-24 Juillet 2008, Salt, Jordanie.
- $2^{ème}$ édition de la Conférence Internationale sur les Systèmes d'Information et Intelligence Économique (SIIE 2009), 12-14 Février 2009, Hammamet, Tunisie.
- $3^{ème}$ édition de la Conférence Internationale sur les Systèmes d'Information et Intelligence Économique (SIIE 2010), 18-20 Février 2010, Sousse, Tunisie.
- Première édition de la Conférence Maghrébine sur l'Extraction et la Gestion des Connaissances (EGC Maghreb 2010), 13-14 Décembre 2010, Alger, Algérie.
- $2^{ème}$ édition de la Conférence Maghrébine sur l'Extraction et la Gestion des Connaissances (EGC Maghreb 2011), 23-25 Novembre 2011, Tanger, Maroc .
- COnférence en Recherche d'Information et Applications (CORIA 2012), 21-23 Mars 2012, Bordeaux, France.
- $16^{th}$ International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, special session on Ontology-Based Information Retrieval (KES 2012), du 10-12 Septembre 2012, San Sebastian, Spain.
- The $21^{st}$ ACM International Conference on Information and Knowledge Management(CIKM 2012), 29 Octobre au 2 Novembre 2012, Maui, USA.
- $3^{ème}$ édition de la Conférence Maghrébine sur l'Extraction et la Gestion des Connaissances (EGC Maghreb 2012) -Chair PhD Session-, 13-15 Novembre 2012, Hammamet, Tunsisie.
- COnférence en Recherche d'Information et Applications (CORIA 2013), 3-5 Avril 2013, Neuchâtel, Suisse.

8. http ://deft.limsi.fr/2012/

## 6.3 Membre de comité de lecture de journaux scientifiques

Editorial member of "*The International Journal of Computational Science and Engineering* (IJCSE)", Inderscience Publishers.

## 6.4 Organisation de conférences francophones et internationales

J'ai participé activement dans l'organisation des conférences scientifiques suivantes :

| **Date** | **Lieu** | **Titre de la conférence** |
|---|---|---|
| Du 18 au 25 Novembre 2004 | Hammamet, Tunisie | $7^{ème}$ Colloque Africain sur la Recherche en Informatique (CARI'04) |
| Du 30 Octobre au 1 Novembre 2006 | Hammamet, Tunisie | The $4^{th}$ International Conference on Concepts Lattices and their Applications (CLA 2006) |
| Du 26 au 29 Janvier 2010 | Hammamet, Tunisie | $10^{èmes}$ Journées francophones EGC'2010 (Co-Présidente du comité d'organisation) |
| Du 01 au 05 Février 2010 | Hammamet, Tunisie | é-EGC Apprentissage statistique et data mining (Présidente du comité d'organisation) |
| Du 13 au 14 Décembre 2010 | Alger, Algérie | Première édition de la Conférence Maghrébine sur l'Extraction et la Gestion des Connaissances (EGC Maghreb 2010) |
| Du 12 au 15 Novembre 2012 | Tunis, Tunisie | Troisième édition de la Conférence Maghrébine sur l'Extraction et la Gestion des Connaissances (EGC Maghreb 2012) |

## 6.5 Activités de recherche menées au sein de l'équipe MRIM du Laboratoire d'Informatique de Grenoble (Ex CLIPS-IMAG)

En tant que membre invitée de l'équipe de recherche MRIM au sein du laboratoire de recherche d'informatique de Grenoble (Ex CLIPS-IMAG), de décembre 2001 à décembre 2003, j'ai effectué plusieurs séjours scientifiques dans le cadre de la préparation de ma thèse de doctorat (décembre 2001, janvier 2002, mars 2002, juillet 2002, décembre 2002, mars 2003, juillet 2003 et décembre 2003). Durant ces séjours, j'ai animé des séminaires de recherche relatifs à mes travaux de thèse. Ces derniers ont permis aussi de développer plusieurs collaborations scientifiques avec des chercheurs de l'équipe MRIM de Grenoble et ont donné lieu à des publications relatives à des axes de recherche communs (*cf.* Partie 4, page 27).

## 6.6 Participation dans le projet CMCU DMP N° 05G1412 (Data Mining Parallèle, 2005-2009)

Ce projet franco-tunisien s'est déroulé en coopération entre la Faculté des Sciences de Tunis (M. Yahya Slimani), le laboratoire CRIL de Lens (M. Engelbert Mephu) et le LGI2A (M. Gilles Goncalves) sur les méta-heuristiques parallèles pour la fouille de données.

Étant membre de ce projet, j'ai effectué un premier séjour scientifique d'une semaine au mois de décembre 2005 au sein du CRIL qui m'a permis de définir le cadre fédérateur de mon habilitation à diriger les recherches, à savoir l'ECT. Le deuxième séjour scientifique, que j'ai effectué durant le mois de juillet 2007 au CRIL, m'a permis d'approfondir mes travaux d'habilitation et la rédaction d'un article intitulé *"Approche hybride de classification supervisée à base de treillis de Galois : application à la reconnaissance de visages"* accepté aux Journées Francophones d'Extraction et Gestion des Connaissances à partir des Données (EGC 2008). Depuis, je continue à collaborer avec M. Engelbert Mephu, responsable du projet CMCU du côté français, sur des problématiques de fouille de données par des méthodes basées sur l'Analyse Formelle de Concepts.

Ceci m'a permis d'assurer, dans le cadre de ce projet, depuis octobre 2006, le co-encadrement de 3 mastères de recherche.

Le projet CMCU a permis de lancer une thèse en co-tutelle que j'ai co-encadrée avec M. Michel Liquière (Laboratoire LIRMM, Montpellier) et M. Yaya Slimani (Laboratoire LISI, INSAT-Tunis) et dont la problématique porte sur la fouille des graphes. La thèse a été soutenue le 27 Novembre 2012. J'ai également participé à l'organisation de la conférence CLA 2006 "Fourth Conference on Concept Lattices and their Applications" qui s'est tenue à Hammamet en Tunisie du 30 octobre au $1^{er}$ novembre 2006, dans le cadre du même projet.

## 6.7 Activités de recherche menées au sein de l'équipe PAROLE du Laboratoire LORIA, Nancy

La collaboration entre les deux équipes PAROLE (Laboratoire LORIA) et LIPAH (FST) a initialement commencé suite à ma rencontre avec le Pr. Kamel Smaïli (LORIA, Université Nancy 2) lors de la conférence internationale SIIE 2008 (Système d'Information et Intelligence Économique) qui a eu lieu en Tunisie en Février 2008. La communication présentée par le Pr. Kamel Smaïli, sur le thème de la traduction automatique à base de méthodes statistiques, a suscité mon intérêt pour le domaine de la Traduction Automatique Statistique (TAS) et nous avons par la suite amorcé un nouvel axe de recherche qui consiste à utiliser des résultats de l'ECT à base de l'AFC dans la TAS, avec pour objectif de se comparer avec l'approche statistique de l'équipe PAROLE.

Depuis Juillet 2008, en tant que chercheur visiteur, j'ai été invitée par le Pr. Kamel Smaïli à plusieurs reprises pour des courts séjours scientifiques, dans le but d'effectuer les évaluations expérimentales de nos contributions en TAS et comparer les résultats avec ceux de l'équipe PAROLE. Ce travail a donné lieu à la publication d'un article dans une revue indexée et trois autres dans des conférences internationales avec comité de lecture (*cf.* Partie 4, page 27).

Cette collaboration a permis également de :

- Lancer un mastère de recherche sur la problématique d'utilisation des règles d'association inter-langues dans la TAS (Soutenu en juillet 2009).
- Démarrer, en Septembre 2010, une thèse en co-tutelle que je co-encadre avec le Pr. Yahya Slimani (Encadreur, ISAMM, Université de la Manouba) et le Pr. Kamel Smaïli et dont la problématique s'intéresse à l'extraction des séquences inter-langues et des triggers inter-langues à base de l'information mutuelle condionnelle, pour la TAS.
- Monter le projet CMCU EXQUI en collaboration avec d'autres structures de recherche françaises.

## 6.8 Participation dans le projet CMCU N° 11G1417 EXQUI : EXtraction, QUalité et Ingénierie des connaissances dans les environnements hétérogènes (2011-2013)

Ce projet CMCU dont je suis membre a été initié entre le laboratoire LIPAH de la Faculté des Sciences de Tunis (M. Sadok Ben Yahia) et le laboratoire LIMOS à Clermont-Ferrand (M. Mephu Engelbert).

Ma participation dans ce projet a permis d'associer l'équipe PAROLE du LORIA (Nancy) et l'équipe SIG-FRI de l'IRIT (Toulouse). Les recherches entamées ainsi que les perspectives de recherche envisagées s'articulent autour de la recherche d'information multilingue, couplant la RI classique et la TAS. Notre objectif final est de pouvoir profiter des résultats de nos contributions précédentes en RI et en TAS pour proposer de nouvelles approches en RI multilingue. Nous

nous intéressons également à une nouvelle orientation de recherche liée à l'Analyse des Réseaux Sociaux (ARS) en collaboration avec M. Michel Liquière (Laboratoire LIRMM, Montpellier), membre du projet.

Ainsi, dans le cadre de ce projet CMCU, ma participation s'adresse principalement aux problématiques suivantes, qui font partie de mon projet de recherche à court et à moyen terme :

- Étudier l'apport des règles associatives inter-langues dans le domaine de la RI multilingue, en terme de gain de performance de la tâche de RI.
- Étudier de nouvelles approches pour la multilinguisation d'ontologies en se basant sur l'extraction de règles d'association entre les syntagmes nominaux inter-langues à partir d'un corpus parallèle.
- Explorer les corpus comparables en TAS par nos techniques d'ECT à base d'AFC.
- Étudier la problématique d'extraction des fermés de cliques maximales pour la complétion de liens et la détection de communautés dans les réseaux sociaux.
- Utiliser nos approches de fouille de sous-graphes fréquents à base d'arc consistence pour la prédiction de liens dans les réseaux sociaux.

## 6.9 Collège EGC Maghreb

Suite à l'organisation de la $10^{ème}$ édition de la conférence francophone EGC 2010 en Tunisie, j'ai entrepris avec un groupe de chercheurs maghrébins et français et en collaboration avec l'association EGC France de créer le collège Maghrébin EGC-M. Le principal objectif du collège est de trouver des opportunités de développement de la recherche entre les pays du Maghreb, et ce par la fédération et la mise en réseau des compétences maghrébines dans le domaine de l'extraction et la gestion des connaissances.

Nous avons crée un comité de pilotage EGC-M dont je suis membre. Ce comité a pour rôle de veiller à l'organisation, à la pérennité et au rayonnement des toutes les activités scientifiques qui seront menées par le collège Maghrébin.

La première édition de la conférence internationale EGC-M a été organisée par l'École Nationale Supérieure d'Informatique d'Alger (Algérie), alors que la deuxième édition s'est tenue à Tanger (Maroc) et a été organisée par l'ENSA de Tanger. La troisième édition s'est tenue à Tunis du 12 au 15 Novembre 2012, organisée par l'Institut Supérieur d'Informatique de Tunis. Il importe de signaler que les rencontres de EGC-M offrent de réelles opportunités de discussions et de collaborations scientifiques.

À travers ce collège Maghrébin, nous œuvrons à mettre en place des projets de coopération bilatéraux qui vont permettre d'instaurer un cadre institutionnel de collaboration (co-encadrements, publications conjointes, échanges de chercheurs, etc.).

## 6.10 Synthèse de la collaboration avec les structures de recherche françaises

| **Période** | **Structure de recherche** | **Contact** | **Type de collaboration** |
|---|---|---|---|
| 2000-2004 | Laboratoire LIG (Équipe MRIM), Grenoble | Catherine Berrut et Jean-pierre Chevallet | Co-encadrement de ma thèse de doctorat et publications communes (*cf.* sous-section 6.5, page 35). |
| 2003-2008 | Centre de Recherche en Informatique de Lens (CRIL) | Engelbert Mephu Nguifo | Montage du projet CMCU "Data Mining Parallèle N° 05G1412", organisation de la conférence internationale CLA 2006 et co-encadrement de mastères de recherche et d'une thèse de doctorat (*cf.* sous-section 6.6, page 35). |
| 2008-2012 | Laboratoire LORIA (Projet PAROLE), Nancy | Kamel Smaïli | Parrainage scientifique de mon projet d'HDR, montage du projet CMCU " N° 11G1417 EXQUI : EXtraction, QUalité et Ingénierie des connaissances dans les environnements hétérogènes", co-encadrement de mastères de recherche et d'une thèse de doctorat et publications scientifiques communes (*cf.* sous-section 6.7, page 36). |
| 2008-2012 | Laboratoire LIRMM (Équipe COCONUT), Montpellier | Michel Liquière | Montage des deux projets CMCU DMP et EXQUI et co-encadrement de deux thèses de doctorat (*cf.* sous-section 6.6, page 35). |
| 2008-2012 | Laboratoire LIMOS, Université Blaise Pascal, Clermont-Ferrand | Engelbert Mephu Nguifo | Montage du projet CMCU " N° 11G1417 EXQUI : EXtraction, QUalité et Ingénierie des connaissances dans les environnements hétérogènes" et animation d'un groupe de recherche en Fouille de données (*cf.* sous-section 6.8, page 36). |
| 2010-2013 | IRIT (Équipe SIG), Toulouse | Lynda Tamine | Montage du projet CMCU " N° 11G1417 EXQUI : EXtraction, QUalité et Ingénierie des connaissances dans les environnements hétérogènes" et co-encadrement d'une thèse de doctorat (*cf.* sous-section 6.8, page 36). |

## Deuxième partie

# Mémoire de Recherche

# Positionnement scientifique

## 1 Cadre fédérateur : Extraction de Connaissances à partir de Textes (ECT)

Pour faire face à l'augmentation vertigineuse, d'année en année, du volume de données disponibles sous forme de corpus de textes ou de collections documentaires, l'un des principaux défis de la communauté de recherche en ECT est de proposer des méthodes et des techniques capables de traiter une telle masse de données textuelles, mais aussi de prendre en compte leur complexité croissante [Berry, 2008]. Très vite, l'ECT s'est trouvée au cœur de plusieurs domaines de recherche.

La revue de la littérature relative au domaine de l'ECT montre que les contributions proposées traitent de nouvelles approches théoriques quant à l'extraction et à la représentation de connaissances à partir de textes et aussi de l'implémentation de solutions algorithmiques performantes, applicables à de grandes masses de données textuelles [Zhong *et al.*, 2012]. Plusieurs communautés scientifiques se sont intéressées aux problématiques posées par l'ECT. Un survol rapide des domaines qui se croisent avec l'ECT, montre que le Traitement du Langage Naturel (TAL) est le domaine le plus privilégié. Aujourd'hui, au moment où le TAL se transforme de plus en plus en ingénierie de langues, l'ECT vient cerner au mieux les limites théoriques et algorithmiques des problèmes que soulève le TAL et proposer des solutions aux problèmes qui pèsent sur l'efficacité des systèmes de TAL proposés. Cette complexité vient du fait que le TAL œuvre à extraire des éléments sémantiques à partir de textes et se trouve par conséquent à la frontière entre la linguistique et l'informatique [Jurafsky and Martin, 2008]. Les applications liées au TAL sont assez nombreuses et diverses, comme par exemple la génération automatique de textes, l'extraction de résumés, la classification et la catégorisation de documents, l'extraction de descripteurs morpho-syntaxiques pour la fouille de textes [Béchet *et al.*, 2009], etc. En plus du traitement linguistique, les méthodes statistiques sont fortement présentes dans des applications du traitement de la langue naturelle, comme le cas du traitement de la parole ou la traduction automatique [Haton *et al.*, 2006].

Dès son apparition, l'ECT s'est trouvée aussi au centre du domaine de la Recherche d'Information (RI). Ce lien entre ECT et RI s'intéresse aux modèles, aux techniques et aux algorithmes permettant de sélectionner l'information pertinente en réponse à un besoin d'information exprimé par un utilisateur à l'aide d'une requête. De nombreux modèles de RI (comprenant des modèles d'indexation et des modèles d'appariement) ont été proposés, tels que le modèle vectoriel [Salton and Buckely, 1988], le modèle probabiliste [Jones *et al.*, 2000] ou encore le modèle de langue [Song and Croft, 1999]. La revue de la littérature en RI montre que ces modèles font souvent appel à la technique d'expansion de requêtes pour pallier au problème d'inadéquation au niveau de la correspondance entre la requête originelle et les document restitués. Généralement, les corrélations entre termes sont évaluées par le biais d'une large panoplie de mesures statistiques

de co-occurrences de termes dans les documents d'une collection [Qui and Frei, 1993, Sun *et al.*, 2006]. La détection de telles corrélations nécessite une analyse de l'ensemble des documents de la collection, détection qui est coûteuse en terme de temps de calcul.

L'axe de recherche qui nous intéresse dans le contexte de l'expansion de requêtes en RI, est la mise en œuvre d'une synergie entre les techniques classiques de RI et une technique d'ECT, à savoir l'extraction de règles d'association [Agrawal and Skirant, 1994]. Certains travaux de recherche dans la littérature ont déjà abordé cette problématique [Tangpong and Rungsawang, 2000, Latiri *et al.*, 2003c, Lin *et al.*, 2008]. L'idée clé est d'utiliser les connaissances additionnelles apportées par les règles d'association entre termes pour étendre les requêtes originelles [Rungsawang *et al.*, 1999, Haddad *et al.*, 2000, Latiri *et al.*, 2003b], dans le but d'améliorer la pertinence système d'un SRI.

D'un autre côté, la RI ne cesse d'évoluer en tenant compte de nouvelles représentations et interprétations de connaissances offertes par l'ECT et par l'Ingénierie de Connaissances (IC). En effet, la majorité des systèmes de recherche d'information (SRI) représentent les documents et les requêtes par des index souvent désignés par "sac de mots" [Baziz *et al.*, 2005]. Cette représentation stipule implicitement que les mots correspondent avec leurs sens. Plusieurs travaux de recherche ont mis en évidence les limites des modèles à base de "sac de mots" [Navigli, 2009], qui sont étroitement liées aux ambiguïtés que peuvent véhiculer le manque d'expressivité des mots singuliers de l'index ainsi que l'imprécision des requêtes utilisateur. Des problèmes, tels que la polysémie et la synonymie ne sont pas pris en compte [Baziz *et al.*, 2005]. Pour pallier à ces limites, des travaux de recherche proposent d'utiliser des structures conceptuelles lors de l'indexation [Popov *et al.*, 2004, Vallet *et al.*, 2005, Andreasen *et al.*, 2009]. La majorité des approches actuelles intègrent l'usage des ressources externes, telles que les ontologies et les hiérarchies des concepts dans le but d'assurer un gain de pertinence dans les SRIs, d'où l'apparition en RI de *l'indexation sémantique* et de *l'indexation conceptuelle* [Mihalcea and Moldovan, 2000, Khan and Luo, 2002, Castells *et al.*, 2007].

Dans le cadre de nos recherches, nous nous intéressons à coupler deux types de connaissances que nous pouvons atteindre par un processus d'ECT, à savoir les règles d'association entre termes qui représentent des connaissances implicites, et les ontologies qui traduisent plutôt des connaissances explicites relatives à un domaine. Le résultat de ce couplage est un réseau sémantique qui prend tout son intérêt dans une problématique d'indexation conceptuelle en RI.

Par ailleurs, l'ECT est fortement utilisée dans le domaine de la Traduction Automatique Statistique (TAS). Plusieurs travaux de recherche en TAS ont confirmé que les modèles fondés sur *des séquences de mots* [Och and Ney, 2000, Koehn, 2004] permettent d'avoir des performances meilleures que ceux fondés sur les mots [Brown *et al.*, 1993]. Toutefois, dans le domaine de la TAS, il est indispensable d'utiliser des corpus d'apprentissage de très grande taille pour obtenir des résultats intéressants. Ce type de corpus représente un vrai défi pour la communauté de l'ECT pour adapter les algorithmes de fouille de séquences à des contextes d'extraction textuels très denses.

De ce fait, l'ECT offre des techniques complémentaires pour contribuer à l'amélioration des modèles de TAS, à savoir : (*i*) l'exploration des séquences de mots par des méthodes de fouille de séquences ; et, (*ii*) l'intégration de ces motifs séquentiels dans un modèle de traduction par le biais de règles d'association. Le croisement de la TAS avec le problème de l'extraction de motifs séquentiels est justifié par l'idée clé, propre à l'extraction de motifs séquentiels, permettant de distinguer à la fois, à l'intérieur des phrases du corpus, *un ordre d'apparition* des termes mais aussi de regrouper certains termes. Dans ce contexte, les règles d'association permettent l'extraction de règles *intra-phrases* alors que la recherche de motifs séquentiels permet l'extraction de règles *inter-phrases*. Ainsi, la notion de motifs séquentiels reste intuitivement applicable à

la TAS, puisqu'il existe une relation d'ordre entre les termes dans les corpus parallèles, et par conséquent l'ordre d'apparition des termes dans une phrase peut être pris en compte.

## 2 Positionnement : Analyse Formelle de Concepts et ECT

L'extraction des connaissances à partir de textes (ECT) a constitué le noyau de ma thèse de doctorat intitulée *"Approche de découverte de règles d'association classiques et floues à partir de textes : Application à la Recherche d'Information"* [Latiri, 2004]. Ces travaux se sont poursuivis durant les huit dernières années avec un objectif bien précis, à savoir extraire d'autres motifs fréquents à partir de larges corpus textuels et montrer l'utilité des connaissances additionnelles découvertes dans le cadre d'applications réelles, telles que la RI ou la TAS.

Une telle démarche s'inscrit dans une double problématique : (*i*) définir les algorithmes adéquats pour la fouille de corpus de grandes tailles en prenant en compte le problème d'adaptation et d'optimisation du processus d'extraction de motifs intéressants ; (*ii*) déployer les motifs découverts dans des applications réelles.

En effet, sur l'ensemble de nos travaux de recherche, nous abordons et nous discutons la notion centrale de "*connaissance extraite à partir de textes*". Nous désignons par *connaissance* tout motif qui peut être découvert à partir d'un corpus textuel. Cette connaissance peut être déclinée en plusieurs motifs fréquents, tels qu'un ensemble de termes fréquents dans un corpus que nous appelons *termset*[9], *une séquence fréquente de termes* ou encore *une règle d'association entre termes* validée par des mesures statistiques, tels que le support et la confiance [Agrawal and Skirant, 1994]. Nous avons ainsi considéré une granularité textuelle variable au niveau de l'analyse d'un corpus qui peut être le document, la phrase ou le mot. Chaque forme de connaissance fait appel à une algorithmique dédiée à son extraction et trouve son usage et son utilité dans des applications diverses liées à des domaines nécessitant l'utilisation de tels motifs textuels fréquents, tels que la RI, l'IC ou encore la TAS.

Nos contributions de recherche s'articulent autour de l'extraction de motifs fréquents à partir de textes. Tout d'abord, nous nous focalisons sur l'aspect algorithmique en utilisant l'Analyse Formelle de Concepts (AFC) [Wille, 1989] comme fondement mathématique pour définir les motifs fréquents et les méthodes d'extraction à partir de textes. Nous proposons également une nouvelle base générique de règles d'association entre termes dédiée à l'ECT [Latiri *et al.*, 2005a, Latiri *et al.*, 2006]. Nous montrons, par la suite, l'utilité et l'intérêt de l'utilisation de cette base générique de règles d'association dans deux applications liées à la RI, à savoir l'expansion de requêtes [Latiri *et al.*, 2012b] et l'indexation conceptuelle basée sur l'enrichissement d'une ontologie de domaine [Ben Ghezaiel *et al.*, 2010, Ben Ghezaiel *et al.*, 2012]. D'autre part, nous étendons la définition de la base générique de règles d'association vers les règles d'association inter-langues (RAILs), où nous proposons d'extraire les séquences inter-langues fréquentes à partir d'un corpus parallèle de grande taille, que nous utilisons par la suite pour définir un nouveau modèle de TAS à base de séquences [Latiri *et al.*, 2010b, Latiri *et al.*, 2011].

Pour chacune de nos propositions, nous nous attachons à définir les concepts associés et à développer les algorithmes permettant leur mise en œuvre. Tous ces travaux ont donné lieu à des évaluations sur des collections de test utilisées par la communauté RI ou celle de la TAS ou encore sur des bases de données synthétiques.

Il importe de préciser que, dans le cadre de nos recherches, nous nous sommes intéressés principalement à l'application des fondements mathématiques de l'Analyse Formelle de Concepts

9. Par analogie à la terminologie *itemset* utilisée dans le domaine de data mining pour désigner un ensemble d'attributs.

(AFC) [Wille, 1989, Ganter and Wille, 1999] dans l'extraction de motifs fréquents à partir de textes. Ainsi, dans le contexte de la fouille de textes, l'AFC définit *un concept formel* par un ensemble d'objets, (*i.e.,* son *extension* qui est un ensemble de documents ou de phrases) auquel s'applique un ensemble d'attributs, (*i.e.,* son *intention* qui est un ensemble de termes ou de séquences de termes). Dans [Wille, 1989], Wille utilise la notion centrale de *treillis de concepts* ou *treillis de Galois* et l'applique tant à la découverte de concepts, qu'à l'acquisition de connaissances, et à la classification d'objets. Ainsi, dans le domaine de l'ECT, le treillis de Galois peut être vu comme un regroupement conceptuel et hiérarchique de documents (à travers les extensions du treillis), et interprété comme une représentation de toutes les implications entre les termes (à travers les intentions) [Stumme *et al.*, 2002].

Durant la dernière décennie, la problématique de découverte de concepts formels à partir de textes a suscité beaucoup d'intérêt dans diverses communautés de recherche [Carpineto and Romano, 1996, Koester, 2006, Zhang *et al.*, 2006]. Le premier cadre applicatif est venu de la RI pour s'étendre par la suite à l'IC et plus particulièrement aux ontologies et enfin à la traduction automatique. Dans [Priss and Old, 2009], les auteurs ont présenté une revue de la littérature sur les idées principales de l'utilisation du treillis de Galois dans la RI et de la traduction automatique.

## 2.1 Application de l'AFC à la Recherche d'Information

La première application directe de l'AFC dans le domaine de la RI a été introduite par Carpineto et Romano [Carpineto and Romano, 1996]. L'idée a été de représenter simplement les objets du contexte formel par les textes d'une collection documentaire, caractérisés par les mots-clés qu'ils contiennent. Le treillis de Galois définit ainsi l'espace de recherche qu'il est possible d'explorer. De la même manière, la requête est formulée par un ensemble de mots-clés. Le calcul de la réponse à une requête consiste à projeter la requête dans le treillis. S'il existe un concept dont l'intention correspond exactement à la requête, la réponse est donnée par l'extension du concept. Sinon, les concepts subsumants, ou plus généralement les concepts voisins du concept de la requête seront des réponses approximatives satisfaisantes. Cette proposition a réussi à contourner le modèle classique de représentation vectorielle des textes pour se baser sur une représentation booléenne (présence/absence d'un mot-clé dans un document) [Carpineto and Romano, 2000].

Ainsi, depuis les travaux de Carpineto et Romano [Carpineto and Romano, 1996], plusieurs systèmes de RI à base d'AFC ont vu le jour. Parmi eux, le système REFINER [Carpineto *et al.*, 1998] qui ne construit que les concepts voisins de la requête au lieu de construire le treillis complet, ainsi que le système CREDO [Carpineto *et al.*, 2004] qui construit un treillis à partir de documents extraits par Google sur le web. Dans la même idée que le système REFINER, nous citons également FooCA [Koester, 2006] ou encore SearchSleuth [Ducrou and Eklund, 2007] qui proposent de reformuler une requête. Plus récemment, Nauer et Toussaint ont développé le système CreChainDo [Nauer and Toussaint, 2010] qui reprend l'idée du système CREDO. Ils se sont basés sur le fait que les concepts du treillis correspondent à une vision synthétique d'un domaine et permettent de guider de façon dynamique la recherche de documents sur le web.

Une autre approche de RI à base d'AFC a été introduite dans [Polaillon *et al.*, 2007]. Il s'agit d'une approche de navigation contextuelle et sémantique pour la RI à travers un treillis de concepts, où chaque concept correspond à un cluster de pages web partageant un ensemble de termes communs. L'aspect sémantique est illustré par l'utilisation d'un thésaurus du domaine permettant d'affecter des étiquettes à chaque concept du treillis.

En revanche, dans un contexte flou, nous avons proposé dans [Latiri *et al.*, 2004] un nouveau

schéma de correspondance en RI qui est basé sur une extension de la connexion de Galois floue. En intégrant les différentes sémantiques que peuvent avoir les degrés attribués aux termes d'une requête ainsi qu'un quantificateur flou dans la nouvelle définition de la connexion de Galois floue, nous avons discuté du choix de l'implication floue afin de sélectionner celle qui est la plus appropriée au domaine de la RI. Plus récemment, une généralisation des opérateurs de dérivation de Galois en recherche d'information basée sur l'AFC a été proposée dans [Djouadi, 2011].

## 2.2 Ontologies et Analyse Formelle de Concepts

Les ontologies ont été largement considérées dans plusieurs domaines tels que l'IC et l'ECT, comme un modèle avancé de représentation et de partage de connaissances. Dans [Cimiano *et al.*, 2005, Zhang *et al.*, 2006], les auteurs montrent la relation étroite qui existe entre les ontologies et l'AFC. Ils expliquent ainsi la complémentarité entre les deux paradigmes d'un point de vue applicatif. En particulier, ils étayent les différentes contributions de l'AFC dans la construction et l'enrichissement d'ontologies, ainsi que leur déploiement dans les applications à base d'AFC. D'après [Cimiano *et al.*, 2005], la synergie entre l'AFC et les ontologies peut être transposée au cycle de vie d'une ontologie comme suit :

1. *L'AFC peut supporter la construction d'ontologies dans la phase d'apprentissage.* À titre d'exemple, dans [Latiri *et al.*, 2005b], nous avons proposé une approche semi-automatique de construction d'ontologie de termes à partir du treillis de l'Iceberg de Galois [Stumme *et al.*, 2002], permettant un passage direct d'une structure hiérarchique partiellement ordonnée vers une structure ontologique. Dans la même famille d'approches, dans [Bendaoud *et al.*, 2008], les auteurs ont présenté une approche semi-automatique de construction d'ontologie à partir de corpus de textes pour un domaine spécifique. Cette approche repose sur la classification d'objets d'après les propriétés qu'ils partagent en utilisant l'AFC, pour la construction d'un treillis de concepts. Ce treillis sert à construire un noyau d'ontologie. Les auteurs ont proposé aussi une méthode originale qui enrichit cette ontologie avec des relations transversales en utilisant l'Analyse Relationnelle de Concepts (ARC) [Moha *et al.*, 2008]. Une autre contribution, décrite dans [Stumme, 2005], aborde la problématique de fusion d'ontologies existantes par le biais de l'AFC. En revanche, dans [Mondary *et al.*, 2008], les auteurs ont présenté une critique des travaux proposés dans [Cimiano *et al.*, 2005] et dans [Bendaoud *et al.*, 2008], en démontrant que l'AFC n'est pas toujours utilisable pour certains corpus. Les auteurs se sont ainsi intéressés à l'évaluation d'une méthode permettant l'acquisition de hiérarchies à partir de textes.
2. *Une ontologie existante peut être analysée et explorée moyennant des techniques de l'AFC.* Par exemple, dans [Cole II *et al.*, 2003], les auteurs ont montré la possibilité de stocker les courriers électroniques dans une ontologie spécialisée et de l'explorer par la suite. L'ontologie utilisée consiste en un treillis de concepts couplé avec un lexique. Les courriers électroniques peuvent être ainsi attribués à plusieurs concepts.
3. *Une ontologie peut être utilisée pour améliorer l'efficacité des applications à base d'AFC.* Dans les définitions standards de l'AFC, l'ensemble des attributs n'obéit à aucune structure. En considérant ce dernier comme un ensemble de concepts d'une ontologie, il est possible de modéliser les relations et les dépendances entre les attributs. Cette modélisation permet d'enrichir la structure conceptuelle du treillis et offre de nouvelles possibilités d'analyse. À titre d'exemple, Hotho *et al.* ont étudié dans [Hotho *et al.*, 2003], l'utilisation d'une ressource ontologique de connaissances pour la classification de documents.

### 2.3 Analyse Formelle de Concepts et Traduction Automatique

La rencontre entre l'AFC et le domaine de la Traduction Automatique (TA) est plus récente et les travaux dans ce contexte sont moins abondants. Elle a débuté avec la contribution publiée dans [Janssen, 2003] où l'auteur a proposé le système SIM$u$LLDA, conçu pour gérer une base de données lexicales multilingue qui permet de générer automatiquement des dictionnaires bilingues. La base de SIM$u$LLDA est pilotée par une structure du treillis de concepts issue de l'AFC. Par ailleurs, dans [Priss and Old, 2007], les auteurs ont mis exergue l'usage de l'AFC dans la construction de réseaux d'associations bilingues pour la traduction automatique. Ces réseaux illustrent en effet les relations existantes entre les mots apparentés à des langues différentes. Une autre contribution plus récente décrite dans [Falk and Gardent, 2011] introduit une nouvelle approche de classification de verbes, englobant l'utilisation du lexique Anglais-Français VerbNet [Kipper *et al.*, 2008], et une fusion de lexiques de sous-catégorisation pour comparer les schémas syntaxiques des verbes. De même, dans [Kiliçaslan and Güner, 2011], les auteurs ont montré que les treillis de concepts peuvent améliorer considérablement les performances des systèmes de TA, lorsqu'ils sont appliqués comme filtres sur leurs résultats. Le treillis de concepts est ainsi utilisé pour accomplir la tâche de désambiguïsation.

## 3 Contributions de recherche

La synthèse de nos différents travaux de recherche a été présentée dans la section 3 (*cf.* page 15) selon trois principaux axes complémentaires. Nos contributions seront détaillées dans les prochains chapitres et sont résumées comme suit :

1. **Définition d'une nouvelle base générique de règles d'association entre termes :** Étant donné que notre objectif principal est d'extraire des règles d'association entre termes à partir de corpus de textes volumineux, nous avons proposé la formalisation et l'extraction d'une nouvelle base générique minimale de règles d'association non-redondantes entre termes, appelée $\mathcal{MGB}$, permettant d'assurer un compromis entre l'informativité et la compacité de cette dernière [Latiri *et al.*, 2012b]. La spécificité de la base proposée est qu'elle est compacte, dans le sens où elle englobe un noyau minimal de règles d'association approximatives et exactes non-redondantes entre termes. Ces règles sont dérivées à partir de la structure ordonnée du *treillis de l'Iceberg de Galois augmenté* [Stumme *et al.*, 2002], *i.e.*, la partie supérieure du treillis de Galois ne conservant que les termsets fermés fréquents [Mephu Nguifo, 1994] et qui sont à leur tour "décorés" par l'ensemble de leurs générateurs minimaux [Latiri *et al.*, 2005a, Latiri *et al.*, 2006]. La caractéristique clé des règles dérivées est qu'elles ont des prémisses minimales illustrées par les générateurs minimaux et des conclusions maximales. Dans ce contexte, les générateurs sont utilisés dans les prémisses des règles d'association découvertes, tandis que les termsets fermés fréquents permettent la dérivation des conclusions de ces règles. Il importe de souligner que la relation de précédence, définie dans le treillis de l'Iceberg de Galois [Mephu Nguifo, 1994], aide à réduire le coût d'extraction des règles d'association, en évitant certaines combinaisons redondantes lors de la génération des candidats. L'étude empirique que nous avons menée sur cinq collections de documents a fait ressortir que la base $\mathcal{MGB}$ apporte des gains substantiels en terme de compacité par rapport aux bases génériques extraites avec et sans perte d'information. Nous avons discuté dans [Latiri *et al.*, 2012b] les résultats des évaluations expérimentales par rapport à la nature du contexte d'extraction textuel, *i.e.,* dense et épars. Cette contribution est détaillée dans le chapitre 2 (*cf.* page 67).

2. **Déploiement de la base générique de règles d'association entre termes pour l'enrichissement d'ontologies et l'indexation conceptuelle en RI :** Nous avons essayé de mettre en exergue, à travers nos contributions, le croisement de l'ECT avec le domaine de l'Ingénierie de Connaissances (IC) et celui de la Recherche d'Information (RI). L'étude proposée est consacrée à l'exploitation de la base générique de règles d'association entre termes $\mathcal{MGB}$ pour l'amélioration des résultats de la RI. Nos contributions abordent, d'une part la problématique d'interrogation en RI travers l'utilisation de la base générique $\mathcal{MGB}$ dans une approche d'expansion automatique de requêtes [Latiri *et al.*, 2012b] et, d'autre part, la problématique d'indexation en RI par la proposition d'une nouvelle approche d'indexation conceptuelle basée sur une ontologie de domaine enrichie par la dite base [Ben Ghezaiel *et al.*, 2010, Ben Ghezaiel *et al.*, 2012].

   Dans un premier temps, il s'agit de dériver la base générique $\mathcal{MGB}$ de règles d'association non-redondantes entre termes à partir d'une collection de documents, et de l'utiliser, dans un deuxième temps, pour étendre la requête originelle de l'utilisateur. Il importe de souligner que la base générique $\mathcal{MGB}$ est mieux adaptée au processus d'expansion automatique de requêtes, dans lesquelles l'ensemble des termes originels sera étendu par les termes des conclusions des règles valides de la base générique $\mathcal{MGB}$, et ayant les termes de la requête initiale dans leurs prémisses respectives. Rappelons que les règles d'association entre termes de la base $\mathcal{MGB}$ sont non-redondantes et qu'elles sont dotées d'une prémisse minimale et une conclusion maximale, ce qui offre plus de termes candidats pour l'expansion. Dans ce cadre et contrairement à la technique d'analyse locale de co-occurrences de termes utilisée en RI, les règles d'association entre termes se dérivent par un processus de fouille globale de toute la collection de documents. Elles offrent, par conséquent, des informations sur les corrélations de termes inter-documents ainsi que sur les relations intra-document. De ce fait, les règles d'association permettent d'expliciter des relations plus fines entre les termes que les approches classiques à base de co-occurrences de termes. L'approche d'expansion proposée est détaillée au niveau du chapitre 3 (*cf.* page 85).

   La deuxième contribution s'adresse à une problématique largement abordée en RI, à savoir l'imperfection de l'indexation automatique des documents. Nous avons ainsi introduit une nouvelle approche d'indexation conceptuelle en RI qui est fondée sur l'enrichissement d'une ontologie de domaine à travers l'utilisation de règles d'association de la base générique $\mathcal{MGB}$ [Ben Ghezaiel *et al.*, 2010, Ben Ghezaiel *et al.*, 2012]. Nous abordons ainsi la question de l'utilisation des ontologies pour la RI du point de vue de l'ingénierie des connaissances, en particulier sur leur mode d'enrichissement et sur la nature de leur contenu en terme de connaissances représentées. Le résultat généré par le processus d'enrichissement d'ontologie à base de règles d'association est exploré en tant que *réseau conceptuel proxémique* [Latiri *et al.*, 2012a]. L'originalité de ce réseau réside dans sa généricité et son exhaustivité, qui sont dûes à la combinaison des connaissances explicites de la structure ontologique d'une part, et des connaissances implicites issues de l'application de la technique d'extraction de règles d'association, d'autre part [Ben Ghezaiel *et al.*, 2010].

   Du point de vue applicatif, nous avons mis en évidence l'utilisation du réseau conceptuel proxémique en RI, par la proposition d'une nouvelle approche d'indexation conceptuelle [Ben Ghezaiel *et al.*, 2012, Latiri *et al.*, 2012a]. En effet, nous suggérons d'utiliser les relations entre les concepts du réseau proxémique, qui sont pondérées par une nouvelle mesure de similarité sémantique que nous avons définie. Ces relations sont utilisées, dans un premier temps, dans le processus de désambiguïsation des termes d'un document lors de la phase d'indexation, et dans un deuxième temps, pour affecter un poids sémantique

aux différents descripteurs des documents [Ben Ghezaiel *et al.*, 2012, Latiri *et al.*, 2012a]. L'approche est détaillée au niveau du chapitre 3 (*cf.* page 85).

3. **Règles d'association inter-langues pour la Traduction Automatique Statistique :** Notre contribution émane de notre conviction qu'il est possible de proposer d'autres approches issues du domaine de l'ECT, capables de constituer une alternative aux méthodes de Traduction Automatique Statistique (TAS) d'IBM [Brown *et al.*, 1993] et de proposer en conséquence un nouveau modèle de traduction à base de séquences [Koehn *et al.*, 2007]. Notre objectif est d'étendre les approches d'extraction des motifs fréquents pour supporter des corpus parallèles de très grande taille et d'extraire des connaissances utiles et pertinentes pour la TAS.

   L'originalité de notre proposition repose ainsi sur le couplage de deux types de connaissances extraites par un processus d'ECT, à savoir les règles d'association entre termes [Latiri *et al.*, 2012b] et les séquences fréquentes fermées de termes [Dong and Pei, 2007]. Ce couplage a donné naissance à la définition d'un nouveau concept que nous avons appelé *Règles d'Association Inter-Langues* (RAILs) [Latiri *et al.*, 2010b]. Ce concept n'est autre qu'une extension de la notion de règle d'association telle qu'elle a été définie par Agrawal et *al.* [Agrawal and Skirant, 1994]. Intuitivement, l'interprétation d'une règle inter-langues dans le domaine de la TAS est que sa *conclusion*, représentant une unité linguistique dans une langue cible, est une traduction potentielle de sa *prémisse*, représentant une autre unité linguistique dans la langue source. L'idée clé est basée sur l'extraction de séquences fréquentes dans le domaine l'ECT en prenant en compte l'ordre d'apparition des mots dans la phrase, et ce à partir de corpus de textes parallèles alignés au niveau de la phrase. Nous nous sommes intéressés particulièrement à la famille des approches dédiées à l'extraction *des motifs séquentiels fermés fréquents* [Yan *et al.*, 2003, Wang and Han, 2004, Chang, 2004], inspirées de la recherche d'itemsets fermés fréquents [Pasquier *et al.*, 1999]. Ainsi, la génération de l'ensemble de règles d'association inter-langues a induit la mise en place d'un nouveau modèle de traduction à base de séquences de mots [Latiri *et al.*, 2010a, Latiri *et al.*, 2011], qui est décrit en détail au niveau du chapitre 4 (*cf.* page 109).

## 4 Organisation du mémoire

L'ensemble de nos contributions sont résumées dans ce mémoire, composé de quatre chapitres, comme suit :

- **Le Chapitre 1** est consacré à l'état de l'art sur l'extraction de motifs fréquents à partir de textes. Il dresse un panorama des approches existantes à base de l'Analyse Formelle de Concepts.
- **Le Chapitre 2** décrit la formalisation et le processus de génération d'une nouvelle base générique. Cette dernière englobe un noyau réduit de règles d'association non-redondantes entre termes, ayant la particularité d'avoir une prémisse minimale et une conclusion maximale. Ce chapitre présente, dans un premier temps, les bases génériques les plus citées dans la littérature. Il expose ensuite notre définition de la base générique minimale ($\mathcal{MGB}$) ainsi que son évaluation empirique sur un ensemble de collections documentaires. Le chapitre se termine par une étude comparative entre notre base et celles de la littérature.
- **Le Chapitre 3** traite de l'utilisation de la base générique minimale $\mathcal{MGB}$ dans le domaine de la Recherche d'Information. Nous abordons les deux problématiques centrales de la RI, à savoir l'expansion de requêtes par les règles d'association de la base $\mathcal{MGB}$, et l'indexation conceptuelle en RI par la proposition d'un réseau conceptuel proxémique résultant de

l'enrichissement d'une ontologie de domaine par la base $\mathcal{MGB}$. Après avoir introduit les travaux liés dans la littérature, le chapitre détaille nos propositions ainsi que leur évaluation expérimentale sur des collections de test largement utilisées par la communauté RI.

– **Le Chapitre 4** met en valeur les liens existants entre l'ECT et le domaine de la Traduction Automatique Statistique (TAS). Après une brève revue sur la TAS, il développe la problématique de fouille de séquences à partir de corpus parallèles. Le chapitre détaille ensuite notre proposition d'extraction de règles d'association inter-langues et leur utilisation pour la définition d'un nouveau modèle de TAS à base de séquences. Une évaluation empirique est présentée sur le corpus parallèle de référence EUROPARL [Rama and Borin, 2011] ainsi qu'une étude comparative avec d'autres modèles de TAS de l'état de l'art.

# Chapitre 1

# État de l'art

**Sommaire**



## 1.1 Objectifs du chapitre

Ce chapitre est consacré à un état de l'art sur l'extraction de motifs fréquents à partir de textes. Il est organisé comme suit : après la présentation des fondements mathématiques issus de l'Analyse Formelle de Concepts (AFC), nous introduirons, dans un premier temps, l'ensemble des concepts et des définitions adaptés à l'Extraction des Connaissances à partir de Textes (ECT). Nous passerons ensuite en revue les principales approches d'extraction des motifs fréquents en considérant un contexte d'extraction textuel. Nous nous intéresserons principalement à trois types de motifs, à savoir : les termsets fermés fréquents, les règles d'association entre termes et les séquences de termes. Par la suite, nous discuterons des problèmes posés par les approches d'extraction dans le contexte de corpus textuels volumineux et non structurés ainsi que les contributions proposées pour étendre l'extraction des motifs fréquents sur du texte et leur utilisation dans des applications réelles.

## 1.2 Fondements mathématiques de l'AFC

L'Analyse Formelle de Concepts (AFC) [Wille, 1989, Ganter and Wille, 1999] définit *un concept formel* par un ensemble d'objets (ou extension), auquel s'applique un ensemble d'attributs (ou intention). Dans [Wille, 1982, Wille, 1989], Wille utilise la notion centrale de *treillis de*

*concepts* ou *treillis de Galois* et l'applique tant à la découverte de concepts, qu'à l'acquisition de connaissances, et à la classification d'objets. Ainsi, le treillis de Galois peut être vu comme un regroupement conceptuel et hiérarchique d'objets (à travers les extensions du treillis), et interprété comme une représentation de toutes les implications entre les attributs (à travers les intentions) [Stumme *et al.*, 2002]. Dans le cadre de nos recherches, nous nous intéressons à l'application de l'AFC dans l'extraction de règles d'association entre termes.

### 1.2.1 Cadre formel et notations

Pour une meilleure compréhension des définitions formelles énoncées dans le reste du présent document, nous résumons dans la TABLE 1.1 les notations utilisées.

| Notation | Description |
|---|---|
| $\mathcal{C}$ | ***l'ensemble de tous*** les documents qui définissent une collection |
| $D$ | un ***ensemble*** de documents appartenant à une collection ($D \subseteq \mathcal{C}$) |
| $d$ | un ***document unique*** de la collection ($d \in \mathcal{C}$) |
| $\mathcal{T}$ | ***l'ensemble de tous*** les termes ***distincts*** de la collection $\mathcal{C}$ |
| $T$ | un ***ensemble*** de termes de la collection ($T \subseteq \mathcal{T}$) |
| $t$ | un ***terme*** de la collection $\mathcal{C}$ ($t \in \mathcal{T}$) |
| $TFF$ | un termset fermé fréquent |
| $\mathcal{TFF}$ | l'ensemble des termsets fermés fréquents |
| $g$ | générateur d'un termset fermé fréquent |
| $\mathcal{TA}$ | treillis de l'Iceberg de Galois augmenté |

TABLE 1.1 – Résumé des notations.

Dans le cadre de nos travaux de recherche portant sur l'ECT, nous utilisons comme cadre théorique l'AFC. De ce fait, nous commençons par formaliser un contexte d'extraction constitué de documents et de termes d'index, appelé *contexte d'extraction textuel*.

**Définition 1** [Latiri *et al.*, 2012b] Un *contexte d'extraction textuel* est un triplet $\mathfrak{M} := (\mathcal{C}, \mathcal{T}, \mathcal{I})$ tel que :

- $\mathcal{C} = \{d_1, d_2, \ldots, d_n\}$ est un ensemble fini des $n$ documents de la collection.
- $\mathcal{T} = \{t_1, t_2, \cdots, t_m\}$ est un ensemble fini des $m$ termes distincts d'une collection. L'ensemble $\mathcal{T}$ englobe en effet, sans les dupliquer, les termes des différents documents de la collection.
- $\mathcal{I} \subseteq \mathcal{C} \times \mathcal{T}$ est une relation binaire (d'incidence). Chaque couple $(d, t) \in \mathcal{I}$ indique que le document $d \in \mathcal{C}$ contient le terme $t \in \mathcal{T}$.
- Un **termset** $T \in \mathcal{T}$ est un sous-ensemble de termes de $\mathcal{T}$ qui co-occurrent ensemble dans une collection de documents $\mathcal{C}$.

**Exemple 1** *Considérons le contexte illustré par la* TABLE *1.2 et qui sera utilisé comme un exemple illustratif tout au long de ce chapitre. Nous avons* $\mathcal{C} = \{d_1, d_2, d_3, d_4, d_5, d_6\}$ *et* $\mathcal{T} = \{A, C, D, T, W\}$. *Le couple* $(d_2, C) \in \mathcal{I}$ *signifie que le document* $d_2$ *contient le terme C.*

| $\mathcal{I}$ | A | C | D | T | W |
|---|---|---|---|---|---|
| $d_1$ | × | × | | × | × |
| $d_2$ | | × | × | | × |
| $d_3$ | × | × | | × | × |
| $d_4$ | × | × | × | | × |
| $d_5$ | × | × | × | × | × |
| $d_6$ | | × | × | × | |

TABLE 1.2 – Contexte d'extraction textuel $\mathfrak{M} := (\mathcal{C}, \mathcal{T}, \mathcal{I})$.

À titre d'exemple, $\{ACW\}$ est un termset composé des termes A, C et W. Le *support* d'un termset est défini comme suit :

**Définition 2** [Latiri *et al.*, 2012b] Soit $T \subseteq \mathcal{T}$. Le support de $T$ par rapport au contexte d'extraction $\mathfrak{M}$ est égal au nombre de documents de $\mathcal{C}$ contenant tous les termes de $T$, soit :

$$Supp(T) = |\{d | d \in \mathcal{C} \ \wedge \ \forall \ t \in T : (d, \ t) \in \mathcal{I}\}| \tag{1.1}$$

$Supp$(T) est appelé le support *absolu* de $T$. Le support *relatif* (fréquence) de $T$ est égal à $\frac{Supp(T)}{|\mathcal{C}|}$.

Un termset est dit *fréquent* (*large*) si les co-occurrences de ses termes dans la collection sont supérieures ou égales à un seuil de support minimal prédéfini appelé *minsupp*. Dans le cas contraire, le termset est dit *non fréquent* (*rare*).

**Exemple 2** *Considérons le contexte textuel de la* TABLE *1.2 et un minsupp égal à 3. Le termset* $\{AC\}$ *est fréquent puisque* $Supp(AC) = 4 \geq 3$. *Par ailleurs, le termset* $\{CDT\}$ *est non fréquent car* $Supp(CDT) = 2 < 3$.

### 1.2.2 Définitions de base

Nous considérons comme contexte formel d'extraction le contexte textuel $\mathfrak{M} = (\mathcal{C}, \mathcal{T}, \mathcal{I})$ auquel nous adaptons les différentes définitions issues de l'AFC [Wille, 1989, Ganter and Wille, 1999].

#### Correspondance de Galois

Deux fonctions sont définies pour correspondre un ensemble de documents à un ensemble de termes et *vice versa*. Pour un termset $T \subseteq \mathcal{T}$, nous définissons :

$$\Psi(T) = \{d | d \in \mathcal{C} \ \wedge \ \forall \ t \in T, (d, \ t) \in \mathcal{I}\} \tag{1.2}$$

comme l'ensemble des documents contenant tous les termes de $T$. Sa cardinalité est alors égale à $Supp(T)$.

Pour un ensemble de documents $D \subseteq \mathcal{C}$, nous définissons :

$$\Phi(D) = \{t | t \in \mathcal{T} \ \wedge \ \forall \ d \in D, (d, \ t) \in \mathcal{I}\} \tag{1.3}$$

comme l'ensemble des termes apparaissant dans tous les documents de $D$.

Les deux fonctions $\Psi$ et $\Phi$ forment *les opérateurs de la connexion de Galois* entre les ensembles $\mathcal{P}(\mathcal{T})$ et $\mathcal{P}(\mathcal{C})$. Il en résulte l'opérateur $\Omega = \Phi \circ \Psi$, appelé *opérateur de fermeture de Galois* qui associe à un termset $T$ l'ensemble de tous les termes qui apparaissent dans *tous* les documents où les termes de $T$ co-occurrent. Cet ensemble de termes est égal à $\Omega(T)$. En effet, $\Omega(T) = \Phi \circ \Psi(T) = \Phi(\Psi\ (T))$. Si $\Psi\ (T) = D$, alors $\Omega(T) = \Phi(D)$.

**Exemple 3** *Considérons le contexte textuel de la* TABLE *1.2. Étant donné que les deux termes A et C apparaissent simultanément dans les documents* $d_1$, $d_3$, $d_4$, *et* $d_5$, *nous avons :* $\Psi(AC) = \{d_1, d_3, d_4, d_5\}$. *D'un autre côté, puisque les documents* $d_1$, $d_3$, $d_4$, *et* $d_5$ *partagent les termes A, C, et W, nous avons :* $\Phi(\{d_1, d_3, d_4, d_5\}) = \{ACW\}$. *Il en résulte que* $\Omega(AC) = \Phi \circ \Psi(AC) = \Phi(\Psi(AC)) = \Phi(\{d_1, d_3, d_4, d_5\}) = \{ACW\}$. *Ainsi,* $\Omega(AC) = \{ACW\}$. *Autrement dit, le terme W apparaît dans tous les documents où A et C co-occurrent.*

### Concept formel

Un concept formel est une paire $c = (D, T)$, tel que $D$ est un ensemble de documents, appelé *extension*, et $T$ est un termset, appelé *intention*. Ainsi, $D$ et $T$ sont reliés à travers les opérateurs de la connexion de Galois, *i.e.,* $\Phi(D) = T$ et $\Psi(T) = D$.

### Termset fermé fréquent

Un termset $T \subseteq \mathcal{T}$ est dit *fermé* si $\Omega(T) = T$. Il représente l'ensemble maximal de termes communs à un ensemble donné de documents. Un termset fermé est appelé *fréquent* par rapport à un seuil de support *minsupp* si $Supp(T) = |\Psi(T)| \geq minsupp$ [Pasquier *et al.*, 2005]. Dans la suite du mémoire, nous notons par TFF un termset fermé fréquent.

**Exemple 4** *Dans la continuité de l'Exemple 3,* $\{ACW\}$ *est un termset fermé puisqu'il n'existe pas un autre terme qui apparaît dans tous les documents contenant* $\{ACW\}$. *Ainsi,* $\{ACW\}$ *est l'ensemble maximal de termes communs aux documents* $\{d_1, d_3, d_4, d_5\}$. *Nous avons ainsi :* $\Omega(ACW) = \{ACW\}$. *Dans le cas où le seuil minsupp est fixé à 3,* $\{ACW\}$ *est dit aussi fréquent puisque* $|\Psi(ACW)| = |\{d_1, d_3, d_4, d_5\}| = 4 \geq 3$.

La Propriété 1 ci-dessous établit la relation entre le support d'un termset et celui de sa fermeture.

**Propriété 1** *Le support d'un termset T est égal au support de sa fermeture* $\Omega(T)$, *qui est le plus petit* TFF *contenant T, i.e.,* $Supp(T) = Supp(\Omega(T))$ [Bastide *et al.*, 2000a].

**Exemple 5** *Puisque* $\Omega(AC) = \{ACW\}$, *nous avons :* $Supp(AC) = Supp(ACW) = 4$.

### Générateur minimal

Un termset $g \subseteq \mathcal{T}$ est un *générateur minimal* d'un termset fermé $T$, si et seulement si $\Omega(g) = T$ et $\nexists\ g' \subset g$ tel que $\Omega(g') = T$ [Bastide *et al.*, 2000a].

**Exemple 6** *Le termset* $\{DW\}$ *est un générateur minimal de* $\{CDW\}$ *étant donné que* $\Omega(DW) = \{CDW\}$ *et aucun de ses sous-ensembles propres a le termset* $\{CDW\}$ *comme fermeture. En effet,* $\Omega(D) = \{CD\}$ *et* $\Omega(W) = \{CW\}$.

**Corollaire 1** *Soit g un générateur minimal d'un termset fermé fréquent T. D'après la Propriété 1, le support de g est égal au support de sa fermeture, i.e.,* $Supp(g) = Supp(T)$.

Ainsi, un termset fermé fréquent apparaît dans le même ensemble de documents et par conséquent il a le même support que celui de ses générateurs. Il représente donc un ensemble maximal de termes partageant les mêmes documents, tandis que ses générateurs minimaux représentent les plus petits éléments décrivant l'ensemble de documents. De ce fait, nous pouvons dire qu'un termset fermé englobe l'expression la plus spécifique qui décrit les documents qui lui sont associés alors que le générateur minimal contient une des expressions les plus générales. Dans la suite du mémoire, pour chaque termset fermé $T$, nous notons par $\mathcal{G}_T$ l'ensemble de ses générateurs minimaux.

**Classe d'équivalence**

L'opérateur de fermeture $\Omega$ induit une relation d'équivalence sur l'ensemble des termsets fermés fréquents noté par $\mathcal{TFF}$, *i.e.,* que l'ensemble $\mathcal{TFF}$ est partitionné en sous-ensembles disjoints appelés aussi *classes*. Les éléments distincts d'une classe d'équivalence donnée apparaissent ainsi dans les mêmes documents et partagent par conséquent la même fermeture et donc le même support. L'unique élément maximal, par rapport à l'inclusion ensembliste, d'une classe d'équivalence est *le termset fermé*, tandis que les éléments minimaux représentent *les générateurs minimaux*.

Ainsi, la localisation d'un termset fermé ou d'un générateur minimal d'un TFF nécessite un voisinage restreint, à savoir ses sur-ensembles immédiats et ses sous-ensembles immédiats, respectivement. Il suffit alors de comparer son support avec ceux des termsets du voisinage associé. Par ailleurs, tout termset est nécessairement compris entre un générateur minimal et le termset fermé associé.

**Treillis de Galois**

Lorsque l'opérateur d'inclusion ensembliste est appliqué sur l'ensemble des concepts formels, noté $\mathcal{C}_{\mathfrak{M}}$, ce dernier forme un treillis complet $\mathcal{L}_c = (\mathcal{C}_{\mathfrak{M}}, \leq)$, appelé *Treillis de Galois* [Ganter and Wille, 1999].

Un ordre partiel peut être défini sur l'ensemble des concepts formels comme suit : $\forall\ c_1, c_2 \in \mathcal{C}_{\mathfrak{M}}, c_1 \leq c_2$ si et seulement si $intention(c_1) \subseteq intention(c_2)$, ou d'une manière équivalente $extension(c_2) \subseteq extension(c_1)$.

Étant donné un concept $c$, nous définissons l'ensemble de ses successeurs immédiats dans le treillis, appelé *couverture supérieure*, comme suit : $Couv^s(c) = \{c_i \in \mathcal{C}_{\mathfrak{M}} | c \preceq c_i\}$, tel que $\preceq$ est la réduction transitive de $\leq$, *i.e.,* $\forall\ c_3 \in \mathcal{C}_{\mathfrak{M}}$, $c_1 \leq c_2 \leq c_3$ implique soit $c_1 = c_3$ ou $c_2 = c_3$.

**Treillis de l'Iceberg de Galois**

Considérons $\mathcal{TFF}$ l'ensemble des termsets fermés fréquents relatifs à un contexte textuel donné. Quand l'ensemble $\mathcal{TFF}$ est partiellement ordonné au sens de l'inclusion ensembliste, la structure hiérarchique résultante est appelée *treillis de l'Iceberg* et désigne le sup-demi treillis de Galois, *i.e.,* la partie supérieure du treillis de Galois ne conservant que les termsets fermés fréquents [Stumme *et al.*, 2002].

Nous appelons *treillis de l'Iceberg de Galois augmenté*, noté par $\mathcal{TA} = (\mathcal{TFF}, \subseteq)$, le treillis de l'Iceberg de Galois standard où chaque termset fermé fréquent lui est associé ses générateurs minimaux [Latiri *et al.*, 2012b].

**Exemple 7** *Considérons le contexte d'extraction de la* Table *1.2. Le seuil de minsupp est fixé à 3. La* Table *1.3 illustre pour chaque termset fermé fréquent, ses générateurs fermés fréquents*

*correspondants ainsi que son support. Le treillis de l'Iceberg de Galois augmenté, associé à cet exemple est donné par la* FIGURE *1.1.*

| **Générateurs minimaux** | **TFF** | **Support** |
|---|---|---|
| C | C | 6 |
| W | CW | 5 |
| D | CD | 4 |
| T | CT | 4 |
| A | ACW | 4 |
| AT/TW | ACTW | 3 |
| DW | CDW | 3 |

TABLE 1.3 – L'ensemble $\mathcal{TFF}$ des termsets fermés fréquents, leurs générateurs minimaux et leurs supports respectifs.

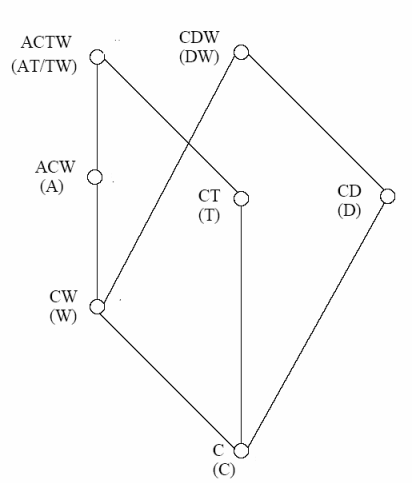


FIGURE 1.1 – Le treillis de l'Iceberg de Galois augmenté.

Chaque termset fermé fréquent $T$, dans le treillis de l'Iceberg, a une *couverture supérieure* qui est constituée des termsets fermés qui sont immédiatement liés à $T$ dans le treillis. Formellement, nous avons :

$$Couv^s(T) = \{T_1 \in \mathcal{TFF} \mid T \subset T_1 \text{ et } \nexists\, T_2 \in \mathcal{TFF} : T \subset T_2 \subset T_1\}$$

**Exemple 8** *Considérons le termset fermé fréquent* $\{CW\}$ *du treillis de l'Iceberg de Galois illustré par la* FIGURE *1.1. Nous obtenons :* $Couv^s(CW) = \{ACW,\ CDW\}$.

Dans ce qui suit, nous allons présenter un survol de la littérature dédiée aux approches d'extraction des termsets fermés fréquents tout en considérant le contexte d'extraction textuel $\mathfrak{M} = (\mathcal{C}, \mathcal{T}, \mathcal{I})$ (*cf.* Définition 1, page 52).

## 1.3 Extraction des termsets fermés fréquents

L'extraction des termsets fréquents à partir de textes constitue une étape primordiale dans divers problèmes de l'extraction des connaissances à partir de textes, tels que l'extraction des règles d'association entre termes. La performance de tout algorithme basé sur une recherche de tels motifs dépend en majeure partie des performances de cette étape.

Dans le domaine du data mining, plusieurs approches ont été proposées dans la littérature, pour trouver les itemsets fréquents dans les bases de données. La première est basée sur l'algorithme Apriori qui fut le premier algorithme, par niveau, à réaliser cette tâche [Agrawal and Skirant, 1994]. Ce dernier identifie les $i$-itemsets à chaque $i^{ème}$ itération puis génère les $(i+1)$-itemsets fréquents à partir des $i$-itemsets. À chaque itération, il requiert un passage sur la base de données pour compter le support des itemsets candidats et ensuite élaguer les candidats infréquents, *i.e.,* ayant des supports strictement inférieurs au seuil minimum *minsupp*. Le processus itératif s'arrête lorsqu'on ne peut plus générer de candidats. La propriété adoptée pour élaguer l'espace des itemsets possibles est une propriété dite *d'anti-monotonie*, qui stipule que *si un itemset n'est pas fréquent alors aucun de ses sur-ensembles ne sera fréquent*. Bien que cette propriété réduise de façon considérable le nombre de candidats à considérer, la tâche reste néanmoins complexe et coûteuse en temps et en espace mémoire [Agrawal and Skirant, 1994].

Toutefois, même si l'algorithme Apriori est très simple et efficace pour des données peu corrélées, il reste inadapté pour des corpus textuels connus par leur densité et leur taille gigantesque dans la majorité des applications manipulant de textes. De plus, il n'est pas toujours possible de charger le corpus de textes en mémoire principale, ce qui occasionne des opérations d'entrées-sorties coûteuses sur des données résidant sur disques.

De ce fait, dans le cadre de nos recherches basées sur l'AFC, nous allons nous intéresser au type d'approches introduisant des représentations plus condensées des itemsets et particulièrement à la recherche de itemsets fermés fréquents dans les bases de données [Pasquier *et al.*, 1999].

Dans le domaine de l'ECT, pour effectuer l'extraction de termsets fermés fréquents à partir d'un contexte d'extraction textuel, il est possible d'adapter différentes approches proposées dans la littérature. En particulier, nous retenons trois types d'approches dont les différences sont liées à la stratégie de parcours de l'espace de recherche, à savoir l'approche par niveau [Pasquier *et al.*, 1999, Bastide *et al.*, 2000a, Bastide *et al.*, 2000b] et l'approche en profondeur d'abord par extension de motifs [Pei *et al.*, 2000, Zaki and Hsiao, 2002].

1. **Parcours de l'espace de recherche en largeur et génération de candidats :**
   La notion de classe d'équivalence introduite précédemment dans la sous-section 1.2.2 (*cf.* page 55) nous permet de déduire que tous les sous-ensembles d'un générateur minimal sont minimaux. La propriété de *minimalité* associée à un générateur est dite ainsi anti-monotone. Par contre, la propriété de fermeture n'est ni monotone, ni anti-monotone. Dans ce cas de figure, comment extraire efficacement les termsets fermés ? Il a été montré que pour extraire les termsets fermés fréquents, il suffisait d'extraire les générateurs minimaux fréquents et de calculer leur fermeture [Pasquier *et al.*, 1999]. Les algorithmes Close [Pasquier *et al.*, 1998], A-Close [Pasquier *et al.*, 1999], Pascal [Bastide *et al.*, 2000b], Titanic [Stumme *et al.*, 2002] et Zart [Szathmary *et al.*, 2007] sont des exemples de ce type d'approches utilisant cette propriété. Le parcours du treillis est effectué par niveau, mais au lieu de retourner les termsets fréquents, les générateurs minimaux fréquents sont dérivés et leurs fermetures sont ensuite calculées, pour générer l'ensemble des termsets fermés fréquents. Ainsi, par rapport à l'algorithme Apriori, tous les termsets fréquents

qui ne sont pas des générateurs minimaux sont élagués, ce qui induit une diminution du nombre de termsets traités. Dans le pire des cas, si tous les termsets sont des générateurs minimaux, il faut donc parcourir autant de termsets qu'avec APRIORI, mais avec un léger sur coût engendré par la vérification de la contrainte sur la minimalité des générateurs. En pratique, sur des données fortement corrélées, comme dans le cas des corpus de textes, beaucoup de termsets sont des générateurs et l'élagage se révèle efficace [Latiri *et al.*, 2012b].

2. **Parcours de l'espace de recherche en profondeur par extension de motifs :**
   Dans ce type d'approches, les algorithmes effectuent un parcours en profondeur de l'espace de recherche. Ils supposent l'utilisation d'une structure de données compacte appelée FP-tree (Frequent Pattern tree). Lorsque l'algorithme a fini de traiter un termset $T$, il examine ses fils par ordre croissant de support. Les fils de $T$ sont les termsets de la forme $t_i$ où chacun de leurs termes n'a pas encore été examiné dans d'autres branches de l'arbre constitué par les termsets. Les algorithmes CLOSET [Pei *et al.*, 2000], CLOSET+ [Wang *et al.*, 2003], FP-CLOSE [Grahne and Zhu, 2003] et GC-GROWTH [Haiquan *et al.*, 2005] sont les plus efficaces concernant l'extraction des termsets fermés fréquents par un parcours en profondeur d'abord.
3. **Parcours hybride de l'espace de recherche :**
   Initialement implémenté dans l'algorithme CHARM [Zaki and Hsiao, 2002], l'idée de ce type d'algorithmes de génération des motifs fermés fréquents est d'explorer l'espace de recherche en profondeur d'abord mais sans diviser le contexte d'extraction en des sous-contextes. Ces algorithmes génèrent un ensemble de candidats comme dans le cas du parcours de l'espace par niveau. Cependant, étant donné que cet ensemble est toujours réduit a un seul élément, ces algorithmes vérifient si cet élément candidat est un termset fermé fréquent ou non. Deux améliorations de l'algorithme CHARM ont été proposées dans la littérature à savoir les algorithmes LCM [Uno *et al.*, 2004] et DCI-CLOSED [Lucchese *et al.*, 2006].

## 1.4 Extraction de règles d'association entre termes

Une règle d'association entre termes $R$ est une implication de la forme $R : T_1 \Rightarrow T_2$, où $T_1$ et $T_2$ sont deux sous-ensembles distincts de $\mathcal{T}$, et $T_1 \cap T_2 = \emptyset$. Les termsets $T_1$ et $T_2$ sont, respectivement, appelés la *prémisse* et la *conclusion* de $R$. La règle $R$ est ainsi générée à partir du termset fermé fréquent $T$ formé de $T_1 \cup T_2$.

Le *support* de la règle $R : T_1 \Rightarrow T_2$ est définie comme suit :

$$Supp(R) = Supp(T) = Supp(T_1 \cup T_2) = |\Psi(T_1 \cup T_2)| \tag{1.4}$$

Alors que sa *confiance* est calculée comme suit :

$$Conf(R) = \frac{Supp(T)}{Supp(T_1)} = \frac{Supp(T_1 \cup T_2)}{Supp(T_1)} = \frac{|\Psi(T_1 \cup T_2)|}{|\Psi(T_1)|} \tag{1.5}$$

Une règle d'association $R$ est dite *valide* si sa valeur de confiance, *i.e.*, $Conf(R)$, est supérieure ou égale à un seuil prédéfini noté par $minconf$[10]. Ce seuil de confiance minimal est utilisé pour exclure les règles dites *non valides*. Par ailleurs, le seuil de support minimal *minsupp* est utilisé pour écarter les règles d'association dérivées à partir du termset $T$ et qui ne sont pas suffisamment fréquentes, *i.e.*, les règles ayant un support $Supp(T) < minsupp$.

10. Dans la suite du chapitre, $T_1 \overset{c}{\Rightarrow} T_2$ indique que la règle $T_1 \Rightarrow T_2$ a une confiance égale à $c$.

**Exemple 9** *En considérant le contexte d'extraction de la* TABLE *1.2 (cf. page 53), la règle d'association $R : W \Rightarrow CD$ peut être dérivée. Dans ce cas de figure, $Supp(R) = Supp(CDW) = 3$, alors que $Conf(R) = \frac{Supp(\text{CDW})}{Supp(\text{W})} = \frac{3}{5}$. Si nous fixons les seuils de minsupp et de minconf, respectivement, à 3 et 0.5, la règle $R$ est valide puisque $Supp(R) = 3 \geq 3$ et $Conf(R) = \frac{3}{5} \geq 0.5$.*

Dans la littérature, deux types de règles d'association sont définis à savoir : *les règles exactes* (avec une confiance égale à 1) et *les règles approximatives* (avec une confiance strictement inférieure à 1) [Zaki, 2004]. Ces deux types d'association s'identifient par deux propriétés différentes comme suit [Zaki, 2004] :

**Propriété 2** *Une règle d'association approximative est de la forme $T_1 \Rightarrow \{T_2 - T_1\}$ et représente une implication entre deux termsets fréquents $T_1$ et $T_2$ tel que $T_1 \subseteq T_2$ et la fermeture de $T_1$ est un sous-ensemble strict de la fermeture de $T_2$, i.e., $\Omega(T_1) \subset \Omega(T_2)$.*

**Exemple 10** *La règle $W \Rightarrow \{CDW\} - W$ est approximative puisque $\Omega(W) = \{CW\} \subset \Omega(CDW) = \{CDW\}$. Sa confiance est égale à $\frac{3}{5}$.*

**Propriété 3** *Une règle d'association exacte est de la forme $T_1 \Rightarrow \{T_2 - T_1\}$ et représente une implication entre $T_1$ et $T_2$, tel que $T_1 \subseteq T_2$ et $T_1$, $T_2$ ont des fermetures identiques, i.e., $\Omega(T_1) = \Omega(T_2)$.*

**Exemple 11** *La règle $\{DW\} \Rightarrow \{CDW\} - \{DW\}$ est dite exacte avec une confiance égale à 1 puisque $\Omega(DW) = \Omega(CDW) = \{CDW\}$.*

Intuitivement, étant donné une collection de documents, le problème de génération des règles d'association entre termes consiste à dériver toutes les associations possibles entre les termes de la collection, en considérant des seuils prédéfinis de support et de confiance (*minsupp* et *minconf*). Ce processus de génération de règles d'association peut être mené en deux étapes, à savoir :

1. **Extraire tous les termsets fréquents qui apparaissent dans la collection avec un support supérieur ou égal à** *minsupp* **:** La dernière décennie a témoigné d'une abondance de travaux de recherche qui proposent des algorithmes efficaces et optimisés pour la découverte des motifs fréquents. Afin d'éviter de dériver un nombre élevé de termsets fréquents, plusieurs approches ont été proposées autour du paradigme des termsets fermés [Pasquier *et al.*, 1999]. Il importe de noter que l'ensemble $\mathcal{TFF}$ des termsets fermés fréquents est souvent plus réduit que l'ensemble de tous les termsets fermés. Une étude comparative détaillée des algorithmes d'extraction des motifs fermés fréquents est donnée dans [Ben Yahia *et al.*, 2006].
2. **Générer les règles d'association valides entre termes à partir des termsets fréquents :** Il s'agit de dériver les associations dont la confiance est supérieure ou égale au seuil prédéfini *minconf*. Ces règles peuvent être générées d'une manière directe et simple sans avoir recours à des parcours supplémentaires du contexte d'extraction, moyennant l'algorithme de référence GEN-RULES [Agrawal and Skirant, 1994]. Toutefois, le nombre de règles d'association générées peut être très élevé selon la taille du corpus ou de la collection de documents et peut atteindre plusieurs millions d'associations [Zaki, 2004], alors qu'une large partie d'entre elles pourrait être redondante [Ashrafi *et al.*, 2007, Ben Yahia *et al.*, 2009, Balcázar, 2010].

Cependant, la problématique d'extraction des règles d'association à partir de textes fait face à de réels défis qui sont étroitement liés à la taille des collections de documents. Ces collections sont souvent volumineuses, voire denses. Le premier défi consiste à trouver une manière efficace pour n'extraire que les associations pertinentes à partir d'un nombre très élevé de possibilités, soit $(2^{|T|} - 2)$ pour un termset fréquent $T$.

Il en résulte que plusieurs travaux de recherche ont abordé le problème de la redondance des règles d'association. Certaines approches ont proposé d'utiliser d'autres mesures de qualité, en plus des deux métriques de support et de confiance, lors du processus de fouille telles que la mesure du *lift* et celle de *conviction* [Guillet and Hamilton, 2007]. D'autres approches introduisent des contraintes définies par l'utilisateur, soit a priori, soit en aval à l'extraction des règles d'association [Liu *et al.*, 2009]. Des techniques plus avancées, fondées sur la fermeture de la connexion Galois, dérivent seulement un ensemble compact er réduit de règles d'association. Ces techniques œuvrent pour l'extraction d'un noyau irréductible de l'ensemble globale des associations valides, appelé *base générique* ; à partir de cette dernière toutes les règles redondantes peuvent être dérivées moyennant un système axiomatique dédié [Bastide *et al.*, 2000a, Kryszkiewicz, 2002, Zaki, 2004, Ben Yahia *et al.*, 2009, Balcázar, 2010]. Nous aborderons plus en détails les bases génériques de règles d'association dans le deuxième chapitre du présent document.

Outre les termsets fermés fréquents et les règles d'association, nous nous intéressons dans le cadre de nos recherches à un autre type de motifs fréquents à savoir les séquences fréquentes de termes que nous décrivons dans la section qui suit.

## 1.5 Fouille de séquences fréquentes et ECT

Initialement introduit dans [Srikant and Agrawal, 1996], le problème de l'extraction de motifs séquentiels est vu comme une extension de la problématique de l'extraction d'itemsets et de règles d'association [Agrawal and Skirant, 1994]. Cette extension s'est imposée en raison d'un élément clé qui est propre à l'extraction de motifs séquentiels, à savoir *la temporalité* [Masseglia *et al.*, 2004]. En effet, cette notion de temporalité permet à la fois de distinguer à l'intérieur des transactions *un ordre d'apparition* mais aussi de regrouper certains attributs. Formellement, les règles d'association permettent l'extraction de règles *intra-transaction* alors que la recherche de motifs séquentiels permet l'extraction de règles *inter-transactions*.

La notion de motifs séquentiels reste intuitivement applicable à tout domaine dans lequel il existe une relation d'ordre entre les éléments, tel que le domaine de l'ECT [Berry, 2008] où l'ordre d'apparition des termes dans une phrase peut être pris en compte.

Dans la littérature, plusieurs approches d'ECT existent et ont été utilisées dans des domaines connexes tels que la RI, l'extraction automatique de résumés, la classification des documents, etc [Berry, 2008]. Basées sur la découverte des termsets fréquents, ces approches considèrent les corpus de textes comme des *sacs de mots* où aucun ordre n'est pris en compte lors du processus de fouille [Feldman *et al.*, 1996, Rungsawang *et al.*, 1999, Haddad *et al.*, 2000, Latiri *et al.*, 2003c].

Dans le cadre de nos recherches, nous avons abordé la problématique d'extraction de séquences fréquentes dans le domaine de l'ECT en prenant en compte l'ordre d'apparition des mots dans une phrase, et ce à partir de corpus de textes parallèles alignés. Notre objectif est de les déployer dans la traduction automatique statistique [Och and Ney, 2000, Koehn, 2004]. Nous consacrons le chapitre 4 pour la description de cette contribution.

Nous proposons, dans ce qui suit, de décrire le cadre formel de la découverte des séquences fréquentes adapté à l'ECT.

### 1.5.1 Cadre formel de l'extraction des séquences fréquentes à partir de textes

Nous retenons comme cadre formel, les énoncés introduits par Agrawal *et al.* dans [Srikant and Agrawal, 1996] en les adaptant au domaine de l'ECT. Nous considérons le contexte d'extraction textuel $\mathfrak{M} = (\mathcal{C}, \mathcal{T}, \mathcal{I})$ (*cf.* Définition 1, page 52). La granularité textuelle retenue dans le corpus est la phrase. Autrement dit, $\mathcal{C}$ représente un ensemble fini de phrases du corpus.

**Définition 3** *Soit $\mathcal{T} = \{t_1, \ldots, t_m\}$ l'ensemble de tous les termes présents dans un corpus. Une séquence $\mathcal{S} = \langle s_i, \ldots, s_n \rangle$, tel que $s_k \in \mathcal{T}$ est une liste ordonnée de termes dans le sens où l'ordre d'apparition des mots dans la phrase est respecté. Notons qu'on peut avoir une duplication de termes dans la même séquence* [Latiri *et al.*, 2010a].

À titre d'exemple, la séquence ⟨Le français⟩ est une ***sous-séquence*** de ⟨Le président français⟩. Cependant, ⟨président le⟩ n'est pas une sous-séquence de ⟨Le président français⟩ puisque l'ordre des termes est pris en compte dans notre approche. La séquence ⟨Le président français⟩ est une ***super-séquence*** de la séquence ⟨Le français⟩.

**Définition 4** *Soit $S$ une séquence du contexte d'extraction textuel $\mathfrak{M} = (\mathcal{C}, \mathcal{T}, \mathcal{I})$. Le support de $S$ est donné par $Supp(S)_{S \subseteq p} = |p \in \mathcal{C}|$ qui n'est autre que le nombre de phrases dans le corpus qui contiennent la séquence $S$ et qui sont des super-séquences de $S$* [Latiri *et al.*, 2010a].

Compte tenu de la densité des corpus textuels et afin d'éviter de dériver un nombre très élevé de séquences, il s'agit d'extraire un ensemble compact de sous-séquences à partir d'un ensemble de séquences candidates tel que leurs supports soient supérieurs ou égaux à un seuil de support minimal fixé, *i.e., minsupp*.

**Définition 5** *Une séquence $S$ est dite **fréquente** si $Supp(S)$ est supérieur ou égal à un seuil de support minimal minsupp* [Latiri *et al.*, 2010a].

Il importe de souligner que, dans les applications d'ECT, le seuil de support minimal *minsupp* est déterminé expérimentalement suite à l'étude de la densité du corpus et de la répartition statistique de ses termes.

**Définition 6** *La longueur d'une séquence $S$, notée par $len(S)$, indique le nombre de termes contenus dans la séquence. Une séquence de $k$ termes est dite **k-séquence*** [Latiri *et al.*, 2010a].

**Définition 7** *Une séquence fréquente $S$ est dite fermée si il n'existe pas de super-séquences fréquentes du contexte d'extraction textuel $\mathfrak{M} = (\mathcal{C}, \mathcal{T}, \mathcal{I})$ ayant le même support que $S$* [Latiri *et al.*, 2010a].

**Proposition 1** *Étant donnée une séquence de termes $S$, son information de position respective, est un couple de la forme (id_ph, pos_seq) tel que id_ph représente le numéro de la phrase dans le corpus et pos_seq la position de la séquence dans la phrase* [Chang, 2004].

Formellement, l'ensemble des motifs séquentiels fermés fréquents peut être défini comme suit [Latiri *et al.*, 2010a] :

**Définition 8** *Soit $\mathfrak{M} = (\mathcal{C}, \mathcal{T}, \mathcal{I})$ un contexte d'extraction textuel, le seuil de support minimal minsupp et $\mathcal{SF}$ l'ensemble des séquences fréquentes correspondantes. L'ensemble des séquences fermées fréquentes $\mathcal{SFF}$ est défini comme suit :*

$$\mathcal{SFF} = \{S \mid S \in \mathcal{SF} \text{ et } \nexists\ S' \text{ tel que } S \subset S' \text{ et } Supp(S') = Supp(S)\} \tag{1.6}$$

Nous présentons, dans ce qui suit, un aperçu général sur les principales approches proposées au niveau de la communauté d'extraction des connaissances à partir des données pour l'extraction des motifs séquentiels fréquents [Han *et al.*, 2000, Zaki, 2001, Pei *et al.*, 2001, Ayres *et al.*, 2002, Wang and Han, 2004, Masseglia *et al.*, 2009].

### 1.5.2 Synthèse sur les approches existantes pour l'extraction des motifs séquentiels fréquents

Par rapport à la recherche des règles d'association, le fait que nous prenons en compte la temporalité des itemsets lors de l'extraction des séquences fréquentes, *i.e.,* l'ordre d'apparition des mots dans une phrase dans le contexte d'ECT, engendre des combinaisons supplémentaires qu'il convient d'examiner.

En se référant au domaine de data mining, dans le cas de la recherche des itemsets fréquents, la taille de l'espace de recherche correspond à $2^I$ où $I$ représente le nombre d'items différents dans la base transactionnelle [Agrawal and Skirant, 1994]. Par ailleurs, si nous considérons une séquence $\langle s_1, \ldots, s_{n_i} \rangle$ et que $n_i = |s_i|$ représente la cardinalité d'un itemset, alors la taille de l'espace de recherche, *i.e.,* de l'ensemble de toutes les séquences potentielles, est $2^{(n_1+\ldots+n_i)}$.

Une méthode intuitive d'extraction de motifs séquentiels serait de générer tous les motifs possibles et de compter leurs supports dans la base transactionnelle. Toutefois, cette approche ne peut pas être traitée en un temps raisonnable vu son caractère combinatoire. Pour résoudre ce problème de parcours de l'espace de recherche, les premières approches d'extraction de motifs ont adopté une stratégie de parcours de l'espace de recherche par niveau, en se basant sur la propriété clé qui stipule que le support décroît de manière monotone lors de l'extension du motif [Agrawal and Skirant, 1994]. Ainsi, si un motif est considéré non fréquent alors ses super-motifs seront aussi non fréquents.

Cependant, dans le cadre de l'extraction de motifs séquentiels, l'étape de génération des candidats est plus complexe que celle des itemsets. Dans le cas des séquences, il existe deux manières d'étendre une séquence par un item, à savoir l'extension de séquence (*S-Extension*) et l'extension d'itemset (*I-Extension*) [Chang, 2004]. Dans une *S-Extension*, l'item est ajouté à la séquence comme nouvel itemset. Dans le cas de la *I-Extension*, l'item est ajouté au dernier itemset de la séquence à étendre.

**Exemple 12** *Soit* $S = \langle (a)(b) \rangle$ *une séquence ; une* I-Extension *de la séquence avec l'item c donne la séquence* $S^{'}$ *suivante :* $S^{'} = \langle (a)(b,c) \rangle$. *Une* S-Extension *de la séquence avec l'item c donne la séquence* $S^{''}$ *suivante :* $\langle S^{''} = (a)(b)(c) \rangle$.

Les principaux algorithmes cités dans la littérature pour l'extraction des séquences fréquentes peuvent être classés selon quatre familles de méthodes [Masseglia *et al.*, 2004]. Ces dernières diffèrent essentiellement sur la manière de parcourir l'espace de recherche et sur les structures de données utilisées pour indexer la base transactionnelle et faciliter une énumération rapide des motifs séquentiels [Kum *et al.*, 2005].

1. **Les méthodes basées sur un parcours en largeur d'abord :** Suite à la première approche d'extraction des motifs fréquents proposée par Agrawal *et al.* [Srikant and Agrawal, 1996], la méthode GSP (*Generalized Sequential Patterns*) [Srikant and Agrawal, 1996] a été l'une des premières propositions pour résoudre la problématique des motifs séquentiels. Les auteurs ont introduit une approche reprenant les principes de l'APRIORI [Agrawal and Skirant, 1994], conçu pour l'extraction des règles d'association. Les difficultés relatives à la prise en compte de la temporalité ont rapidement conduit à la mise en place d'une méthode

de génération de candidats adaptée à ce contexte. Celle-ci maintient les principes d'une recherche par niveau, puisque les candidats sont générés en fonction de leur longueur et non de leur préfixe. Les méthodes basées sur un parcours en *largeur d'abord* sont itératives et effectuent un parcours de bas en haut du treillis des itemsets en déterminant à chaque niveau l'ensemble des *k-séquences* fréquentes. Dans [Masseglia *et al.*, 1998], l'algorithme PSP (*Prefix Tree for Sequential Pattern*) a été proposé avec le but de mettre en place une structure d'arbre de préfixes, pour gérer les candidats. L'idée clé de l'algorithme PSP est de factoriser les séquences candidates en fonction de leur préfixe. Cette factorisation, inspirée de celle mise en place dans [Srikant and Agrawal, 1996], pousse plus loin l'exploitation des préfixes communs que présentent les candidats.

2. **Les méthodes basées sur un parcours en profondeur d'abord :** L'algorithme SPADE [Zaki, 2001] est le plus cité dans cette famille de méthodes et il permet une extraction plus efficace des motifs séquentiels. Son originalité est de considérer une représentation verticale du contexte d'extraction, qui est une extension de l'approche de génération des itemsets fréquents introduite dans [Zaki and Hsiao, 2002]. Dans les méthodes verticales, le contexte d'extraction devient un ensemble de $n$-uplets de la forme $\langle itemset : ID(sequence), ID(item)\rangle$. L'ensemble des paires $ID$ d'un itemset donné forme la liste d'identifiants $liste(ID)$ de l'itemset. Pour découvrir les $k$-séquences (séquences contenant $k$ items), l'algorithme SPADE joint les $liste(ID)$ de deux éléments de l'ensemble des ($k$-1)-séquences fréquentes. La longueur de la liste résultante est égale au support de la $k$-séquence générée. La procédure s'arrête quand aucune séquence fréquente ne peut être générée ou qu'aucune séquence ne peut être jointe. Notons que les approches verticales d'exploration des motifs fréquents permettent d'améliorer nettement l'étape de vérification des séquences candidates. L'algorithme SPAM qui est assez proche de SPADE, mais avec une méthode de représentation différente, a été proposé dans [Ayres *et al.*, 2002]. Il gère la présence ou l'absence d'un item dans une séquence par l'intermédiaire de vecteurs de bits. Ainsi, un arbre de représentation est créé pour chaque extension d'item et les résultats sont stockés sous la forme de vecteurs de bits. La vérification des candidats est immédiate, car il suffit de compter les bits positionnés à 1 dans la structure et de les comparer avec le support minimum.

3. **Les méthodes basées sur une projection de la base :** D'autres propositions ont considéré la stratégie d'exploration "*diviser pour régner*". C'est le principe adopté dans l'algorithme FREESPAN [Han *et al.*, 2000] où l'idée générale est d'utiliser des projections récursives du contexte d'extraction en fonction des items fréquents. Le contexte est projeté en plusieurs sous-contextes et ainsi les temps de réponses sont améliorés, car chaque contexte projeté est plus petit. L'algorithme FREESPAN a amorcé d'autres méthodes par projection du contexte d'extraction pour la recherche de motifs séquentiels (paradigme *pattern-growth*), notamment l'algorithme PREFIXSPAN [Pei *et al.*, 2001].

4. **Les méthodes de recherche des motifs séquentiels fermés :** Les méthodes horizontales, verticales ou par projection du contexte permettent d'explorer la totalité des sous-séquences fréquentes satisfaisant la contrainte de *minsupp*. Elles génèrent ainsi un nombre très élevé de sous-séquences fréquentes pour les motifs longs. L'idée, inspirée de la recherche d'itemsets fermés [Pasquier *et al.*, 1999], consiste à extraire un ensemble compact de séquences, à savoir les séquences fermées fréquentes. Le premier algorithme dans cette famille de méthodes est CLOSPAN [Yan *et al.*, 2003] qui dérive des séquences ne contenant aucune super-séquence ayant le même support. Un des inconvénients de l'algorithme CLOSPAN est qu'il conserve l'historique des séquences candidates. Ainsi, il ne s'avère pas

efficace dans le cas de contextes d'extraction contenant un nombre important de séquences fermées. Pour pallier à ce problème, l'algorithme BIDE (*BI-Directional Extension*) a été proposé [Wang and Han, 2004]. Son idée clé est d'étendre les séquences dans les deux directions, *i.e.,* en avant (*extension en avant*) et en arrière (*extension en arrière*). Ainsi, dans [Wang and Han, 2004], les auteurs montrent que pour une séquence $S$, s'il n'existe pas d'extension avant ni d'extension arrière alors $S$ est une séquence fermée. Dans la même famille de méthodes dédiées à l'extraction des séquences fermées, l'algorithme BFSM (*Breadth-First Sequence Mining*) [Chang, 2004] est un algorithme performant qui utilise des listes décrivant *l'information de position* de chaque séquence. Une première étape consiste à parcourir la base des séquences de données pour trouver les 1-séquences fréquentes avec leurs informations de position ; ensuite les deux séquences fréquentes sont générées en faisant la jointure entre les 1-séquences fréquentes trouvées. La génération des séquences de longueur 2 permet la génération des séquences de longueur supérieure, tout en réduisant l'espace de recherche. L'algorithme BFSM déploie ensuite deux techniques d'élagage permettant d'obtenir à la fin un treillis compact des séquences fermées fréquentes.

Nous avons pu constater, lors de l'étude de l'état de l'art relatif à l'extraction des motifs séquentiels, que depuis la définition de la problématique dans [Srikant and Agrawal, 1996], de nombreux travaux de recherche ont été menés [Dong and Pei, 2007]. Initialement, les premiers travaux ont consisté à améliorer les performances des algorithmes d'extraction des séquences. Dans ce cadre, de nouvelles structures de données ou de représentations des données sources ont été mises au point. D'autres algorithmes ont proposé de nouvelles techniques de stockage afin de pouvoir gérer la base en mémoire vive [Masseglia *et al.*, 2009]. Des études comparatives détaillées des approches d'extraction des motifs séquentiels sont proposées dans [Masseglia *et al.*, 2004, Kum *et al.*, 2005].

## 1.6 Discussion et conclusion

Nous avons mis l'emphase dans ce chapitre sur l'extraction de trois types de motifs fréquents à partir de textes, à savoir : les termsets fermés fréquents, les règles d'association et les séquences fréquentes. Ces connaissances sont l'aboutissement d'un processus de fouille à partir d'un contexte d'extraction textuel. Cependant, l'extraction de ces connaissances induit une large panoplie d'algorithmes dont les stratégies de fouille diffèrent selon le type de parcours de l'espace de recherche, le type de structures de données adoptées ainsi que la nature des données explorées. En fonction de la granularité de l'unité textuelle choisie, *i.e.,* document ou phrase, la recherche de motifs séquentiels à partir de textes met en évidence des associations *inter-unités textuelles*, contrairement à celle des termsets fréquents et des règles d'association qui extrait des combinaisons *intra-unités textuelles*.

En effet, la dérivation des règles d'association entre termes se réalise à partir de l'ensemble des termsets fréquents extraits d'un contexte d'extraction textuel $\mathfrak{M}$. Cependant, l'approche à la APRIORI a plusieurs limitations, à savoir : (*i*) Les algorithmes d'extraction des termsets fréquents ne sont pas efficaces sur des données fortement corrélées et/ou pour des seuils de support très faibles intéressants pour l'utilisateur tels que le cas des applications liées au traitement automatique de la langue ; (*ii*) Les règles d'association obtenues à partir de ces motifs fréquents sont très nombreuses. La découverte de règles réellement intéressantes est donc d'autant plus délicate que de nombreuses règles s'avèrent inintéressantes ou redondantes ; (*iii*) Les métriques statistiques d'élagage, tels que le *support* et la *confiance* ne permettent pas de s'affranchir de la redondance liée au domaine d'application.

À l'échelle d'un domaine tel que la RI ou la Traduction Automatique Statistique (TAS), nombreuses règles générées sont redondantes et viennent parasiter la phase d'exploration des résultats, rendant d'autant plus difficile la découverte de règles intéressantes. Cette redondance peut être abordée par les approches introduisant les représentations condensées, appelées *bases génériques* et basées sur les *termsets fermés fréquents* [Pasquier *et al.*, 2005] et les *générateurs minimaux* [Bastide *et al.*, 2000a]. Dans le cadre de nos travaux liés à l'ECT, nous nous intéressons à ce type de représentations condensées et nous proposerons dans le chapitre 2 une nouvelle définition d'une base générique minimale de règles d'association entre termes, que nous déploierons dans deux applications différentes liées à la RI (*cf.* chapitre 3) et dans la traduction automatique statistique (*cf.* chapitre 4).

Par ailleurs, les algorithmes d'extraction de règles d'association connaissent de grandes difficultés d'adaptation aux problèmes d'extraction de motifs séquentiels. En effet, si le problème de la recherche de règles d'association est proche de celui des motifs séquentiels, Srikant et Agrawal [Srikant and Agrawal, 1996] ont montré que l'adaptation est possible au détriment des temps de réponse qui demeurent inacceptables. Depuis la définition du problème dans [Srikant and Agrawal, 1996], de nombreuses approches destinées à résoudre la problématique de l'extraction de motifs séquentiels ont émergé [Masseglia *et al.*, 2009].

Toutefois, même si les différentes approches d'extraction de séquences citées dans la littérature ont permis d'améliorer efficacement les performances et les temps de réponse des algorithmes, les chercheurs de la communauté d'ECT font généralement face au même obstacle. Tant pour les règles d'association entre termes que pour les séquences de termes, il n'est pas possible de décider quel algorithme est meilleur que les autres, étant donné que leurs performances sont étroitement liées à la taille et à la nature du corpus de texte manipulé. La densité ou l'éparsité du contexte d'extraction textuel constitue un des paramètres essentiels dans l'analyse des performances d'un algorithme de fouille de motifs fréquents. Ainsi, les algorithmes du type GSP [Srikant and Agrawal, 1996] ou PSP [Masseglia *et al.*, 1998] seront très efficaces dans le cas de grandes collections de documents avec des séquences moyennement longues. Par contre, les algorithmes comme SPADE [Zaki, 2001], SPAM [Ayres *et al.*, 2002] ou PREFIXSPAN [Pei *et al.*, 2001] seront eux très efficaces dans le cas où un très grand nombre de candidats de même taille sont générés [Masseglia *et al.*, 2004, Masseglia *et al.*, 2009].

Dans le cadre de nos travaux, nous nous intéressons particulièrement à la famille des approches dédiées pour l'extraction *des motifs séquentiels fermés fréquents* [Yan *et al.*, 2003, Wang and Han, 2004, Chang, 2004]. En effet, l'extraction de motifs séquentiels devient problématique selon la longueur des motifs séquentiels extraits. Étant donné que l'algorithme BFSM tient compte de l'ordre d'apparition des items dans une séquence et introduit la notion de l'information de position, nous l'avons utilisé comme algorithme de référence dans notre approche d'extraction des séquences fermées de termes à partir de corpus de textes. Cette approche sera présentée dans le chapitre 4.

Ainsi, en considérant comme cadre fédérateur l'ECT, nous orientons nos contributions, qui seront décrites dans les chapitres qui suivent, vers un volet algorithmique à base de l'AFC et des fondements du treillis de Galois pour la définition d'une nouvelle base générique de règles d'association entre termes, et un deuxième volet applicatif abordant des applications qui se croisent avec l'ECT, telles que la RI et la TAS. Dans le chapitre qui suit, nous abordons le problème de la redondance des règles d'association entre termes et la représentation concise des associations via les bases génériques. Nous proposons un compromis entre l'efficacité d'extraction des règles d'association entre termes et la pertinence lors de leurs usages en aval.

# Chapitre 2

# Définition d'une base générique de règles d'association entre termes

**Sommaire**



## 2.1 Objectifs du chapitre

Ce chapitre s'adresse à la problématique de la redondance des règles d'association. Dans la pratique, le nombre de règles d'association, pouvant être extraites à partir de corpus de textes, s'avère très élevé du fait de la présence de règles redondantes. Pour réduire le nombre de règles extraites, certains travaux ont puisé dans les fondements mathématiques de l'Analyse Formelle de Concepts (AFC) pour proposer des approches de sélection d'un noyau compact de règles d'association non-redondantes, appelé *base générique* [Pasquier *et al.*, 2005, Balcázar, 2010].

Dans le présent chapitre, après avoir donné un aperçu sur les bases génériques de règles d'association les plus citées dans la littérature, nous proposons la formalisation et l'extraction d'une nouvelle base générique minimale de règles d'association non-redondantes entre termes, appelée $\mathcal{MGB}$ [Latiri *et al.*, 2012b]. La définition de cette base générique est basée sur les fondements mathématiques de l'AFC. Elle contient ainsi un nombre plus réduit de règles valides suite à l'élagage de celles qui sont redondantes. La caractéristique clé de ces règles est qu'elles ont des prémisses minimales et des conclusions maximales. Ce chapitre se termine par une évaluation

empirique de la base $\mathcal{MGB}$ ainsi qu'une étude comparative avec les bases génériques pionnières de la littérature [Latiri *et al.*, 2012b].

## 2.2 Aperçu sur les bases génériques de règles d'association

Dans ce qui suit, nous nous focalisons sur les travaux issus de l'AFC [Wille, 1989, Ganter and Wille, 1999] pour la dérivation des bases génériques [Bastide *et al.*, 2000a, Kryszkiewicz, 2002, Ben Yahia *et al.*, 2009, Balcázar, 2010]. Certaines bases génériques proposées dans la littérature englobent des règles d'association qui se présentent sous forme d'implications entre les générateurs minimaux et les termsets fermés, tout en garantissant l'obtention de règles d'association avec une prémisse minimale et une conclusion maximale. Ces règles véhiculent le maximum d'information, et sont donc considérées comme les plus informatives [Bastide *et al.*, 2000a]. En effet, une base générique doit remplir les conditions suivantes [Kryszkiewicz, 2002] :

- *Informativité* : la base générique de règles d'association doit permettre de retrouver avec **exactitude** le support et la confiance des règles dérivées.
- *Dérivabilité* : la base générique doit être dotée d'un mécanisme d'inférence (*i.e.,* un système axiomatique), permettant la dérivation des règles redondantes. Ce système doit être **correct** (*i.e.,* le système ne permet de dériver que les règles d'association valides) et **complet** (*i.e.,* l'ensemble de toutes les règles valides peut être retrouvé).
- *Compacité* : l'ensemble de règles d'association dérivé doit être réduit et minimal tout en permettant la dérivation de toutes les règles valides, *i.e.,* les règles redondantes.

Dans la littérature, deux principales classes d'approches ont été explorées pour l'extraction de bases génériques. La première contient celles qui proposent des bases *avec perte d'information*, *i.e.,* elles ne remplissent pas la condition de dérivabilité ou celle de l'informativité, tandis que la deuxième classe couvre les approches qui utilisent des bases génériques *sans perte d'information*. Une discussion intéressante sur les principales bases génériques de règles d'association est proposée dans [Ben Yahia *et al.*, 2009].

Dans ce qui suit, nous allons présenter la principale base générique représentante de chacune des classes susmentionnées. Notons que les définitions associées sont adaptées à notre contexte d'ECT, et ce à travers l'utilisation de *termsets* au lieu d'*itemsets*.

### 2.2.1 Extraction de bases génériques sans perte d'information

Dans la littérature, plusieurs approches se sont intéressées à la réduction de l'ensemble de règles d'association extraites sans aucune perte d'information [Ben Yahia *et al.*, 2009, Balcázar, 2010]. Toutefois, comme cela a été mentionné dans [Kryszkiewicz, 2002, Ben Yahia *et al.*, 2009], la principale base générique représentante de cette classe est celle de Bastide *et al.* [Bastide *et al.*, 2000a]. Dans leurs travaux, les auteurs ont défini une règle d'association *redondante* comme suit :

**Définition 9** *Soit $\mathcal{VAR}$ l'ensemble de toutes les règles d'association valides, découvertes à partir d'un contexte textuel $\mathfrak{M} = (\mathcal{C}, \mathcal{T}, \mathcal{I})$ pour un seuil de support minimal minsupp et un seuil de confiance minimal minconf. Une règle d'association $R_1 : T_1 \Rightarrow T_2 \in \mathcal{VAR}$ est dite redondante par rapport à (ou dérivable à partir) d'une règle $R_2 : T'_1 \Rightarrow T'_2 \in \mathcal{VAR}$, si et seulement si :*

1. *$Supp(R_1) = Supp(R_2)$ et $Conf(R_1) = Conf(R_2)$, et,*
2. *$T'_1 \subseteq T_1$ et $T_2 \subset T'_2$.*

Nos exemples illustratifs seront basés sur le contexte d'extraction donné dans la TABLE 1.2 (*cf.* page 53).

**Exemple 13** *Considérons les deux règles $R_1$ : $W \Rightarrow A$ et $R_2$ : $W \Rightarrow AC$. Étant donné, $\Omega(AW) = \Omega(ACW) = \{ACW\}$ et en se basant sur la Propriété 1 (cf. page 54), nous avons $Supp(AW) = Supp(ACW)$. Ainsi, $Supp(R_1) = Supp(R_2)$. De plus, puisque les deux règles ont la même prémisse, elles ont donc la même confiance, soit : $Conf(R_1) = Conf(R_2)$. Par conséquent, $R_1$ est dite redondante par rapport à la règle $R_2$ puisqu'elles ont les mêmes valeurs de support et de confiance ainsi que la même prémisse, tandis que la conclusion de $R_1$, à savoir $A$, est un sous-ensemble propre de celle de $R_2$, soit $\{AC\}$.*

En se référant à la Définition 9, étant donnée une règle d'association $R_1 : T_1 \Rightarrow T_2$, s'il n'existe pas une autre règle de la forme $R_2 : T_1' \Rightarrow T_2'$, tel que, $Supp(R_1) = Supp(R_2)$, $Conf(R_1) = Conf(R_2)$, $T_1' \subseteq T_1$, et $T_2 \subset T_2'$, alors $R_1 : T_1 \Rightarrow T_2$ est dite *minimale non-redondante* [Bastide *et al.*, 2000a].

Notons que cette définition garantit que les règles d'association non-redondantes découvertes ont des *prémisses minimales* et des *conclusions maximales*. Les auteurs distinguent deux types de bases à savoir : $(i)$ la base générique pour les règles *exactes*, notée par $\mathcal{GBE}$ ; et, $(ii)$ la base générique pour les règles *approximatives*, notée par $\mathcal{GBA}$. Les bases génériques $\mathcal{GBE}$ et $\mathcal{GBA}$ sont formalisées comme suit [Bastide *et al.*, 2000a] :

**Définition 10** *Soit $\mathcal{TFF}$ l'ensemble des termsets fermés fréquents extrait à partir d'un contexte d'extraction textuel et, pour chaque termset fermé fréquent $T$, $\mathcal{G}_T$ désigne l'ensemble de ses générateurs minimaux. La base générique pour les règles exactes $\mathcal{GBE}$ est définie comme suit :*

$$\mathcal{GBE} = \{R : g \Rightarrow (T - g) |\ T \in \mathcal{TFF} \ \wedge \ g \in \mathcal{G}_T \ \wedge \ g \neq T\}. \tag{2.1}$$

*La base générique pour les règles approximatives $\mathcal{GBA}$ est définie comme suit :*

$$\mathcal{GBA} = \{R : g \Rightarrow (T - g) |\ T, T_1 \in \mathcal{TFF} \ \wedge \ g \in \mathcal{G}_{T_1} \ \wedge \ T_1 \subset T \ \wedge \ Conf(R) \geq minconf\}. \tag{2.2}$$

Pour remédier aux faiblesses liées à la grande taille et à la faible compacité de la base générique $\mathcal{GBA}$, notamment pour les contextes d'extraction épars [Bastide *et al.*, 2000a], Bastide *et al.* ont proposé un *réduction transitive* de la base générique de règles d'association approximatives, notée $\mathcal{TGBA}$, comme suit :

**Définition 11** *La base $\mathcal{TGBA}$ est formalisée par :*

$$\mathcal{TGBA} = \{R : g \Rightarrow (T - g) |\ T, T_1 \in \mathcal{TFF} \wedge T \in Couv^s(T_1) \wedge g \in \mathcal{G}_{T_1} \wedge Conf(R) \geq minconf\}. \tag{2.3}$$

**Exemple 14** *Supposons un seuil de confiance minimale $minconf = 0.5$. Puisque $\{ACW\} \in Couv^s(CW)$ et $W$ est un générateur minimal de $\{CW\}$, la règle $W \Rightarrow AC$ appartient à $\mathcal{TGBA}$ et a une valeur de confiance égale à $\frac{4}{5} \geq minconf$. Notons par ailleurs que la règle $W \Rightarrow ACT$ appartenant à $\mathcal{GBA}$ n'est pas incluse dans $\mathcal{TGBA}$ puisque $\{ACTW\} \notin Couv^s(CW)$.*

Dans [Kryszkiewicz, 2002], l'auteur prouve que le couple $(\mathcal{GBE}, \mathcal{GBA})$ forme une base générique valide et informative de règles d'association, *i.e.*, leurs support et confiance respectifs sont inférés avec exactitude. Toutefois, la base $(\mathcal{GBE}, \mathcal{GBA})$ souffre de la génération d'un nombre important de règles surtout pour les contextes d'extraction denses. Ce constat est renforcé par le fait que pour les contextes épars, l'extraction du couple $(\mathcal{GBE}, \mathcal{GBA})$ n'apporte aucun gain en terme de compacité.

**Exemple 15** *Nous nous référons dans cet exemple au treillis de l'Iceberg augmenté $\mathcal{TA}$, illustré dans la* FIGURE *1.1 (*cf. *page 56). Considérons le termset fermé fréquent $\{ACTW\}$ et son générateur minimal $\{AT\}$. La règle d'association induite à partir de ces motifs est : $AT \Rightarrow CW$ appartient à $\mathcal{GBE}$. D'un autre côté, admettons un seuil de minconf égal à 0.5 et considérons les deux* TFF*s $\{CW\}$ et $\{ACTW\}$. Puisque $\{CW\} \subset \{ACTW\}$ et $W$ est un générateur minimal de $\{CW\}$, la règle d'association $W \Rightarrow ACT$ appartient donc à $\mathcal{GBA}$ avec une valeur de confiance égale à $\frac{3}{5} \geq$ minconf.*

### 2.2.2 Extraction de bases génériques avec perte d'information

D'autres bases génériques de règles d'association avec perte d'information ont été proposées dans le littérature [Ben Yahia *et al.*, 2009]. La base de référence a été introduite par Zaki dans [Zaki, 2004]. Il a défini une base générique appelée *base de règles d'association non-redondantes*, notée $\mathcal{NRR}$. Son approche est basée sur un système axiomatique, tenant compte du support et de la confiance, pour la génération de tout l'ensemble de règles d'association valides et ce à partir d'une base minimale. L'auteur définit la redondance d'une règle d'association comme suit :

**Définition 12** *Soit $\mathcal{VAR} = \{R_1,\ldots,R_n\}$ l'ensemble de toutes les règles valides dérivées à partir d'un contexte d'extraction $\mathfrak{M} = (\mathcal{C}, \mathcal{T}, \mathcal{I})$. $R_1 : T_1 \Rightarrow T_2 \in \mathcal{VAR}$ couvre la règle $R_2 : T_1' \Rightarrow T_2'$ (ou d'une manière équivalente, $R_2$ est redondante par rapport à $R_1$), notée par $R_1 \preceq R_2$, si et seulement si les conditions suivantes sont vérifiées* [Zaki, 2004] :

1. $T_1 \subset T_1'$ et $T_2 \subset T_2'$ ;
2. $Supp(R_1) = Supp(R_2)$ et $Conf(R_1) = Conf(R_2)$.

De ce fait, une règle d'association $R_2$ est considérée comme redondante si et seulement si il existe une règle d'association $R_1$ telle que $R_1 \preceq R_2$ ; autrement elle est dite non-redondante.

**Exemple 16** *Considérons le treillis de l'Iceberg augmenté illustré par la* FIGURE *1.1 (cf. page 56). Selon la Définition 12, la règle $R_1 : W \Rightarrow T$ est couverte par la règle $R_2 : CW \Rightarrow AT$ puisque, d'une part, $W \subseteq \{CW\}$ et $T \subseteq \{AT\}$ et, d'autre part, $Supp(R_1) = Supp(R_2) = 3$ et $Conf(R_1) = Conf(R_2) = \frac{3}{5}$. La règle d'association $R_2$ est ainsi redondante par rapport à $R_1$.*

La notion de non-redondance considérée par Zaki [Zaki, 2004] est basée sur le système d'inférence composé de l'axiome de transitivité de Luxenburger [Luxenburger, 1991] et celui de l'augmentation de Armstrong [Armstrong, 1974]. Ainsi, toutes les règles générées ont une prémisse minimale ainsi qu'une conclusion minimale. Toutefois, comme nous allons le montrer plus loin dans ce chapitre, cette forme minimale de règles d'association n'est pas toujours adaptée pour représenter toutes les connaissances implicites cachées dans une collection de documents.

En se référant à la Définition 12, Zaki a introduit la base générique $\mathcal{NRR}$ comme suit :

**Définition 13** *Soit $\mathcal{VAR} = \{R_1,\ldots,R_n\}$ l'ensemble des règles d'association valides dérivées à partir du contexte d'extraction $\mathfrak{M} = (\mathcal{C}, \mathcal{T}, \mathcal{I})$,*

$$\mathcal{NRR} = \{R_i \in \mathcal{VAR} \mid \nexists\ R_j \in \mathcal{VAR} :\ R_j \preceq R_i\ \wedge\ i \neq j\}. \tag{2.4}$$

**Exemple 17** *Pour minsupp = 3 et minconf = 0.5, la règle $W \Rightarrow T \in \mathcal{NRR}$ puisque cette règle est valide et il n'existe pas une autre règle qui la couvre.*

Néanmoins, dans [Ben Yahia *et al.*, 2009], les auteurs ont souligné que la base $\mathcal{NRR}$ ne couvre pas tout l'ensemble de règles d'association valides. Certaines règles valides n'appartiennent pas à la base $\mathcal{NRR}$ et ne sont pas dérivables par le système axiomatique proposé dans [Zaki, 2004]. Par ailleurs, selon la Définition 12, si une règle d'association $R_2$ est considérée comme redondante par rapport à $R_1$, alors elle devrait hériter des mêmes valeurs de support et de confiance de $R_1$. Néanmoins, une règle d'association $R_3$, déduite en appliquant l'axiome de transitivité de Luxenburger sur $R_1$ et $R_2$ de la base $\mathcal{NRR}$, peut avoir une valeur de confiance différente de celles des deux règles $R_1$ et $R_2$.

Il convient de rappeler que l'application de règles d'association dans le contexte d'applications connexes à l'ECT telles que la RI ou la traduction automatique est loin d'être une tâche facile, principalement en raison du nombre très élevé de règles d'association potentiellement intéressantes, découvertes à partir d'une collection de documents. Dans l'objectif de proposer un compromis entre l'informativité et la compacité d'une base générique, nous introduisons dans la section qui suit une nouvelle approche pour l'exploration d'une *base générique minimale de règles d'association non-redondantes entre termes* à partir d'un contexte d'extraction textuel [Latiri *et al.*, 2012b].

## 2.3 $\mathcal{MGB}$ : Nouvelle base générique minimale de règles d'association entre termes

La revue de la littérature des bases génériques de règles d'association nous a permis de dégager leurs limites dans le contexte de l'ECT, qui sont principalement liées au manque de compacité et d'informativité dans le cas de contextes d'extraction denses tel qu'un corpus de textes.

Ainsi, en se basant sur la Définition 10 (*cf.* page 69) du couple de bases générique ($\mathcal{GBE}$,$\mathcal{GBA}$), nous considérons que, étant donné un treillis de l'Iceberg $\mathcal{TA}$, représentant la relation de précédence au sein de l'ensemble de termsets fermés fréquents $\mathcal{TFF}$, une base générique de règles d'association peut être dérivée d'une manière directe. Nous supposons que dans une telle structure ordonnée, chaque termset fermé fréquent est augmenté avec la liste de ses générateurs minimaux. De ce fait, les règles approximatives (RAs) représentent des implications *inter-nœuds*, d'un sous-termset fermé vers un super-termset fermé, tout en commençant à partir d'un nœud donné dans la structure du treillis de l'Iceberg $\mathcal{TA}$. D'autre part, les règles exactes (REs) traduisent des implications *intra-nœuds*, extraites de chaque nœud du treillis.

En partant de cette idée, nous proposons la définition d'une nouvelle *base générique minimale*, appelée $\mathcal{MGB}$, basée sur l'extraction du *treillis de l'Iceberg augmenté* $\mathcal{TA}$ [Latiri *et al.*, 2012b]. Dans ce contexte, les générateurs minimaux sont utilisés dans la génération de la partie prémisse des règles d'association, tandis que les termsets fermés fréquents constituent les conclusions. La relation de précédence structurant le treillis de l'Iceberg augmenté $\mathcal{TA}$ permet de limiter le coût de l'extraction des règles d'association et ce en évitant certaines combinaisons redondantes. Bien que la base générique proposée $\mathcal{MGB}$ est dédiée à être utilisée dans plusieurs applications, il sera montré plus loin qu'elle est plus appropriée pour l'expansion de requêtes en RI et pour l'enrichissement d'ontologie (*cf.* chapitre 3) et aussi pour la traduction automatique statistique (*cf.* chapitre 4). Ceci est possible grâce à sa formalisation qui tient compte des caractéristiques suivantes [Latiri *et al.*, 2012b] :

1. La base générique $\mathcal{MGB}$ a une taille réduite par rapport à l'ensemble de toutes les règles d'association valides et même en la comparant avec d'autres bases génériques proposées dans la littérature. Cette caractéristique est importante car elle résout le problème de

redondance de règles d'association, tout en conservant un noyau de règles valides et pertinentes.

2. Les règles de la base générique $\mathcal{MGB}$ ont une forme intéressante. En effet, chaque règle retenue traduit une prémisse minimale, *i.e.*, contenant un générateur minimal, qui implique une conclusion maximale par rapport au critère de validité de la confiance. La règle est alors basée sur un termset fermé fréquent, qui est le plus grand ensemble de termes dont la présence dans la collection de documents dépend de celle de l'ensemble des termes de la prémisse avec une probabilité supérieure ou égale à $minconf$.
3. La base générique $\mathcal{MGB}$ vérifie les propriétés de compacité et d'informativité. Ainsi, pour un seuil de support minimal *minsupp* et un seuil de confiance minimal *minconf*, il est possible de dériver toutes les règles d'association valides sans perte d'information. Il est intéressant de noter que la compacité de la base générique $\mathcal{MGB}$ garantit l'utilisation d'un espace de stockage minimal et rend plus facile les manipulations futures lors de son utilisation dans des applications utilisant des corpus de textes volumineux.

Les principales caractéristiques de la base générique $\mathcal{MGB}$ seront davantage détaillées dans les sections suivantes.

### 2.3.1 Découverte des règles d'association non-redondantes

Habituellement, une règle d'association $R_2$ est considérée comme redondante par rapport à une règle $R_1$ si l'information véhiculée par $R_1$ induit celle véhiculée par $R_2$.

Nous considérons, dans le cadre de l'ECT, que la définition de la redondance proposée par Zaki [Zaki, 2004] n'est pas appropriée vu que les règles d'association ont des conclusions minimales, ce qui peut limiter leurs possibilités d'utilisation. Cette limite est expliquée par le fait que les connaissances additionnelles induites par les conclusions minimales s'avèrent insuffisantes et incomplètes lors de leur utilisation dans une application réelle.

Par ailleurs, la définition de Bastide *et al.* [Bastide *et al.*, 2000a], souffre d'une permissivité élevée. En effet, le seul critère utilisé pour élaguer une règle redondante par rapport à une autre non-redondante, est que les deux règles ont les mêmes mesures de support et de confiance. De ce fait, l'approche de Bastide *et al.* ne parvient pas à exclure les règles d'association redondante avec des prémisses identiques et des conclusions comparables par rapport à l'inclusion ensembliste.

Dans notre contribution, nous considérons des règles d'association qui maximisent le nombre de termes dans la partie conclusion. Notre principale motivation est d'obtenir des connaissances additionnelles, pouvant être exploitées dans d'autres applications.

Nous définissons la redondance comme suit [Latiri *et al.*, 2012b] :

**Définition 14** *Une règle d'association valide $R_1 : T_1 \Rightarrow T_2$ est redondante par rapport à une règle valide $R_2 : T'_1 \Rightarrow T'_2$ si et seulement une des conditions suivantes est vérifiée :*

1. $\Omega(T'_1 \cup T'_2) = \Omega(T_1 \cup T_2)$ *et* $T'_1 \subset T_1$
2. $T'_1 = T_1$ *et* $T_2 \subset T'_2$.

**Exemple 18** *Considérons le termset fermé fréquent $\{ACTW\}$ ainsi que les deux règles $R_1 : A \Rightarrow CTW$ et $R_2 : AC \Rightarrow TW$. Selon la Définition 14, $R_2$ est dite redondante par rapport à $R_1$ puisque la découverte de $R_1$ implique nécessairement celle de $R_2$. En effet, les deux règles ont la même valeur de support qui est égale à $Supp(ACTW)$ et les valeurs de confiances respectives suivantes : $Conf(R_1) = \frac{Supp(\mathrm{ACTW})}{Supp(\mathrm{A})}$ et $Conf(R_2) = \frac{Supp(\mathrm{ACTW})}{Supp(\mathrm{AC})}$. Nous déduisons que $Conf(R_2) \geq Conf(R_1)$. De ce fait, si $R_1$ est une règle valide, nécessairement $R_2$ l'est aussi.*

Nous définissons dans ce qui suit le concept de *prémisse valide* dans le cadre de la base générique $\mathcal{MGB}$. Nous formalisons en premier lieu l'ensemble des prémisses potentielles comme suit [Latiri *et al.*, 2012b] :

**Définition 15** *Soit $T$ un termset fermé fréquent. L'ensemble des prémisses potentielles relatives aux règles d'association valides générées à partir de $T$ contient ses générateurs minimaux ainsi que ceux de tous les* TFF*s inclus dans $T$. Cet ensemble est défini comme suit :*

$$all\mathcal{G}_T = \{g \subseteq \mathcal{T} | \Omega(g) = T_1 \subseteq T\}. \tag{2.5}$$

**Exemple 19** *Soit le termset fermé fréquent $\{ACW\}$, selon la Définition 15, $all\mathcal{G}_{\{ACW\}} = \{A,C,W\}$.*

Étant donné que dans notre contribution, nous visons à ne retenir que les règles d'association valides avec des prémisses minimales et des conclusions maximales par rapport aux deux critères de sélection *minsupp* et $minconf$, *les prémisses valides* retenues pour chaque termset fermé fréquent sont alors définies comme suit [Latiri *et al.*, 2012b] :

**Définition 16** *En considérant le treillis de l'Iceberg augmenté $\mathcal{TA}$, l'ensemble des prémisses valides dérivées à partir d'un* TFF*, noté $min\mathcal{G}_T$, est défini par :*

$$\begin{aligned} min\mathcal{G}_T = \{g \in all\mathcal{G}_T | (\nexists\ g_1 \in all\mathcal{G}_T :\ g_1 \subset g\ \wedge\ \frac{Supp(T)}{Supp(g_1)} \geq minconf) \\ \wedge\ (\nexists\ s \in Couv^s(T) :\ \frac{Supp(s)}{Supp(g)} \geq minconf)\} \end{aligned} \tag{2.6}$$

Autrement dit, l'ensemble $min\mathcal{G}_T$ contient uniquement les termsets minimaux, au sens de l'inclusion ensembliste, permettant de générer des règles valides à partir du TFF $T$. De plus, les éléments de $min\mathcal{G}_T$ ne peuvent pas être utilisés comme étant des prémisses valides relatives aux règles d'association dérivées à partir des termsets fermés couvrant le TFF $T$. En effet, dans ce dernier cas de figure, les règles d'association générées à partir de $T$ n'auront pas de conclusions maximales candidates, compte tenu des prémisses qui leur sont associées.

**Exemple 20** *Soit le treillis de l'Iceberg de Galois augmenté $\mathcal{TA}$ illustré par la* Figure *1.1 (cf. page 56). En faisant référence à l'exemple précédent, nous avons l'ensemble des prémisses potentielles $all\mathcal{G}_{\{ACW\}} = \{A,C,W\}$. Nous analysons ainsi les éléments de l'ensemble $min\mathcal{G}_{ACW}$ par rapport aux différentes valeurs de minconf, à savoir :*

- *Pour minconf = 1.0 : Seul le générateur $A$ est retenu dans $all\mathcal{G}_{\{ACW\}}$. En effet, $A$ ne peut pas être la prémisse d'une règle d'association ayant une conclusion plus large que $\{CW\}$ i.e., $(\{ACW\} \setminus A)$. Aussi, les autres prémisses potentielles induisent des règles d'association non valides par rapport au seuil de minconf. Seule une règle dérivée à partir de $\{ACW\}$ est retenue, soit $A \overset{1.0}{\Rightarrow} CW$.*
- *Pour minconf = 0.8 : Les deux générateurs minimaux $A$ et $W$ sont retenus dans l'ensemble $all\mathcal{G}_{\{ACW\}}$. De la même manière, ils ne peuvent pas être utilisés dans des règles d'association à conclusion plus large. La troisième prémisse potentielle,* i.e., *$C$, n'est pas retenue puisqu'elle induit une association non valide, i.e., $(Conf\ (C \Rightarrow AW) = 0.66 < 0.8)$. Deux règles d'association dérivées à partir du termset fermée fréquent $\{ACW\}$ sont retenues, à savoir $A \overset{1.0}{\Rightarrow} CW$ et $W \overset{0.8}{\Rightarrow} AC$.*

- *Pour minconf = 0.6 : Seul le générateur C est retenu dans l'ensemble all$\mathcal{G}_{\{ACW\}}$ puisqu'il permet la génération de la règle valide $C \overset{0.66}{\Rightarrow} AW$, sans pour autant être la prémisse d'une règle ayant une conclusion plus large. Notons que, bien que les deux générateurs A et W permettent de dériver des règles d'association valides à partir du TFF $\{ACW\}$, ils ne seront pas retenus dans l'ensemble all$\mathcal{G}_{\{ACW\}}$, puisque ils sont déjà considérés dans des associations à conclusions plus larges et générées à partir du TFF $\{ACTW\}$ au lieu du TFF $\{ACW\}$. Ces règles d'association sont : $A \overset{0.75}{\Rightarrow} CTW$ et $W \overset{0.6}{\Rightarrow} ACT$.*
- *Pour minconf = 0.5 : Dans ce cas, l'ensemble min$\mathcal{G}_{ACW}$ est vide et aucune règle n'est générée à partir du TFF $\{ACW\}$. En effet, toutes les prémisses potentielles sont utilisées dans des règles valides dérivées à partir du TFF $\{ACTW\}$ et ayant des conclusions plus larges, soient : $A \overset{0.75}{\Rightarrow} CTW$, $C \overset{0.5}{\Rightarrow} ATW$ et $W \overset{0.6}{\Rightarrow} ACT$.*

Dans le contexte de la base générique proposée $\mathcal{MGB}$, nous présentons les définitions relatives aux règles d'association *approximative* et *exacte*, respectivement [Latiri *et al.*, 2012b].

**Définition 17** *Soit T un termset fermé fréquent. Une règle d'association non-redondante approximative dérivée à partir de T, est de la forme $R : g \Rightarrow T$ tel que $g \in min\mathcal{G}_T$ et $\Omega(g) \subset T$.*

**Exemple 21** *Considérons les résultats obtenus à partir de l'Exemple 20 (cf. page 73) pour un seuil de minconf = 0.6. La règle d'association $A \Rightarrow CTW$, ayant une confiance égale à 0.75, est une règle non-redondante approximative. De ce fait, $\Omega(A) = \{ACW\} \subset \{ACTW\}$.*

Autrement dit, *une règle d'association approximative* traduit le fait qu'il n'existe pas un termset fermé fréquent $s \in Couv^s(T)$ qui induit une règle approximative sachant que $g$ représente déjà la prémisse d'une règle d'association générée à partir de $s$ (*i.e.*, la règle de la forme $R : g \Rightarrow (s - g)$ n'est pas valide pour tous les TFF $s \in Couv^s(T)$).

**Définition 18** *Soit T un* TFF. *Une règle d'association non-redondante dérivée à partir de T est dite exacte si et seulement si elle est de la forme $R : g \Rightarrow T$ avec $g \in min\mathcal{G}_T$ et $\Omega(g) = T$, i.e., $g \in \mathcal{G}_T$.*

**Exemple 22** *Considérons les résultats obtenus à partir de l'Exemple 20 (cf. page 73) pour un seuil minconf = 0.8. La règle non-redondante $A \Rightarrow CW$, ayant une valeur de confiance égale à 1.0, est dite exacte. Ainsi, $\Omega(A) = ACW$ (puisque A est un générateur minimal de $\{ACW\}$).*

En se basant sur les définitions précédentes, l'ensemble $min\mathcal{G}_T$ est ainsi divisé en deux parties, à savoir : la première contient les générateurs minimaux ayant comme fermeture un sous-ensemble propre de $T$ et, par conséquent, induisent des règles approximatives valides ; par contre la deuxième partie englobe les générateurs minimaux ayant comme fermeture $T$ et génèrent ainsi des règles exactes. Notons qu'une règle exacte est nécessairement valide puisque sa valeur de confiance est égale 1.0.

### 2.3.2 Définition de la base générique minimale $\mathcal{MGB}$

En se basant sur les définitions ci-dessus, nous donnons la formalisation de la base générique minimale de règles d'association entre termes [Latiri *et al.*, 2012b].

**Définition 19** *Soient le contexte d'extraction textuel $\mathfrak{M} = (\mathcal{C}, \mathcal{T}, \mathcal{I})$, $T$ un termset fermé fréquent et ses générateurs minimaux $min\mathcal{G}_T$. La base générique minimale $\mathcal{MGB}$ est définie comme suit :*

$$\mathcal{MGB} = \{R : g \Rightarrow (T - g) \mid T \in \mathcal{TFF} \ \wedge \ g \in min\mathcal{G}_T\} \tag{2.7}$$

Selon l'équation (2.7), la Définition 17 et la Définition 18 (*cf.* page 74), les règles approximatives non-redondantes (RANRs) sont de la forme $g_1 \Rightarrow (T_2 - g_1)$ liées au générateur minimal $g_1$ du TFF $T_1$ et au second TFF $T_2$, tel que $T_1 \subset T_2$ (*i.e.*, $g_1$ "implique" $T_2$, qui est situé plus haut dans le treillis de l'Iceberg de Galois $\mathcal{TA}$, avec une confiance égale à $\frac{Supp(T_2)}{Supp(g_1)}$). Par ailleurs, les règles exactes non-redondantes dérivées (RENRs) sont de la forme $g \Rightarrow (\Omega(g)$ - $g)$, sachant que $g$ n'apparaît dans aucune prémisse des règles approximatives valides, ayant une conclusion plus large que $(\Omega(g)$ - $g)$. À titre d'exemple, pour un seuil $minconf = 0{,}6$, la règle A $\Rightarrow$ CW ne fait pas partie des RENRs de $\mathcal{MGB}$, puisqu'elle est redondante par rapport à la règle approximative : A $\Rightarrow$ CTW (*cf.* Exemple 20, page 73).

Nous décrivons dans ce qui suit l'algorithme Gen-MGB qui permet la construction la base générique $\mathcal{MGB}$ [Latiri *et al.*, 2012b].

### 2.3.3 Description de l'algorithme Gen-MGB

Dans le cadre de notre approche, le treillis de l'Iceberg de Galois $\mathcal{TA}$ supporte la découverte des règles d'association non-redondantes entre termes. En effet, l'utilisation de l'ordre de précédence sur l'ensemble des termsets fermés fréquents $\mathcal{TFF}$, permet de dériver directement les règles non-redondantes exactes et approximatives entre termes, sans aucun calcul additionnel de la mesure de confiance.

Le pseudo-code de l'algorithme Gen-MGB est décrit dans l'Algorithme 1. Il effectue un parcours sur l'ensemble des termsets fermés fréquents $\mathcal{TFF}$ du treillis de l'Iceberg de Galois $\mathcal{TA}$, en commençant par le TFF le plus petit et en le parcourant du haut vers le bas par rapport à l'inclusion ensembliste $\subseteq$.

L'algorithme considère comme entrée le treillis de l'Iceberg augmenté $\mathcal{TA}$ et génère comme résultat les règles approximatives et exactes non-redondantes (*i.e.*, RANRs et RENRs). En se basant sur la Définition 17 et pour un nœud donné dans le treillis de l'Iceberg $\mathcal{TA}$, nous considérons que les RANRs représentent les implications qui englobent d'un côté les générateurs minimaux du sous-termset fermé, associés au nœud traité, et d'un autre côté un super-termset fermé. Par ailleurs, selon la Définition 18 (*cf.* page 74), les RENRs sont des implications dérivées à partir des générateurs minimaux et de leurs fermetures respectives, appartenant au même nœud dans le treillis augmenté $\mathcal{TA}$.

Cependant, la génération des règles d'association non-redondantes entre termes avec l'algorithme Gen-MGB se réalise en deux étapes décrites ci-dessous.

#### Étape 1 : Génération des conclusions candidates

L'objectif de cette étape est de trouver, pour un TFF donné $T_i$, les TFFs qui représentent les conclusions candidates pour les règles d'association ayant $g_i$ comme prémisse, où $g_i \in \mathcal{G}_{T_i}$. Les TFFs ciblés sont ceux qui englobent le TFF $T_i$. Ainsi, une règle d'association $R$ impliquant les générateurs minimaux $g_i$ de $T_i$ et un TFF $T_j \in Couv^s(T_i)$ est valide si et seulement si :

$$Conf(R) = \frac{Supp(T_j)}{Supp(g_i)} = \frac{Supp(T_j)}{Supp(T_i)} \geq minconf \tag{2.8}$$

L'équation (2.8) stipule que $Supp(T_j) \geq Supp(T_i) \times minconf$. Comme premier filtre sur la liste des TFFs couvrant $T_i$, nous utilisons la mesure *seuil-Supp* qui est égale à $(minconf \times Supp(T_i))$ pour ne retenir que les TFFs ayant un support supérieur ou égal au *seuil-Supp*. De ce fait, au lieu de calculer, pour chaque TFF $T_j$, la confiance donnée par l'équation (2.8), l'algorithme Gen-MGB vérifie uniquement si $Supp(T_j) \geq$ *seuil-Supp*. La valeur du *seuil-Supp* est mise à jour en parcourant le treillis de l'Iceberg de Galois du haut vers le bas, en partant par le termset fermé fréquent le plus petit.

La fonction Gen-Conclusion ajoute le TFF $T_j \supset T_i$ (ligne 7) à l'ensemble *limit-fermés* si $T_j$ est un TFF maximal, par rapport à l'inclusion ensembliste, parmi les termsets fermés fréquents retenus après l'application du filtre susmentionné. L'ensemble *limit-fermés* englobe les TFFs, à partir desquels sont dérivées les règles d'association valides et ayant comme prémisse $g_i \in \mathcal{G}_{T_i}$, tout en considérant les conditions spécifiées dans la Définition 16 (*cf.* page 73).

1: **Algorithme** Gen-MGB(**Entrée :** $\mathcal{TA}$ : le treillis de l'Iceberg augmenté, **Sortie :** la base générique minimale $\mathcal{MGB}$)
2: **Pour tout** TFF $T_i \in \mathcal{TA}$ **Faire**
3: *seuil-Supp* $\leftarrow$ *minconf* $\times$ *Supp*($T_i$)
4: limit-fermés $\leftarrow \emptyset$
5: **Pour tout** TFF $T_j \in \mathcal{TFF}$ tel que $T_i \subset T_j$ **Faire**
6: **Si** *Supp*($T_j$) $\geq$ *seuil-Supp* **alors**
7: Gen-Conclusion($T_j$, limit-fermés)
8: **Fin Si**
9: **Fin Pour**
10: **Si** limit-fermés $\neq \emptyset$ **alors**
11: Générer-RANR(limit-fermés, $T_i$, RANRs)
12: **Sinon**
13: Générer-RENR($T_i$, RENRs)
14: **Fin Si**
15: **Fin Pour**
16: **Retourner** ($\mathcal{MGB} = \{\text{RANR}s \cup \text{RENR}s\}$)

**Algorithme 1:** L'algorithme Gen-MGB.

### Étape 2 : Génération des règles non-redondantes

Durant cette étape, pour chaque TFF $T_i$, deux cas de figure sont distingués par rapport au contenu de sa liste respective *limit-fermés*, à savoir :

1. Si la liste *limit-fermés* est non vide (ligne 9), l'algorithme génère les RANRs. Afin de ne retenir que les règles non-redondantes, l'algorithme Gen-MGB gère une liste, dite *list-prohibée* pour chaque TFF contenant les prémisses candidates des associations dérivables à partir de ce TFF. De ce fait, avant de générer de nouvelles règles, l'algorithme Gen-MGB vérifie si un sous-ensemble propre du termset $g_i \in \mathcal{G}_{T_i}$ existe déjà dans la *list-prohibée* de $T_j$ ($T_j \in$ *limit-fermés*). La vérification de ce cas induit que la règle dérivée à partir de $T_j$, et ayant $g_i$ comme prémisse est redondante par rapport à une autre règle non-redondante déjà extraite. Dans le cas contraire, la règle approximative $g_i \Rightarrow (T_j - g_i)$ est générée. Le termset $g_i$ est ainsi ajouté à la *list-prohibée* du TFF $T_j$.
2. Dans le cas où la fonction Get-Conclusion retourne une liste vide, *i.e.*, l'ensemble *limit-fermés* est vide (ligne 11), l'algorithme génère les RENRs associées au TFF $T_i$. Chaque règle exacte dérivée est de la forme $g_i \Rightarrow T_i - g_i$.

**Exemple 23** *Considérons le treillis de l'Iceberg de Galois illustré par la* FIGURE *2.1 pour un seuil de minconf = 0.6 et un seuil de minsupp égal à 3. Toutes les règles d'association approximatives non-redondantes sont indiquées dans la* FIGURE *2.1. Dans ce cas, aucune règle exacte n'est générée puisqu'elles sont toutes redondantes par rapport aux règles approximatives appartenant à $\mathcal{MGB}$. À titre d'exemple, en partant du nœud $\{CDW\}$, la règle $DW \overset{1}{\Rightarrow} C$ n'est pas dérivée étant donné qu'elle considérée comme redondante par rapport à la règle approximative $D \overset{0.75}{\Rightarrow} CW$, selon la Définition 14 (cf. page 72).*

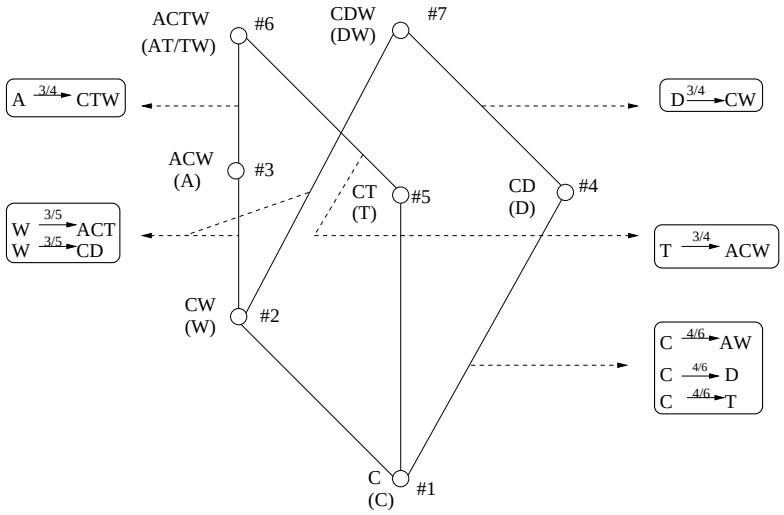


FIGURE 2.1 – Règles d'association non-redondantes relatives au contexte d'extraction textuel $\mathfrak{M}$.

### 2.3.4 Dérivation des règles d'association redondantes

La base générique minimale proposée $\mathcal{MGB}$ permet de dériver toutes les règles valides, y compris celles qui sont redondantes. Ceci est possible grâce aux deux axiomes de dérivation relatifs au système d'inférence défini comme valide et complet dans [Ben Yahia *et al.*, 2009], soit :

1. *L'axiome d'augmentation :* Si $T_1 \Rightarrow T_2 \in \mathcal{MGB}$ et $T_3 \subset T_2$, alors $T_1 \cup T_3 \Rightarrow T_2$ - $T_3$ est une règle d'association valide.
2. *L'axiome de décomposition conditionnelle :* Si $T_1 \Rightarrow T_2 \in \mathcal{MGB}$ et $T_3 \subset T_2$, alors $T_1 \Rightarrow T_3$ est une règle d'association valide.

Cependant, notons bien que la base $\mathcal{MGB}$ est extraite sans perte d'information, elle n'est pas informative. À titre d'exemple, le support et la confiance de la règle d'association AT $\Rightarrow$ CW ne peuvent pas être dérivés à partir de la règle non redondante la couvrant et qui a été retenu dans $\mathcal{MGB}$, *i.e.*, A $\Rightarrow$ CTW. Ceci est dû au fait que les règles de la base $\mathcal{MGB}$ ne permettent pas nécessairement de localiser le termset fermé fréquent correspondant à un termset donné, *i.e.*, sa fermeture. Par contre, ceci est possible pour le termset $\{AT\}$ dont le support n'est pas connu avec exactitude.

## 2.4 Comparaison des bases génériques de règles d'association avec la base $\mathcal{MGB}$

Dans cette section, nous proposons une étude comparative de notre base générique minimale $\mathcal{MGB}$, en considérant l'exemple illustratif de la TABLE 1.3, avec celles proposées respectivement, par *Bastide et al.* [Bastide *et al.*, 2000a] et *Zaki* [Zaki, 2004]. La comparaison nous a permis de tirer les conclusions suivantes :

– La définition de *Bastide et al.* [Bastide *et al.*, 2000a] (*cf.* sous-section 2.2.1, page 68) souffre d'une forte rigidité. En effet, la condition d'avoir exactement les mêmes mesures de support et de confiance comme critères d'élimination, échoue pour exclure les règles redondantes à prémisses identiques et à conclusions comparables selon l'inclusion ensembliste. Ceci induit la génération d'un nombre élevé de règles d'association, parmi lesquelles, certaines associations véhiculent la même information. Par exemple, la lecture de la TABLE 2.1 (page 78), montre que cette approche retient les règles C $\overset{5/6}{\Rightarrow}$ W et C $\overset{4/6}{\Rightarrow}$ AW. Par contre, notre approche ne dérive que la dernière règle. La motivation de ne garder que la règle C $\overset{4/6}{\Rightarrow}$ AW est qu'elle a une conclusion maximale en la comparant avec la première, ce qui peut être plus utile dans d'autres applications, telle que la recherche d'information.

| $\mathcal{GBA}$ |
|---|
| C $\overset{5/6}{\Rightarrow}$ W |
| C $\overset{4/6}{\Rightarrow}$ AW |
| C $\overset{4/6}{\Rightarrow}$ T |
| C $\overset{4/6}{\Rightarrow}$ D |
| W $\overset{4/5}{\Rightarrow}$ AC |
| W $\overset{3/5}{\Rightarrow}$ ACT |
| W $\overset{3/5}{\Rightarrow}$ CD |
| T $\overset{3/4}{\Rightarrow}$ ACW |
| D $\overset{3/4}{\Rightarrow}$ CW |
| A $\overset{3/4}{\Rightarrow}$ CTW |

| $\mathcal{GBE}$ |
|---|
| D $\overset{1,0}{\Rightarrow}$ C |
| T $\overset{1,0}{\Rightarrow}$ C |
| W $\overset{1,0}{\Rightarrow}$ C |
| A $\overset{1,0}{\Rightarrow}$ CW |
| DW $\overset{1,0}{\Rightarrow}$ C |
| AT $\overset{1,0}{\Rightarrow}$ CW |
| TW $\overset{1,0}{\Rightarrow}$ AC |

| $\mathcal{NRR}$ | |
|---|---|
| C $\overset{4/6}{\Rightarrow}$ D | TW $\overset{1,0}{\Rightarrow}$ A |
| C $\overset{4/6}{\Rightarrow}$ T | A $\overset{1,0}{\Rightarrow}$ W |
| C $\overset{5/6}{\Rightarrow}$ W | W $\overset{1,0}{\Rightarrow}$ C |
| W $\overset{4/5}{\Rightarrow}$ A | T $\overset{1,0}{\Rightarrow}$ C |
| W $\overset{3/5}{\Rightarrow}$ D | D $\overset{1,0}{\Rightarrow}$ C |
| A $\overset{3/4}{\Rightarrow}$ T | |
| D $\overset{3/4}{\Rightarrow}$ W | |
| T $\overset{3/4}{\Rightarrow}$ W | |

TABLE 2.1 – Règles générées à partir du contexte textuel donné dans la TABLE 1.2 (*cf.* page 53) pour $minsupp = 3$ et $minconf = 0.6$ : **(Gauche)** Règles d'association découvertes par l'approche de *Bastide et al.* ; **(Droite)** Règles d'association découvertes par l'approche de *Zaki*.

– D'un autre côté, *Zaki* [Zaki, 2004] a introduit une approche axiomatique afin de dériver l'ensemble de toutes les règles redondantes (*cf.* sous-section 2.2.2, page 70), sans garantir aucune minimalité, *i.e.*, la compacité de la base. Notons que l'auteur a considéré que les règles dérivées à partir des termsets fermés voisins dans le treillis de l'Iceberg $\mathcal{TA}$. Il a affirmé que les règles non retenues peuvent être déduites en appliquant l'axiome de transitivité à cette base ainsi que la relation d'ordre dans cette structure laticielle [Zaki, 2004]. Toutefois, la base $\mathcal{NRR}$ de Zaki ne couvre pas l'ensemble de toutes les règles valides, *i.e.,* que des règles d'association valides pourraient ne pas être générées avec le mécanisme d'inférence proposé. Ainsi, le système axiomatique est utilisé afin de minimiser la taille des prémisses et des conclusions des règles. Cependant, ceci conduit à l'élimination de règles

ayant des conclusions plus larges, donc plus intéressantes, en faveur de règles ayant des conclusions minimales. Il va sans dire que la même information véhiculée par une règle ayant la conclusion la plus large, peut être extraite à partir d'un ensemble de règles à conclusions plus petites, en appliquant l'axiome de transitivité [Luxenburger, 1991].
À titre d'exemple, à partir des résultats de l'approche de Zaki, illustrés dans la TABLE 2.1 (Droite), nous pouvons conclure que même si la règle C $\overset{4/6}{\Rightarrow}$ AW ne fait pas partie de la base générique proposée $\mathcal{NRR}$ par Zaki, il est possible de la découvrir en composant C $\overset{5/6}{\Rightarrow}$ W et W $\overset{4/6}{\Rightarrow}$ A.

## 2.5 Évaluation empirique de la base générique $\mathcal{MGB}$

Afin d'évaluer expérimentalement le taux de réduction réalisé par la base générique $\mathcal{MGB}$ par rapport à l'ensemble global de toutes les règles valides $\mathcal{VAR}$, nous avons mené une série d'expérimentations, en quatre étapes décrites comme suit [Latiri *et al.*, 2012b] :

1. Pour extraire l'ensemble des termsets fermés fréquents, *i.e.*, $\mathcal{TFF}$ associés à leurs générateurs minimaux respectifs, nous avons adapté l'algorithme GC-GROWTH [Haiquan *et al.*, 2005] à notre contexte d'extraction textuel $\mathfrak{M} = (\mathcal{C}, \mathcal{T}, \mathcal{I})$.
2. Les règles d'association non-redondantes approximatives et exactes entre termes sont ensuite générées à partir de l'ensemble des termsets fermés fréquents $\mathcal{TFF}$, *i.e.*, la base générique $\mathcal{MGB}$ est dérivée, en utilisant l'algorithme GEN-MGB décrit précédemment.
3. L'ensemble de toutes les règles d'association valides $\mathcal{VAR}$ est généré, *i.e.*, celles qui sont redondantes, en utilisant l'implémentation de l'algorithme APRIORI [Agrawal and Skirant, 1994] proposée par Bart Goethals [11].
4. Le taux de réduction que réalise la base $\mathcal{MGB}$ par rapport à l'ensemble total des règles valides $\mathcal{VAR}$, est calculé comme suit pour les différentes collections de documents :

$$Taux\_réduction = \frac{|\mathcal{VAR}| - |\mathcal{MGB}|}{|\mathcal{VAR}|} \tag{2.9}$$

L'équation (2.9) représente en effet le taux de compression de l'ensemble total de règles d'association valides $\mathcal{VAR}$.

Dans l'objectif de comparer le taux de réduction donné par notre base générique $\mathcal{MGB}$ avec ceux réalisés avec les autres bases génériques décrites dans la section 2.2, nous avons présenté dans [Latiri *et al.*, 2012b] les résultats des évaluations expérimentales menées sur des bases de données de référence, fréquemment utilisées pour évaluer les performances des méthodes dédiées à l'extraction des règles d'association dans le domaine de data mining [12]. Le choix de ces benchmarks de test est motivé par le fait que les approches d'extraction de bases génériques citées dans la littérature ne sont pas dédiées pour les contextes d'extraction volumineux, surtout par rapport au nombre de termes distincts, comme c'est le cas pour les corpus de textes ou les collections de documents que nous utilisons dans le cadre de nos contributions en ECT. Ceci explique les optimisations que nous avons introduites dans l'implémentation de l'algorithme GEN-MGB pour l'adapter à l'exploration de la base générique de règles d'association non-redondantes entre termes.

11. *http ://www.adrem.ua.ac.be/~goethals/software/*
12. Les benchmarks de test sont disponibles sur le site : *http ://fimi.cs.helsinki.fi/data.*

### Description des collections

Nos expérimentation ont été conduites sur cinq collections de documents, à savoir [Latiri *et al.*, 2012b] :

- Les collections OFIL et INIST de la deuxième campagne d'évaluation Amaryllis [13].
- Les deux collections Le Monde 94 et ATS 94 de la campagne CLEF 2003 (Collection 2001) [14].
- La cinquième collection représente la composition des deux collections Le Monde 94 et ATS 94.

Un ensemble de requêtes est associé à chaque collection et, pour chaque requête, un ensemble de documents pertinents lui est affecté. La Table 2.2 donne plus de détails sur les collections utilisées.

| **Campagne** | **Collection** | **Taille (Mo)** | **# Doc.** | **# Termes ≠** | **# Req.** |
|---|---|---|---|---|---|
| Amaryllis II | OFIL | ≈ 33 | 11. 016 | 119. 434 | 26 |
| | INIST | ≈ 68 | 163. 307 | 174. 659 | 30 |
| CLEF 2003 | Le Monde 94 | ≈ 158 | 44. 013 | 106. 558 | 50 |
| | ATS 94 | ≈ 86 | 43. 178 | 55. 526 | 50 |
| | Le Monde 94 & ATS 94 | ≈ 244 | 87. 191 | 113. 422 | 50 |

Table 2.2 – Caractéristiques des collections utilisées.

Notons que les collections de documents OFIL, Le Monde 94 et ATS 94 sont composées d'articles extraits de journaux français, tandis que la collection INIST contient des résumés d'article scientifiques extraits à partir des bases de données *PASCAL* et *FRANCIS* [15]. Du point de vue des caractéristiques des différentes collections utilisées, les collections OFIL, Le Monde 94 et ATS 94 sont moins volumineuses que la collection INIST en terme de nombre de documents. Cependant, ses documents sont plus courts, *i.e.,* en terme de nombre de mots dans un document, étant donné que le vocabulaire est de nature scientifique.

### Pré-traitement des collections de documents

Afin d'extraire les termes les plus représentatifs des différentes collections, un prétraitement linguistique est effectué sur chaque collection de documents en utilisant l'analyseur morpho-syntaxique Cordial [16]. Nous avons effectué cette analyse en considérant uniquement les termes qui sont classés grammaticalement, soit comme des substantifs propres ou des substantifs communs. Une liste de mots "outils" est utilisée pour éliminer les termes fonctionnels de la langue française, qui s'avèrent souvent très fréquents.

13. Amaryllis est une Action de Recherche Concertée (ARC), organisée par l'Institut National français de l'Information Scientifique et Technique (INIST), avec le soutien de l'Agence Francophone pour l'Enseignement Supérieur et la Recherche (AUPELF-UREF) et le Ministère français de l'Education Nationale de la Recherche et de la Technologie (MERT). Deux cycles du projet ont déjà eu lieu, l'un en 1996-1997 et l'autre en 1998-1999. La méthodologie employée dans le projet Amaryllis est très proche de celle de Trec.

14. Le "Cross-Language Evaluation Forum" (CLEF) offre des collections de données afin d'évaluer les systèmes de recherche d'information (*http ://www.clef-campaign.org/*).

15. L'Inist-Cnrs produit les bases de données bibliographiques, multilingues et multidisciplinaires, *PASCAL* et *FRANCIS* qui, avec 20 millions de références, recensent l'essentiel de la littérature scientifique internationale (*http ://www.inist.fr/*).

16. Distribué par *Synapse Development* Corporation.

Le contexte d'extraction textuel document-terme $\mathfrak{M} = (\mathcal{C}, \mathcal{T}, \mathcal{I})$ est alors construit en conservant uniquement les termes correspondants aux catégories grammaticales sélectionnées. Les règles d'association sont ensuite générées en utilisant l'algorithme Gen-MGB. Le seuil de confiance minimale est fixé à 0%[17] et nous avons varié les seuils minimal et maximal du support, *i.e., minsupp* et *maxsupp*[18]. Ces seuils de support sont définis par rapport à la taille de la collection de documents et de la distribution statistique des termes. Pour cela, nous nous sommes basés sur la distribution de *Zipf* [Lafouge and Boukacem, 2004] relative à chaque collection pour déterminer expérimentalement le seuil *maxsupp*, afin d'élaguer les termes triviaux qui apparaissent dans la plupart des documents de la collections. D'autre part, le seuil de support minimal est également défini à partir de la représentation *Zipfienne*, permettant d'écarter les termes marginaux qui ne sont pas statistiquement significatifs, *i.e.,* les termes rares qui apparaissent seulement dans quelques documents de la collection.

**Résultats de la réduction**

La Table 2.3 résume l'ensemble des résultats expérimentaux, en terme de nombre de règles d'association découvertes et de taux de réduction pour les différents intervalles de support, *i.e.*, [*minsupp*, *maxsupp*], respectivement pour les différentes collections de documents utilisées [Latiri *et al.*, 2012b].

| **Intervalle de support** | **Taille($\mathcal{VAR}$)** | **Taille($\mathcal{MGB}$)** | **Taux de réduction (en %)** |
|---|---|---|---|
| OFIL | | | |
| [5, 50] documents | 235. 806 | 5. 761 | 97.56 |
| [50, 1000] documents | 291. 062 | 85. 878 | 70.49 |
| [1000, 5000] documents | 374 | 257 | 31.28 |
| INIST | | | |
| [3, 30] documents | 5. 154 | 3. 062 | 40.59 |
| [30, 250] documents | 472 | 273 | 42.16 |
| [250, 16000] documents | 11. 012 | 8. 949 | 18, 73 |
| Le Monde 94 | | | |
| [150, 1500] documents | 1. 965. 766 | 716. 842 | 63.53 |
| [200, 2000] documents | 709. 904 | 300. 493 | 57.67 |
| [300, 3000] documents | 171. 846 | 89. 072 | 48.17 |
| ATS 94 | | | |
| [150, 1500] documents | 113. 220 | 42. 393 | 62.56 |
| [200, 2000] documents | 44. 574 | 22. 297 | 49.98 |
| [300, 3000] documents | 15. 624 | 8. 816 | 43.57 |

Table 2.3 – Résultats de la réduction sur les collections de documents OFIL, INIST, Le Monde 94 et ATS 94 (toutes les règles valides *vs* $\mathcal{MGB}$).

Nous constatons que pour la collection OFIL, un nombre très élevé de règles d'association entre termes est découvert pour l'intervalle de support entre 5 et 1, 000 documents. Ce constat

17. Aucune règle d'association de confiance égale à 0 n'est dérivée.
18. *maxsupp* signifie que le termset doit apparaître dans la collection au maximum autant de fois que la valeur du seuil prédéfinie.

n'est pas en contradiction avec la distribution statistique des termes de la collection OFIL, où le support des termes est considéré comme élevé entre 25 et 200 documents. Notons que pour la collection INIST, un nombre important d'associations entre terms a été généré pour l'intervalle de support entre 25 et 500 documents. Ceci se justifie par l'importance des fréquences des termes dans la collection INIST dans cet intervalle.

Par ailleurs, la collection de documents Le Monde 94 présente le contexte d'extraction "pire des cas", par rapport à l'opérateur de la fermeture de Galois, où chaque termset fermé fréquent (TFF) est exactement égal à son générateur minimal. Ceci est vérifié pour toutes les valeurs de *minsupp* testées, même pour des seuils très faibles. À titre d'exemple, pour un seuil de $minsupp = 150$, il existe autant de TFFs que de générateurs minimaux, soit 310. 181. De plus, chaque termset fermé, représente généralement un générateur minimal, à l'exception de quatre termsets (sur 310, 185 termsets). Par conséquent et contrairement aux précédentes collections de documents, un nombre très réduit de règles d'association *exactes* avec une conclusion non-vide est généré à partir de cette collection, puisque chaque générateur minimal fréquent $g$ est lui même un TFF. Une règle de la forme $g \Rightarrow \emptyset$ est en effet considérée comme non-informative et elle est donc éliminée. Il est intéressant de noter que l'espace de recherche associé à la collection Le Monde 94, tend à être un espace fermé antimatroïde [Pfaltz and Taylor, 2002]. Ceci correspond au cas où chaque sous-ensemble de documents du contexte d'extraction partage des motifs communs, *i.e.,* un termset fermé, qui est à son tour un générateur minimal.

Nous remarquons aussi d'après les évaluations expérimentales, que la collection ATS 94 a les mêmes caractéristiques que celles de la collection Le Monde 94. En effet, le nombre de générateurs minimaux est approximativement toujours égal à celui des TFFs. Par exemple, pour un seuil de $minsupp = 100$, nous avons généré 60, 709 TFFs, tandis que le nombre de générateurs minimaux est égal à 60, 949. La différence de 240 générateurs constitue moins que 0.4% de l'ensemble global de générateurs minimaux. Notons aussi, que pour un seuil très faible *minsupp*, le nombre de termsets fréquent atteint 62, 847. Ainsi, nous pouvons déduire que le nombre de termsets qui ne sont pas fermés ne dépasse pas 3.52% de l'ensemble global des termsets fréquents. Il en résulte de cette caractéristique de la collection ATS 94, qu'un nombre très réduit de règles d'association exactes, avec des conclusions non-vides, est généré.

## 2.6 Bilan des contributions

L'abondance d'algorithmes d'extraction de termset fermés fréquents, de plus en plus efficaces, a permis d'améliorer les performances du processus d'exploration de corpus de textes très volumineux. Cependant, le nombre de règles d'association entre termes découvertes reste prohibitif et freine l'utilisation en aval de ces associations dans d'autres applications connexes à l'ECT. Ceci a motivé des travaux pour la proposition de nouvelles approches d'extraction de sous-ensembles réduits de règles d'association, appelés bases génériques, contenant les règles qui véhiculent le maximum d'informations utiles. Une revue de littérature relative aux différentes approches a permis de dégager les couples de bases génériques $(\mathcal{GBE}, \mathcal{GBA})$ et $(\mathcal{GBE}, \mathcal{TGBA})$ [Bastide *et al.*, 2000a]), la base générique $\mathcal{IGB}$ [Ben Yahia *et al.*, 2009] et les bases introduites par Balcazar [Balcázar, 2010] peuvent être qualifiées d'approches sans perte d'information. En outre, nous avons noté que l'approche d'extraction avec perte d'information de la base $\mathcal{NNR}$, proposée par Zaki [Zaki, 2004], présente quelques limites du fait qu'elle ne couvre pas l'ensemble de toutes les règles valides et qu'elle exige la vérification de l'égalité des mesures de support et de confiance tester la redondance d'une règle par rapport à une autre. Bien que la base $\mathcal{NNR}$ soit informative, l'utilisation des axiomes d'augmentation et de transitivité ne garantit pas la

génération de toutes les règles valides. Par conséquent, le mécanisme d'inférence utilisé par Zaki n'est pas complet et l'extraction de la base $\mathcal{NNR}$ se fait avec perte d'information. Il importe de noter que la faiblesse principale dans le contexte d'ECT est que cette base englobe des règles d'association ayant des conclusions minimales.

En se positionnant par rapport aux approches d'extraction de bases génériques de règles d'association de l'état de l'art, nous avons proposé une nouvelle définition d'une base générique minimale, appelée $\mathcal{MGB}$, dérivée à partir de corpus de textes volumineux, et permettant d'assurer un compromis entre l'informativité et la compacité de cette dernière. La spécificité de la base proposée est qu'elle est compacte dans le sens où elle englobe un noyau minimal de règles d'association entre termes, qui sont approximatives et exactes non-redondantes. Ces règles sont générées à partir de la structure ordonnée du treillis de l'Iceberg de Galois $\mathcal{TA}$. Leur caractéristique clé est qu'elles ont des prémisses minimales illustrées par les générateurs minimaux et des conclusions maximales. En effet, cette conclusion maximale, *i.e.,* englobant le maximum de termes, peuvent exprimer les termes candidats pour étendre une requête originelle ou encore les traductions potentielles des termes de la prémisse dans un contexte de traduction automatique. Nous montrerons dans les chapitres qui suivent l'utilité et l'apport de la base générique minimale $\mathcal{MGB}$ dans des applications réelles telles que la RI ou la TAS.

L'étude empirique menée sur cinq collections de documents a confirmé le fait que la base $\mathcal{MGB}$ apporte des gains en terme de compacité par rapport aux bases génériques extraites avec et sans perte d'information.

# Chapitre 3

# Règles d'association entre termes et ontologie au service de la RI

**Sommaire**



## 3.1 Objectifs du chapitre

Les contributions exposées dans ce chapitre se situent dans le croisement de l'ECT et de la Recherche d'Information (RI). L'étude proposée est consacrée à l'exploitation de la base générique de règles d'association entre termes $\mathcal{MGB}$ pour l'amélioration des résultats de la recherche d'information [Latiri *et al.*, 2012b, Latiri *et al.*, 2012a]. Notre principale motivation découle du fait que, dans la pratique, les Systèmes de Recherche d'Information (SRIs) font apparaître des

écarts notables entre la pertinence utilisateur et la pertinence système. Ces écarts, mesurés en terme de rappel et de précision, sont liés essentiellement à deux problématiques très abordées en RI : (*i*) l'imperfection de l'indexation automatique des documents, due aux problèmes d'interprétation du langage naturel et au traitement des ambiguïtés ; et, (*ii*) les problèmes posés par une interrogation effectuée avec une requête originelle de l'utilisateur qui manque souvent de précision.

Bien entendu, ces problèmes ne sont pas disjoints. Les travaux décrits dans ce chapitre traitent des deux problématiques à travers l'utilisation de la base générique $\mathcal{MGB}$ dans une approche d'expansion automatique de requêtes [Latiri *et al.*, 2012b], d'une part, et dans la proposition d'une nouvelle approche d'indexation conceptuelle, d'autre part. Cette dernière est basée sur une ontologie de domaine enrichie par la base $\mathcal{MGB}$ [Ben Ghezaiel *et al.*, 2010, Ben Ghezaiel *et al.*, 2011, Latiri *et al.*, 2012a].

## 3.2 Expansion de requêtes en RI par la base générique $\mathcal{MGB}$

La RI étudie le processus d'adéquation entre la requête d'un utilisateur et une collection de documents, dont le résultat est souvent un sous-ensemble de documents pertinents contenant les mêmes termes de la requête originelle. Le modèle classique de RI [Salton and McGill, 1983] consiste à attribuer, à chaque document d'une collection, des termes d'indexation, dits *index* du document, limitant les requêtes à l'ensemble global des termes de l'index, et utilisant des mesures de correspondance entre les requêtes et les documents.

Une des difficultés rencontrées au cours d'une session de recherche documentaire est liée aux choix des termes d'interrogation. En effet, dans bien des cas, les termes de la requête, exprimée par l'utilisateur, ne correspondent pas exactement aux descripteurs des documents retenus par le modèle d'indexation. De ce fait, afin d'avoir des documents pertinents, l'utilisateur est contraint d'utiliser le "*vocabulaire de description du document*" propre au système. Face à cette contrainte, il est possible de faire appel à la technique *d'expansion de requêtes* [Buckley *et al.*, 1994] afin d'améliorer la correspondance requête/document, et ce en étendant la requête par des termes additionnels, corrélés à ceux de la requête originelle. Intuitivement, l'apport d'une telle technique ne se réduit pas à l'amélioration du rappel en récupérant des documents pertinents qui ne peuvent pas être trouvés par la requête utilisateur, mais également à améliorer la précision des documents trouvés en les plaçant en haut de la liste des documents pertinents.

L'issue de recherche que nous suggérons est le déploiement des règles d'association entre termes [Agrawal and Skirant, 1994] dans un processus d'expansion de requêtes en RI [Haddad *et al.*, 2000, Lin *et al.*, 2008, Rungsawang *et al.*, 1999, Tangpong and Rungsawang, 2000]. Intuitivement, une règle d'association traduit la probabilité d'avoir les termes de la conclusion dans un document, sachant que ceux de la prémisse y sont. Ainsi, l'utilisation de telles dépendances dans un processus d'expansion de requêtes améliore sensiblement la pertinence d'un SRI, car elles reflètent des corrélations fortes et implicites découvertes à partir de la collection de documents. Toutefois, face au nombre très important de règles d'association entre termes qui peuvent être découvertes à partir d'une collection de documents, nous proposons un nouveau processus d'expansion automatique de requêtes moyennant la base générique minimale $\mathcal{MGB}$ [Latiri *et al.*, 2012b].

Avant de décrire le nouveau processus d'expansion automatique de requêtes proposé, nous allons présenter, dans ce qui suit, quelques travaux de l'état de l'art relatifs à l'expansion de requêtes.

### 3.2.1 Travaux reliés à l'expansion de requêtes en RI

La problématique d'expansion de requêtes a été largement abordée par la communauté RI depuis deux décennies [Buckley *et al.*, 1994, Ruthven, 2003, Joho *et al.*, 2004, Kumaran and Allan, 2008]. Les différentes approches proposées peuvent être groupées en deux principales classes selon le type de connaissances utilisé lors de l'expansion, à savoir :

1. **Expansion à partir des collections de documents :** Ces approches d'expansion utilisent des connaissances dérivées à partir des collections de textes. Elles observent généralement la régularité des termes dans un contexte déterminé d'une collection de textes. Elles sont basées sur l'hypothèse qui stipule que "*L'emploi de deux termes en co-occurrence est l'expression d'une relation sémantique entre eux*" [Rijsbergen, 1979]. L'avantage de ces approches est qu'elles sont faciles à mettre en œuvre tout en étant indépendantes du corpus.

   Parmi les premiers travaux relatifs à cette classe d'approches, Grefensette [Grefenstette, 1992] contribue par une approche syntaxique d'extraction de contextes de mots à partir des corpus textuels pour produire la liste des mots reliés à n'importe quel mot du corpus. Ces mots reliés seront utilisés dans l'expansion de requêtes.

   D'autres approches s'appuient sur une analyse statistique des collections par l'extraction de règles d'association entre termes, afin d'ajouter des termes voisins à la requête originelle [Tangpong and Rungsawang, 2000, Haddad *et al.*, 2000]. Les associations sont généralement basées sur la co-occurrence des termes dans les documents [Rungsawang *et al.*, 1999, Sun *et al.*, 2006, Lin *et al.*, 2008]. L'usage d'une telle technique a montré que les liens inter-termes renforcent la notion de pertinence des documents par rapport aux requêtes.

2. **Expansion à base de ressources externes existantes :** Certaines approches d'expansion de requêtes utilisent un vocabulaire contrôlé issu de ressources lexicales et sémantiques externes, tels que les thésaurus [Croft and Yufeng, 1994, Voorhees, 1994, Hu *et al.*, 2009] et les ontologies [Song *et al.*, 2007]. Un panorama d'approches d'expansion de requêtes à base de ressources externes de connaissances est présenté dans [Bhogal *et al.*, 2007].

   Les premiers travaux, qui ont fait appel à des ressources lexicales externes de type dictionnaire ou thésaurus, ont d'abord tenté d'utiliser la base lexicale WordNet pour réaliser l'expansion de requêtes [Voorhees, 1993]. Plus récemment, d'autres travaux ont proposé des approches d'enrichissement de requêtes utilisant Wikipédia comme collection externe [Koolen and Kamps, 2011].

   En outre, parmi les approches utilisant les connaissances du domaine pour l'expansion de requêtes, les auteurs dans [Bodner and Song, 1996, Yang and Yao, 2005] ont proposé des mécanismes d'expansion de requêtes par des termes liés dans un réseau sémantique, représentant une base de connaissances spécifique d'un domaine. Dans [Revuri *et al.*, 2006], Revuri *et al.* définissent différents cas d'expansion de requêtes avec divers éléments d'une ontologie de domaine tels que les concepts, les propriétés et les instances. Une autre technique d'expansion sémantique de requêtes combinant les règles d'association entre termes et une ontologie de domaine est proposée dans [Song *et al.*, 2007].

   Par ailleurs, dans le domaine de la RI médicale, K. Friberg [Friberg, 2007] s'appuie sur l'ontologie médicale MeSH pour étendre une requête par des mots simples intervenant dans la construction de certains mots composés. Également, dans [Díaz-Galiano *et al.*, 2008], les auteurs utilisent l'ontologie MeSH pour étendre les requêtes utilisateurs avec des termes médicaux.

En se positionnant par rapport aux approches d'expansion de requêtes de l'état de l'art, nous proposons, dans ce qui suit, une approche statistique d'expansion automatique de requêtes, basée

sur les co-occurrences de termes. Les corrélations entre les termes sont extraites par une technique de fouille globale de la collection de documents. Notre approche a pour caractéristique, comme les autres méthodes d'expansion de requêtes susmentionnées, de représenter un document et une requête par des vecteurs de pondérations, dans lesquels chaque pondération est associée à un terme.

### 3.2.2 Processus d'expansion automatique de requêtes par la base $\mathcal{MGB}$

Notre approche d'expansion automatique de requêtes consiste à dériver, dans un premier temps, la base générique $\mathcal{MGB}$ de règles d'association non-redondantes entre termes à partir d'une collection de documents, et de l'utiliser, dans un deuxième temps, pour étendre la requête originelle de l'utilisateur. Il importe de souligner que la base générique $\mathcal{MGB}$ est mieux adaptée au processus d'expansion automatique de requêtes, dans lesquelles l'ensemble des termes originels sera étendu par les conclusions des règles valides de la base générique $\mathcal{MGB}$, et ayant les termes de la requête initiale dans leurs prémisses respectives. Rappelons que les règles d'association entre termes de la base $\mathcal{MGB}$ sont non-redondantes et qu'elles sont dotées d'une prémisse minimale et une conclusion maximale, ce qui offre plus de termes candidats pour l'expansion. Il est intéressant de mentionner que le processus de génération des règles d'association, à partir d'une collection de documents, est en effet réalisé en amont. Par conséquent, même si la collection de documents est volumineuse, cela n'affectera pas le processus d'expansion et d'interrogation d'une requête donnée.

Notre approche automatique d'expansion de requêtes peut donc être considérée comme une approche plus élaborée basée sur les co-occurrences de termes. Ceci est justifié par le fait qu'une règle d'association entre termes infère une relation globale entre les termes, qui ne dépend pas d'un document donné mais implique plutôt un ensemble de documents de la collection et caractérisant un ensemble de termes liés, *i.e.,* un termset. Dans ce cadre, contrairement à la technique d'analyse locale de co-occurrences de termes qui permet de dériver les corrélations entre les termes d'un document donné, *i.e.,* relations intra-document, les règles d'association entre termes se dérivent par un processus de fouille globale de toute la collection de documents et offrent par conséquent, des informations sur les corrélations de termes inter-documents ainsi que sur les relations intra-document. De ce fait, les règles d'association permettent d'expliciter des relations plus fines entre les termes que les approches classiques à base de co-occurrences de termes.

#### Étapes du processus d'expansion

Le processus d'expansion automatique de requêtes par la base générique $\mathcal{MGB}$ se déroule en trois étapes [Latiri *et al.*, 2012b] :

**Étape 1 : Évaluation des résultats pour les requêtes sans expansion** Il s'agit de déterminer la base d'évaluation comparative (baseline). La pertinence des documents retournés est estimée selon les mesures d'évaluation suivantes :

- Précision de la requête originelle (OQ) à 11 points de rappel (P@11).
- Les précisions à P@5, P@10, P@15, et P@30 documents pertinents restitués.
- La précision moyenne MAP (Mean Average Precision). Comme la courbe de la précision à 11 points de rappel, la MAP définit la performance globale d'un SRI.

**Étape 2 : Expansion automatique des requêtes par les règles d'association de la base $\mathcal{MGB}$** Il s'agit d'étendre chaque requête originelle de la collection par tous les termes qui apparaissent dans les conclusions des règles d'association, qui ont dans leurs prémisses respectives les termes de la requête originelle. Tous les champs de la requête, *i.e.,* titre, champs descriptifs et narratifs, sont utilisés lors du processus d'expansion.

À partir des règles d'association entre termes de la base $\mathcal{MGB}$, chaque terme de la requête est traité individuellement en le cherchant dans les prémisses minimales des règles valides. La requête sera ensuite enrichie par la conclusion maximale de chaque règle ayant comme prémisse le ou les termes de la requête originelle. De part sa forme maximale, la conclusion de la règle offre plus de termes candidats pour l'expansion.

Étant donnée une requête originelle OQ = $\{t_1,\ldots,t_n\}$, la requête étendue par les règles d'association de la base $\mathcal{MGB}$, notée EQ, est exprimée comme suit [Latiri *et al.*, 2012b] :

$$\forall\ R : T_1 \Rightarrow T_2, \text{une règle non-redondante} \in \mathcal{MGB};\ \textit{si}\ T_1 \subseteq \text{OQ},\ \textit{alors}\ \text{EQ} = \{\text{OQ} \cup T_2\}. \tag{3.1}$$

L'équation (3.1) signifie que si la prémisse de $R$ est contenue dans OQ, les termes de la conclusion seront ajoutés à EQ.

**Étape 3 : Évaluation des résultats avec expansion** Les requêtes étendues (EQ) sont évaluées selon les mêmes mesures que les requêtes originelles (OQ). Les résultats obtenus sont ainsi comparés avec ceux obtenus avec les requêtes sans expansion, afin de calculer l'amélioration apportée par l'injection des termes provenant des règles d'association entre termes.

Notons que la variation de la pertinence système, notée $\Delta$, est calculée comme suit [Latiri *et al.*, 2012b] :

$$\Delta = \frac{(\textit{Résultat avec expansion}) - (\textit{Résultat sans expansion})}{(\textit{Résultat sans expansion})} \tag{3.2}$$

## 3.3 Évaluation expérimentale de l'approche d'expansion

L'évaluation expérimentale a été menée avec le SRI Lemur [19], en utilisant les collections de test décrites dans la Table 2.2 (*cf.* page 80). Nous avons testé notre approche d'expansion avec trois schémas de pondérations, à savoir :

- Le modèle vectoriel $tf{\times}idf$ [Salton and Buckely, 1988].
- La pondération BM25$tf$ qui est une variante du schéma de pondération $tf{\times}idf$ basé sur le modèle probabiliste de RI [Robertson and Walker, 1994].
- La pondération Okapi BM25 qui est une méthode d'ordonnancement de la méthode Okapi, la plus connue des méthodes probabilistes, et ayant pour but de construire un modèle probabiliste qui prend en compte la fréquence des termes ainsi que la taille des documents [Jones *et al.*, 2000].

Nous discutions, dans ce qui suit, les résultats des différentes stratégies d'évaluation que nous avons menées [Latiri *et al.*, 2012b].

19. *http ://sourceforge.net/projects/lemur/*

### 3.3.1 Résultats et discussion

Lors de la génération de la base générique $\mathcal{MGB}$, nous avons fait varier les seuils minimal et maximal de support, *i.e., minsupp* et *maxsupp*. Rappelons que ces seuils sont définis pour éliminer, respectivement, les règles très rares et celles qui sont très fréquentes. Il importe de préciser que les valeurs de ces seuils ont été déterminées empiriquement, en étudiant la régularité sur la fréquence d'apparition des termes, moyennant les distributions de *Zipf* [Lafouge and Boukacem, 2004] respectives aux corpus des différentes collections. En effet, la représentation graphique de la distribution *zipfienne* relative au nombre d'occurrences des termes montre une courbe décroissante qui est découpée traditionnellement en trois zones : (*i*) une première zone décrivant l'information triviale et représentée par les descripteurs principaux définissant la collection ; (*ii*) une deuxième zone contenant l'information intéressante représentée par les termes qui permettent de construire les structures et les relations des différents sujets traités dans le corpus de textes ; et, (*iii*) une troisième zone représentant l'information marginale et le bruit, illustrée par les termes rares. Notons que les trois zones sont contiguës et que la deuxième zone constitue la cible des approches d'extraction de règles d'association à partir de textes. Ainsi, la distribution de *zipf* des termes traduit fidèlement la dispersion des termes dans le corpus. Ceci permet, dans le cadre de nos recherches, de délimiter l'intervalle de support qui favorise la découverte de règles d'association non triviales et non marginales. Nous préconisons une troncature qui consiste à supprimer les fortes et les faibles occurrences des termes.

Les résultats pour les différentes collections sont rapportés dans la Table 3.1. Dans cette table, "RANRs" désignent les règles non-redondantes approximatives, alors que "RENRs" désignent les règles exactes non-redondantes.

La Table 3.1 montre, pour les différentes collections testées, l'amélioration de la pertinence système en terme de précision moyenne à 11 points de rappel, réalisée avec les requêtes étendues par les règles d'association de la base générique $\mathcal{MGB}$. Pour les trois schémas de pondération considérés, les différentes stratégies d'expansion que nous avons réalisées améliorent la précision à 11 points de rappel et place la pondération Okapi BM25 devant celles de BM25$tf$ et de $tf \times idf$. Nous notons également une amélioration intéressante pour les larges collections, comme la cinquième collection qui est composée des documents du journal Le Monde 94 et de ATS 94.

La Figure 3.1 illustre l'évolution de la précision à 11 points de rappel (P@11) selon différentes stratégies d'expansion de requêtes. Nous remarquons que l'utilisation de toutes les règles d'association de la base générique $\mathcal{MGB}$ induit une amélioration de la P@11, qui est en moyenne meilleure que celle obtenue avec les règles d'association exactes, qui traduisent des corrélations fortes entre termes avec une confiance totale égale à 1 (*cf.* les courbes relatives à la collection OFIL dans la Figure 3.1). Ce constat est justifié par le fait que les règles d'association approximatives expriment des corrélations entre termes qui n'apparaissent pas toujours ensemble et elles ne sont pas explicitées par les règles exactes.

Par ailleurs, nous remarquons que, pour la collection INIST, les différentes stratégies d'expansion de requêtes avec les règles d'association de $\mathcal{MGB}$, *i.e.*, en considérant uniquement les règles exactes (RENRs), que les règles approximatives (RANRs) et toutes les règles de la base générique $\mathcal{MGB}$ donnent des variations significatives de la précision à 11 points de rappel avec un écart faible entre les trois stratégies. Ceci se justifie comme suit :

- INIST est une collection scientifique dans laquelle les termes ont de très faibles distributions et co-occurrences.
- Le fait que l'analyseur linguistique ne peut pas identifier les termes scientifiques de la collection INIST, implique qu'une grande partie du vocabulaire n'est pas utilisée puis-

| **Évaluation** | $tf \times idf$ | | **BM25*tf*** | | **Okapi BM25** | |
|---|---|---|---|---|---|---|
| **Collection** | **OFIL (Campagne Amaryllis II )** | | | | | |
| **Baseline (OQ)** | 24.83% | | 31.71% | | 31.44% | |
| $\mathcal{MGB}$**-5-50** | (EQ) | Δ | (EQ) | Δ | (EQ) | Δ |
| Utilisant les RANRs et RENRs | 32.06% | +29.11% | 36.82% | +16.11% | 37.61% | +19.62% |
| Utilisant uniquement les RANRs | 29.59% | +19.17% | 35.42% | +11.69% | 35.73% | +13.64% |
| Utilisant uniquement les RENRs | 28.12% | +13.25% | 33.78% | +06.52% | 34.09% | +08.42% |
| **Collection** | **INIST (Campagne Amaryllis II )** | | | | | |
| **Baseline (OQ)** | 15.47% | | 15.25% | | 15.48% | |
| $\mathcal{MGB}$**-3-30** | (EQ) | Δ | (EQ) | Δ | (EQ) | Δ |
| Utilisant les RANRs et RENRs | 17.34% | +12.21% | 18.10% | +18.68% | 18.91% | +22.15% |
| Utilisant uniquement les RANRs | 17.32% | +12.92% | 18.10% | +18.68% | 18.90% | +22.09% |
| Utilisant uniquement les RENRs | 17.36% | +12.21% | 18.13% | +18.88% | 18.91% | +22.15% |
| **Collection** | **Le Monde 94 (Campagne Clef 2003 )** | | | | | |
| **Baseline (OQ)** | 41.01% | | 42.56% | | 43.54% | |
| $\mathcal{MGB}$**-300-3000** | (EQ) | Δ | (EQ) | Δ | (EQ) | Δ |
| Utilisant les RANRs et RENRs | 42.59% | +03.85% | 43.73% | +02.74% | 45.04% | +03.44% |
| **Collection** | **ATS 94 (Campagne Clef 2003 )** | | | | | |
| **Baseline (OQ)** | 53.03% | | 55.89% | | 56.56% | |
| $\mathcal{MGB}$**-300-3000** | (EQ) | Δ | (EQ) | Δ | (EQ) | Δ |
| Utilisant les RANRs et RENRs | 53.48% | +0.84% | 55.92% | +0.05% | 57.01% | +0.79% |
| **Collection** | **Le Monde 94 & ATS 94 (Campagne Clef 2003 )** | | | | | |
| **Baseline (OQ)** | 44.49% | | 47.51% | | 48.48% | |
| $\mathcal{MGB}$**-300-3000** | (EQ) | Δ | (EQ) | Δ | (EQ) | Δ |
| Utilisant les RANRs et RENRs | 45.45% | +02.15% | 47.99% | +1.01% | 49.24% | +01.56% |

TABLE 3.1 – Apport de la base $\mathcal{MGB}$ dans l'expansion de requêtes en terme de P@11 (Les seuils de support minimal et maximal sont représentés par $\mathcal{MGB}$-*minsupp-maxsupp*).

qu'elle n'est pas correctement analysée. Par conséquent, ces termes ne seront pas pris en considération lors de l'extraction de la base de règles d'association entre termes $\mathcal{MGB}$.

Nous remarquons également que l'amélioration de la précision à 11 points de rappel est moins significative pour des seuils de support minimal élevés. En effet, l'extraction de règles d'association entre termes avec des seuils de support minimal élevés provoque la génération de règles triviales entre les termes qui sont jugés très fréquents dans la collection de documents. Par conséquent, l'expansion de requêtes, à l'aide de ces termes, n'améliore ni le rappel, ni la précision.

La TABLE 3.2 présente l'évolution des précisions exactes à faibles taux de rappel ($P$@5, $P$@10, $P$@15 et $P$@30 documents), réalisée par notre approche d'expansion et pour toutes les collections testées. Cette variation significative traduit une augmentation du nombre de documents pertinents restitués et réordonnés parmi les documents les mieux classés. À titre d'exemple, pour la collection composée conjointement des deux collections LE MONDE 94 & ATS 94, la précision exacte à 5 documents est égale à 97.60% avant l'expansion, ce qui est considéré comme une pertinence très élevée. Il importe de noter, que même dans ce cas de figure, notre approche d'expansion de requêtes à base de $\mathcal{MGB}$ permet de réaliser une variation significative de 0.4%.

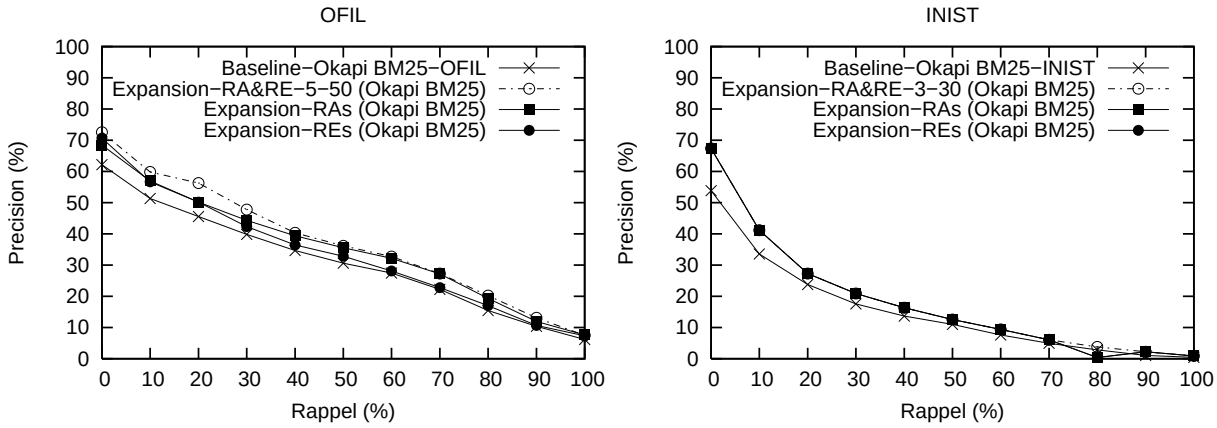


FIGURE 3.1 – Courbes Rappel/Précision relatives aux collections OFIL et INIST avec la pondération OKAPI BM25.

| **Évaluation** | P@5 | P@10 | P@15 | P@30 |
|---|---|---|---|---|
| **Collection OFIL (Campagne Amaryllis II)** | | | | |
| Baseline (OQ) | 40.77% | 37.69% | 33.33% | 27.31% |
| $\mathcal{MGB}$-5-50 (EQ) | 50.02% | 46.92% | 41.79% | 30.26% |
| Variation $\Delta$ | +22.68% | +24.48% | +25.38% | +10.80% |
| **Collection INIST (Campagne Amaryllis II )** | | | | |
| Baseline (OQ) | 32.67% | 30.00% | 28.22% | 21.44% |
| $\mathcal{MGB}$-3-30 (EQ) | 40.67% | 34.33% | 30.67% | 24.67% |
| Variation $\Delta$ | +24.48% | +14.43% | +8.68% | +15.06% |
| **Collection Le Monde 94 (Campagne Clef 2003 )** | | | | |
| Baseline (OQ) | 93.60% | 88.80% | 86.13% | 78.67% |
| $\mathcal{MGB}$-300-3000 (EQ) | 95.20% | 92.00% | 89.47% | 81.67% |
| Variation $\Delta$ | +1.70% | +3.60% | +3.87% | +3.81% |
| **Collection ATS 94 (Campagne Clef 2003)** | | | | |
| Baseline (OQ) | 93.33% | 91.04% | 90.56% | 84.10% |
| $\mathcal{MGB}$-300-3000 (EQ) | 96.0% | 94.60% | 94.00% | 86.33% |
| Variation $\Delta$ | +2.86% | +3.91% | +3.79% | +2.65% |
| **Collection Le Monde 94 & ATS 94 (Campagne Clef 2003 )** | | | | |
| Baseline (OQ) | 97.60% | 96.00% | 95.07% | 92.80% |
| $\mathcal{MGB}$-300-3000 (EQ) | 98.00% | 96.40% | 95.20% | 93.20% |
| Variation $\Delta$ | +0.40% | +0.41% | +0.13% | +0.43% |

TABLE 3.2 – Précision exacte à 5, 10, 15 et 30 documents en considérant le schéma de pondération OKAPI BM25.

### 3.3.2 Tests de significativité

Afin de mesurer la performance de notre approche d'expansion de requêtes par la base $\mathcal{MGB}$, nous avons choisi la précision moyenne MAP (Mean Average Precision), en utilisant les 1000 premiers documents restitués. Pour vérifier objectivement si la différence de performance, entre les approches sans expansion *vs* avec expansion, est statistiquement significative, nous avons adopté le test de rangs signés Wilcoxon [Smucker *et al.*, 2007], appliqué au cas d'échantillons

| **Tests** | **MAP** | ***p-value*** |
|---|---|---|
| ***tf*×*idf*** | | |
| OFIL-$\mathcal{MGB}$-5-50 | 0.3039 | 0.1060 |
| INIST-$\mathcal{MGB}$-3-30 | 0.1492 | 0.3190 |
| Le Monde 94-$\mathcal{MGB}$-300-3000 | 0.4104 | 0.0010 |
| ATS 94-$\mathcal{MGB}$-300-3000 | 0.5277 | 0.4300 |
| Le Monde 94 & ATS 94-$\mathcal{MGB}$-300-3000 | 0.4441 | 0.0310 |
| **BM25*tf*** | | |
| OFIL-$\mathcal{MGB}$-5-50 | 0.3497 | 0.0010 |
| INIST-$\mathcal{MGB}$-3-30 | 0.1629 | 0.0438 |
| Le Monde 94-$\mathcal{MGB}$-300-3000 | 0.4267 | $<$ 0.0001 |
| ATS 94-$\mathcal{MGB}$-300-3000 | 0.5599 | 0.0395 |
| Le Monde 94 & ATS 94-$\mathcal{MGB}$-300-3000 | 0.4778 | 0.0053 |
| **Okapi BM25** | | |
| OFIL-$\mathcal{MGB}$-5-50 | 0.3536 | 0.0696 |
| INIST-$\mathcal{MGB}$-3-30 | 0.1666 | 0.0840 |
| Le Monde 94-$\mathcal{MGB}$-300-3000 | 0.4394 | 0.0060 |
| ATS 94-$\mathcal{MGB}$-300-3000 | 0.5688 | 0.0240 |
| Le Monde 94& ATS 94-$\mathcal{MGB}$-300-3000 | 0.4884 | $<$ 0.0001 |

TABLE 3.3 – Résultats du test de Wilcoxon pour $\alpha = 5\%$.

appariés, qui est largement utilisé par la communauté de recherche en RI pour effectuer les tests de significativité.

La TABLE 3.3 rapporte les résultats des tests statistiques pour un niveau de signification $\alpha$ =5%. Nous remarquons que selon la mesure d'évaluation MAP, la meilleure variation significative, pour toutes les collections, est obtenue avec le schéma de pondération OKAPI BM25. Les *p-values* données par le test de Wilcoxon indiquent que les améliorations, pour les différentes stratégies d'expansion de requêtes par la base générique $\mathcal{MGB}$, sont significatives. Toutefois, en considérant le schéma de pondération $tf \times idf$, le test de Wilcoxon n'est significatif que pour la collection LE MONDE 94 & ATS 94 (avec une *p-value* égale à 0.0310) et pour la collection LE MONDE 94 (avec une *p-value* égale à 0.0010).

En effet, notre étude a montré globalement que les améliorations de la MAP, en considérant les pondérations BM25*tf* et OKAPI BM25, sont meilleures que celles obtenues avec la pondération $tf \times idf$.

Après avoir montré l'intérêt de l'utilisation de la base générique de règles d'association entre termes $\mathcal{MGB}$ dans l'expansion de requêtes en RI, nous proposons une deuxième contribution qui s'adresse plutôt à une problématique largement abordée en RI, à savoir l'imperfection de l'indexation automatique de documents. Notre contribution introduit une nouvelle approche d'indexation conceptuelle en RI, qui est fondée sur l'enrichissement d'une ontologie de domaine $\mathcal{O}$ par les règles d'association de la base générique $\mathcal{MGB}$ [Ben Ghezaiel *et al.*, 2011, Ben Ghezaiel *et al.*, 2012].

## 3.4 Enrichissement d'une ontologie de domaine par la base $\mathcal{MGB}$

Durant la dernière décennie, les travaux menés en RI ont largement bénéficié de l'apport de l'intelligence artificielle et l'ingénierie de connaissances (IC) [Nagypál, 2005, Hernandez *et al.*,

2007]. En outre, une des priorités de l'IC est de disposer de modèles structurés, définissant les notions-clés d'un domaine et permettant de raisonner sur ses connaissances, d'où l'utilisation des *ontologies* [Maedche and Staab, 2000, Mondary *et al.*, 2008]. Ces dernières sont considérées comme un moyen de représentation et de gestion de connaissances partageables au sein d'un domaine spécifique. Ainsi, en RI, les ontologies sont vues comme une version plus formelle des thésaurus, permettant d'élargir les possibilités de caractérisation des documents et des besoins d'information des utilisateurs [Charlet *et al.*, 2005].

Dans le cadre de nos recherches, nous abordons la question de l'utilisation des ontologies pour la RI du point de vue de l'IC, et plus particulièrement, sur leur mode d'enrichissement et sur la nature de leur contenu en terme de connaissances représentées. Nous pensons que les ontologies sont des représentations de connaissances pertinentes en RI du fait qu'elles comportent une dimension sémantique, et qu'elles peuvent être en outre rattachées à d'autres motifs fréquents issus de l'ECT. À cet égard, notre contribution définit un réseau conceptuel proxémique pour la représentation des connaissances d'un domaine, qui sera utilisé en aval pour l'indexation conceptuelle en RI.

### 3.4.1 Techniques d'enrichissement d'ontologies

La définition d'une ontologie, la plus citée dans la littérature, est donnée dans [Gruber, 1993] qui la présente comme un ensemble de *spécifications formelles explicites des termes d'un domaine et des relations entre eux*. Constituées de concepts liés par des relations, et souvent structurées hiérarchiquement, les ontologies permettent d'organiser des connaissances en fonction du domaine considéré. Ainsi, *une ontologie de domaine* traduit une conceptualisation formelle des connaissances implicites et explicites inhérentes au domaine. Plusieurs travaux de la littérature, liés au domaine de la RI et de l'ECT, ont prouvé que cette représentation s'avère utile pour améliorer l'efficacité de certaines applications tels que les modèles de RI à base d'ontologies [Castells *et al.*, 2007], l'expansion de requêtes [Bhogal *et al.*, 2007, Song *et al.*, 2007], l'indexation sémantique [Popov *et al.*, 2004], la fouille de textes [Müller *et al.*, 2004] et la classification de documents [Hotho *et al.*, 2003].

Cependant, se trouvant au cœur de plusieurs applications, les ontologies font l'objet de nombreux travaux de recherche [Castells *et al.*, 2007, Di-Jorio *et al.*, 2008, Benz *et al.*, 2010]. En particulier, l'évolution permanente des ressources textuelles nécessite la mise au point de techniques permettant l'enrichissement des ontologies et leurs mises à jours [Valarakos *et al.*, 2004, Di-Jorio *et al.*, 2008].

Dans la littérature, de nouvelles approches ont intégré l'utilisation de techniques d'ECT dans le processus d'enrichissement d'ontologies [Hoser *et al.*, 2006]. Certains travaux de recherche ont introduit des méthodes statistiques pour la découverte de nouveaux concepts candidats à l'enrichissement (un concept désigne un terme ou un termset), qui se basent généralement sur le nombre d'occurrences d'un terme. Ces méthodes sélectionnent souvent les termes en fonction de leur distribution dans le corpus [Faatz and Steinmetz, 2002, Parekh *et al.*, 2004], ainsi que sur d'autres mesures statistiques comme l'information mutuelle [Velardi *et al.*, 2001], ou encore des mesures calculant la probabilité d'occurrences d'un termset [Neshatian and Hejazi, 2004]. Toutefois, ces différentes approches ne réalisent pas le placement des nouveaux concepts dans l'ontologie initiale [Di-Jorio *et al.*, 2008].

D'autre part, des méthodes syntaxiques ont été aussi proposées pour l'identification des nouveaux concepts et relations. Elles se basent sur l'analyse grammaticale d'un terme ou d'une séquence de termes au sein d'une phrase [Bendaoud *et al.*, 2008]. Dans [Maedche *et al.*, 2002, Navigli and Velardi, 2006], les auteurs utilisent les motifs morpho-syntaxiques pour désigner

des relations ontologiques. Cependant, pour faire face au nombre très élevé de termes reliés, des techniques issues de l'ECT sont déployées, comme par exemple la découverte de règles d'association à partir de dépendances syntaxiques [Stumme and Maedche, 2001, Benz *et al.*, 2010]. L'avantage inhérent aux méthodes syntaxiques, par rapport aux méthodes statistiques, est le fait qu'elles permettent le placement automatique de nouveaux termes dans l'ontologie existante. Toutefois, ces méthodes ne permettent pas d'étiqueter les nouvelles relations.

D'autres travaux de la littérature ont abordé la problématique de placement des nouveaux concepts dans une ontologie originelle [Ganter and Stumme, 2003, Hernandez *et al.*, 2007, Di–Jorio *et al.*, 2008]. À ce sujet, dans [Faatz and Steinmetz, 2002], les auteurs ont suggéré d'utiliser la co-occurrence de termes, impliquant un ou plusieurs concepts de l'ontologie. Dans [Jorio *et al.*, 2007], l'approche proposée utilise les motifs séquentiels afin d'extraire les termes candidats à l'enrichissement, et de les corréler à la structure ontologique. D'autres approches issues de l'ECT, telles que l'extraction de règles d'association sont également utilisées pour placer précisément des concepts dans une ontologie, sans que les relations extraites ne soient étiquetées [Di-Jorio *et al.*, 2008]. Ce processus d'enrichissement reste semi-automatique, car d'une part, le nombre de règles d'association dérivées est très important, et, une intervention humaine est nécessaire pour définir sémantiquement les relations découvertes et les nommer, d'autre part.

### 3.4.2 Nouvelle approche d'enrichissement d'ontologies

Comme contribution à la problématique d'enrichissement d'ontologie, nous proposons d'utiliser la base générique minimale de règles d'association $\mathcal{MGB}$ [Latiri *et al.*, 2012b], afin d'enrichir automatiquement une ontologie de domaine et obtenir un *réseau conceptuel proxémique* [Ben Ghezaiel *et al.*, 2011]. Ce réseau intègre deux types de connaissances : des connaissances *explicites* provenant de l'ontologie initiale et des connaissances *implicites* issues de la base générique $\mathcal{MGB}$.

Formellement, nous considérons un contexte d'extraction textuel $\mathfrak{M} = (\mathcal{C}, \mathcal{T}, \mathcal{I})$ (*cf.* Définition 1, page 52) propre à un domaine spécifique et doté d'une ontologie $\mathcal{O}$. Par souci de conformité avec l'état de l'art, nous adoptons la définition suivante pour désigner une ontologie de domaine $\mathcal{O}$ [Stumme and Maedche, 2001, Cimiano *et al.*, 2005] :

**Définition 20** *Une ontologie est un quadruplet noté $\mathcal{O} := \langle \mathcal{C}_\mathcal{D}, \leq_\mathcal{C}, \mathcal{R}, \leq_\mathcal{R} \rangle$, tel que $\mathcal{C}_D$ est un ensemble de concepts d'un domaine spécifique, $\leq_\mathcal{C}$ est un ordre partiel sur $\mathcal{C}_\mathcal{D}$ (i.e., une relation binaire is-a $\subseteq \mathcal{C}_\mathcal{D} \times \mathcal{C}_\mathcal{D}$), $\mathcal{R}$ est un ensemble de relations, et $\leq_\mathcal{R}$ est la fonction qui attribut à chaque relation de $\mathcal{R}$ son arité.*

D'après [Stumme and Maedche, 2001], la définition ci-dessus considère les éléments clés de la plupart des langages de représentation d'ontologies.

Le processus général d'enrichissement automatique d'une ontologie que nous proposons, considère en entrée une ontologie de domaine $\mathcal{O}$ et une collection de documents associée au même domaine. Il se déroule en deux phases, que nous détaillons dans ce qui suit.

#### Phase 1 : Extraction de la base de règles d'association entre termes $\mathcal{MGB}$

Lors de cette phase, nous utilisons l'algorithme Gen-Mgb [Latiri *et al.*, 2012b] pour l'extraction de la base générique de règles d'association entre termes $\mathcal{MGB}$, décrit dans le chapitre 2 (*cf.* Algorithme 1, page 76).

En considérant le contexte d'extraction textuel $\mathfrak{M} = (\mathcal{C}, \mathcal{T}, \mathcal{I})$, nous adaptons la définition de la base $\mathcal{MGB}$ à la problématique d'enrichissement d'ontologie de domaine. Pour cela, nous

proposons de ne retenir que les règles d'association non-redondantes ayant un seul terme du domaine dans la prémisse, moyennant l'axiome de décomposition conditionnelle [Ben Yahia *et al.*, 2009], soit :

$$\begin{array}{c} \mathcal{MGB} = \{R : t \Rightarrow T_k \mid Conf(R) \geq minconf\} \\ \text{tel que } T_k = \{t_1, \ldots, t_k\} \subset \mathcal{T}\} \end{array} \tag{3.3}$$

À l'issue de cette phase, les règles d'association dérivées traduisent à travers de nouvelles corrélations entre termes, les concepts potentiellement candidats pour l'enrichissement d'ontologie.

**Phase 2 : Enrichissement de l'ontologie $\mathcal{O}$ par les règles d'association de la base $\mathcal{MGB}$**

Étant donné un domaine spécifique, nous supposons que nous disposons d'une ontologie de domaine $\mathcal{O}$ (telles que l'ontologie médicale MeSH [Díaz-Galiano *et al.*, 2008] ou Emwis traitant du domaine des eaux [Di-Jorio *et al.*, 2008]). Une telle ontologie inclut les primitives de base d'une structure ontologique, à savoir les concepts et les relations taxonomiques tel que le lien de subsomption "*is-a*". L'évaluation du lien sémantique entre les concepts de l'ontologie $\mathcal{O}$ est calculée à partir de la mesure de similarité de *Wu et Palmer's* [Wu and Palmer, 1994], qui prend en compte à la fois la profondeur des concepts dans la hiérarchie de concepts et la structure de cette dernière.

Ainsi, la similarité entre deux concepts $C_1$ and $C_2$ d'une ontologie $\mathcal{O}$ est estimée en se basant sur : $(i)$ les distances $depth(C_1)$ et $depth(C_2)$ qui séparent, respectivement, les concepts $C_1$ et $C_2$ de la racine de l'ontologie ; et, $(ii)$ la distance $depth(C)$ qui sépare l'ancêtre commun le plus proche $(C)$ de $C_1$ et $C_2$ à partir de la racine. Cette similarité se calcule comme suit [Wu and Palmer, 1994] :

$$Sim_{Wu}(C_1, C_2) = \frac{2 \times depth(C)}{depth(C_1) + depth(C_2)} \tag{3.4}$$

En utilisant la base générique de règles d'association entre termes $\mathcal{MGB}$, l'enrichissement d'une ontologie $\mathcal{O}$ consiste à rapprocher les concepts initiaux de cette dernière des termes qui figurent dans les prémisses des règles d'association de la base $\mathcal{MGB}$. Par la suite, les termes des parties conclusion de ces règles sont identifiés comme les concepts candidats pour l'enrichissement. Nontons que les liens sémantiques entre les concepts de l'ontologie enrichie, notée dans la suite $\mathcal{O}_{\mathcal{MGB}}$, sont pondérés selon une nouvelle mesure de similarité, que nous appelons $Sim_{\mathcal{O}_{\mathcal{MGB}}}$.

L'enrichissement se réalise itérativement selon les étapes suivantes :

1. **Étape 1 : Sélection des concepts candidats pour l'enrichissement**
   Pour chaque concept $C_o$ de l'ontologie initiale $\mathcal{O}$, l'ensemble des concepts candidats qui lui seront reliés est identifié. Cet ensemble englobe les termes figurant dans les conclusions des règles d'association valides dont la prémisse est $C_o$.
2. **Étape 2 : Placement des nouveaux concepts**
   Cette étape consiste à placer les concepts candidats tout en préservant la cohérence des concepts et des relations pré-établies dans l'ontologie initiale $\mathcal{O}$. Ceci permet de ne pas ajouter des redondances relationnelles dans le cas où un concept est candidat à être lié à plusieurs concepts de l'ontologie $\mathcal{O}$. Autrement dit, étant donnée une règle d'association

valide appartenant à la base $\mathcal{MGB}$, de la forme $R: C_o \Rightarrow C_j$, le concept candidat $C_j$ de l'association $R$ est retenu pour être relié au concept $C_o$ de l'ontologie $\mathcal{O}$ si et seulement si la règle d'association $R$ a la confiance la plus élevée parmi celles appartenant à $\mathcal{MGB}$ et ayant $C_o$ comme prémisse.

3. **Étape 3 : Calcul des mesures de similarité et du voisinage d'un concept dans $\mathcal{O}_{\mathcal{MGB}}$**
   Parmi les termes candidats à l'enrichissement, extraits des conclusions des règles d'association valides, certains sont des concepts initiaux de l'ontologie $\mathcal{O}$. Afin de ne relier que de nouveaux termes avec les concepts existants, nous définissons la notion de *voisinage* d'un concept $C_i$ appartenant à l'ontologie $\mathcal{O}_{\mathcal{MGB}}$ comme suit [Ben Ghezaiel *et al.*, 2010] :

   **Définition 21** *Le voisinage d'un concept $C_i$, noté $\mathcal{V}(C_i)$, représente l'ensemble des concepts de l'ontologie enrichie $\mathcal{O}_{\mathcal{MGB}}$, qui sont liés à $C_i$ soit par la hiérarchie dans l'ontologie initiale $\mathcal{O}$, soit par une ou plusieurs règles d'association de la base $\mathcal{MGB}$.*

   En effet, les relations entre un concept $C_i$ de l'ontologie initiale $\mathcal{O}$ et son voisinage sont évaluées par une mesure statistique que nous appelons *mesure de similarité* entre $C_i$ et son voisinage, que nous notons par $Sim_{\mathcal{O}_{\mathcal{MGB}}}$. Pour ce faire, en considérant une ontologie $\mathcal{O}$ et $\mathcal{C}_{\mathcal{O}}$ l'ensemble des ses concepts, il est nécessaire de disposer d'une fonction de similarité, notée par $Sim_C : \mathcal{C}_{\mathcal{O}} \rightarrow [0,1]$, si et seulement si $Sim_C(C) = 1$ et $0 \leq Sim_C(C_j) < 1 \ \forall \ C_j \neq C \in \mathcal{C}_{\mathcal{O}}$.
   Il importe de préciser que la mesure similarité sémantique que nous considérons n'est pas une distance, parce qu'elle ne satisfait pas la symétrie et l'inégalité triangulaire [Ventresque *et al.*, 2008]. Elle est calculée en fonction de : (*i*) la mesure de confiance des règles d'association retenues pour l'enrichissement de l'ontologie $\mathcal{O}$ ; et, (*ii*) la mesure de similarité calculée entre les concepts $\mathcal{C}_{\mathcal{O}}$ de la structure ontologique initiale $\mathcal{O}$, estimées par la mesure de *Wu et Palmer* [Wu and Palmer, 1994] (*cf.* équation (3.4)).
   Nous suggérons ainsi de calculer la similarité entre deux concepts $C_i$ et $C_j$ appartenant à l'ontologie enrichie $\mathcal{O}_{\mathcal{MGB}}$ selon trois cas de figures possibles, énonçés dans les propositions ci-dessous.

   **Proposition 2** *Soit un concept $C_i$ de l'ontologie initiale $\mathcal{O}$. S'il existe un concept $C_j$ en relation avec le concept $C_i$ et provenant d'une règle d'association valide de la base $\mathcal{MGB}$, alors :*

$$Sim_{\mathcal{O}_{\mathcal{MGB}}}(C_i, C_j) = Conf(R : C_i \Rightarrow C_j) \tag{3.5}$$

   **Proposition 3** *Si les deux concepts $C_i$ et $C_j$ appartiennent à l'ontologie initiale $\mathcal{O}$ alors :*

$$Sim_{\mathcal{O}_{\mathcal{MGB}}}(C_i, C_j) = Sim_{Wu}(C_i, C_j) \tag{3.6}$$

   **Proposition 4** *Si $C_i$ est un concept ajouté à l'ontologie $\mathcal{O}_{\mathcal{MGB}}$ et si il est lié au concept $C_k \in \mathcal{O}$ tel que $Sim_{\mathcal{O}_{\mathcal{MGB}}}(C_k, C_i) = Conf(R : C_k \Rightarrow C_i) = \alpha \geq minconf$, alors tout concept $C_j \in \mathcal{O}$ en relation avec $C_k$, tel que $Sim_{Wu}(C_k, C_j) = \beta$, est aussi en relation avec le concept $C_i$. Dans ce cas, la mesure de similarité est dite mixte et elle est calculée comme suit :*

$$Sim_{\mathcal{O}_{\mathcal{MGB}}}(C_i, C_j) = \alpha \times \beta \tag{3.7}$$

   Ainsi, nous considérons que le voisinage d'un concept $C_i$ englobe l'ensemble des $k$ concepts du réseau proxémique conceptuel $\mathcal{O}_{\mathcal{MGB}}$ en relation avec le concept $C_i$, tel que la mesure de similarité entre eux est supérieure ou égale à un seuil fixé au préalable $\delta$, soit :

$$\mathcal{V}(C_i) = \{C_j \mid Sim_{\mathcal{O}_{\mathcal{MGB}}}(C_i, C_j) \geq \delta, \ j \in [1..k]\} \tag{3.8}$$

En guise d'application de notre approche d'enrichissement d'ontologie, nous avons choisi le domaine de la neurologie pédiatrique, et plus particulièrement la maladie des dystonies [20]. Nous avons détaillé dans [Ben Ghezaiel *et al.*, 2010] les résultats de la validation expérimentale de l'approche d'enrichissement proposée dans ce domaine.

### 3.4.3 $\mathcal{O}_{\mathcal{MGB}}$ : Un réseau conceptuel proxémique pour la représentation des connaissances

Le résultat généré par le processus d'enrichissement d'ontologie à base de règles d'association est exploré en tant que *réseau conceptuel proxémique* dans l'objectif de représenter les connaissances implicites et explicites du domaine. L'originalité de notre réseau réside dans sa généricité et son exhaustivité, qui sont dûes à la combinaison des connaissances de la structure ontologique d'une part, et aux connaissances implicites issues de l'application de la technique d'extraction de règles d'association, d'autre part.

Nous considérons qu'un concept du réseau conceptuel est ainsi doté de trois niveaux de proximité sémantique [Bachimont, 2000, Ben Ahmed, 2007] :

1. **Une sémantique référentielle** : Elle associe à chaque concept une référence intentionnelle. Il s'agit de trouver le sens qui correspond au mieux à un concept de $\mathcal{O}_{\mathcal{MGB}}$. Un processus de désambiguïsation des concepts est donc nécessaire pour déterminer la sémantique référentielle et ce en sélectionnant le sens le plus approprié pour chaque concept.
2. **Une sémantique différentielle** : Elle associe à chaque concept ses concepts voisins, *i.e.,* ceux qui sont utilisés en même temps que lui dans les contextes du domaine pour le définir par les similarités et les différences qu'il entretient avec ses voisins. Cette sémantique est assurée par le calcul du voisinage d'un concept (*cf.* Définition 21, page 97).
3. **Une sémantique inférentielle** : Elle est induite par les règles d'association. Elle associe à chaque concept un potentiel inférentiel, en reliant des concepts de l'ontologie initiale $\mathcal{O}$ aux concepts issus des règles d'association valides de la base générique $\mathcal{MGB}$.

En se basant sur les trois niveaux de sémantique cités ci-dessus, nous pouvons dire que, autour de chaque concept $C_i$ du réseau proxémique conceptuel $\mathcal{O}_{\mathcal{MGB}}$, gravite un sous-espace sémantique propre à $C_i$ et qui reflète ces différents niveaux de sémantique à travers des relations pondérées par la mesure de similarité sémantique, *i.e.,* $Sim_{\mathcal{O}_{\mathcal{MGB}}}$. Par conséquent, cette représentation conceptuelle des connaissances d'un domaine peut être qualifiée d'exhaustive, étant donné qu'elle permet de : (*i*) couvrir le vocabulaire des spécialistes d'un domaine ; et, (*ii*) expliciter des relations implicites entre les concepts du domaine. Ceci étend les possibilités d'interprétation et d'application des connaissances issues du réseau proxémique conceptuel $\mathcal{O}_{\mathcal{MGB}}$ dans d'autres problématiques.

Sur le plan applicatif, nous proposons d'utiliser le réseau conceptuel proxémique $\mathcal{O}_{\mathcal{MGB}}$ en RI, par la proposition d'une nouvelle approche d'indexation conceptuelle reposant sur les trois niveaux de sémantiques qu'il identifie. En effet, nous suggérons d'utiliser les relations entre les concepts du réseau proxémique, qui sont déjà pondérées par la mesure de similarité sémantique dans le processus de désambiguïsation des termes d'un document lors de la phase d'indexation. Nous proposons de l'utiliser ensuite, pour affecter un poids sémantique aux différents descripteurs des documents.

20. Cette phase expérimentale du travail a été effectuée en collaboration avec les médecins spécialistes du Service de Neurologie de l'Enfant et de l'Adolescent de l'Institut de Neurologie de Tunis.

## 3.5 Nouvelle approche d'indexation conceptuelle en RI

Durant la dernière décennie, de nombreux travaux en RI se sont orientés vers la prise en compte de la sémantique des termes dans le processus d'indexation [Baziz *et al.*, 2005, Castells *et al.*, 2007, Navigli, 2009]. Les méthodes utilisées sont censées améliorer les performances d'un SRI en termes de rappel et de précision pour le rendre capable de traiter l'ambiguïté des termes. Dans [Baziz *et al.*, 2005], les auteurs distinguent deux types d'approches d'indexation, à savoir : l'indexation *sémantique* et l'indexation *conceptuelle*.

L'indexation sémantique se base sur des techniques de désambiguïsation contextuelle des termes dans les documents et les requêtes [Sanderson, 1994, Mihalcea and Moldovan, 2000, Navigli, 2009, Dinh and Tamine, 2011]. L'idée est que le sens d'un mot est complètement déterminé par les autres mots occurrant dans le même contexte [Hernandez *et al.*, 2007]. Plus récemment, dans [Dinh and Tamine, 2011], les auteurs présentent un modèle d'indexation sémantique adapté aux dossiers électroniques de patients fondé sur l'utilisation de l'ontologie MeSH [Díaz-Galiano *et al.*, 2008].

Nous nous intéressons particulièrement à l'indexation conceptuelle qui se base sur des concepts issus de ressources externes telles que les ontologies [Vallet *et al.*, 2005] et les taxonomies [Maedche *et al.*, 2002], pour indexer les documents [Andreasen *et al.*, 2009], contrairement aux index de mots simples. Généralement, le processus d'indexation est effectué en deux étapes : $(i)$ projeter un document sur une ontologie afin de représenter au mieux le contenu sémantique d'un document et détecter ses termes les plus représentatifs [Baziz *et al.*, 2005] ; ensuite, $(ii)$ désambiguïser les termes sélectionnés précédemment.

Parmi les premiers travaux de l'indexation conceptuelle à base de l'ontologie WordNet, nous citons ceux proposés dans [Gonzalo *et al.*, 1998]. Les *synsets* sont ainsi intégrés dans le modèle vectoriel pour représenter l'espace d'indexation. Dans [Khan and Luo, 2002], les auteurs utilisent la notion de concept et proposent un algorithme permettant d'attacher les termes d'un texte aux concepts de l'ontologie en se basant sur la notion de région d'ontologie et de distance sémantique entre concepts. Une approche similaire est proposée dans [Baziz *et al.*, 2004], dans laquelle les termes d'un texte sont attachés aux concepts de WordNet en se basant sur la notion de similarité sémantique entre concepts. Un document est finalement représenté par un réseau sémantique de concepts et de relations sémantiques entre eux. Une autre approche est proposée dans [Navigli and Velardi, 2006] pour relier les concepts d'une ontologie de domaine à ceux d'un document. Les auteurs proposent aussi une mesure de similarité inter-concepts à travers le réseau sémantique de l'ontologie. Une autre idée est introduite dans [Amirouche *et al.*, 2008], où les auteurs proposent une approche d'indexation conceptuelle basée sur les *CP-Nets* (Conditional Preferences Networks). Le formalisme *CP-Net* est utilisé d'une part, pour la représentation graphique de requêtes flexibles exprimant des préférences qualitatives, et pour l'évaluation flexible de la pertinence des documents, d'autre part.

Nous décrivons, dans la suite de cette section, notre approche d'indexation conceptuelle en RI avec une proposition originale pour la représentation des documents [Ben Ghezaiel *et al.*, 2011, Ben Ghezaiel *et al.*, 2012]. Cette représentation se base sur le réseau conceptuel proxémique $\mathcal{O}_{\mathcal{MGB}}$. Il importe de souligner que les deux principales caractéristiques de notre réseau conceptuel proxémique sont : $(i)$ le modèle de représentation de connaissances est le résultat d'un processus complètement automatisé faisant appel à une ontologie enrichie par des règles d'association entre termes, dans le but d'identifier les concepts d'un document et de calculer les liens de proximité entre eux ; et, $(ii)$ les relations entre les concepts du réseau conceptuel ne sont pas étiquetées mais pondérées en fonction de la mesure de similarité sémantique, *i.e.,* $Sim_{\mathcal{O}_{\mathcal{MGB}}}$.

L'approche d'indexation conceptuelle de documents que nous proposons comprend trois

phases principales que nous détaillons dans ce qui suit.

### 3.5.1 Phase 1 : Identification et pondération des concepts représentatifs d'un document

L'objectif de cette phase est de retrouver, pour chaque terme d'indexation d'un document, le concept correspondant dans le réseau conceptuel proxémique $\mathcal{O}_{\mathcal{MGB}}$. Dans un premier temps, l'index d'un document est projeté sur $\mathcal{O}_{\mathcal{MGB}}$ pour identifier les concepts correspondants aux termes d'indexation. Ensuite, ces différents termes sont pondérés par leur degré de représentativité dans un document.

Nous définissons ainsi une nouvelle mesure de pondération des termes qui prend en compte à la fois la représentativité statistique et la représentativité sémantique des termes dans un document.

#### Identification des concepts candidats représentatifs d'un document

Nous considérons qu'un document $d$ d'une collection est représenté par l'ensemble des termes fréquents par rapport au seuil de *minsupp*, soit $d = \{t_1, \ldots, t_m\}$. Un terme noté par $t = \{w_1, \ldots, w_n\}$ est composé d'un ou de plusieurs mots. La longueur d'un terme $t$, notée $|t|$, est alors définie comme le nombre de mots dans $t$.

Le but de cette étape est d'identifier les termes et les multi-termes représentatifs d'un document, qui correspondent à des entrées dans le réseau conceptuel proxémique $\mathcal{O}_{\mathcal{MGB}}$. Pour cela, nous utilisons la méthode d'identification de concepts proposée dans [Amirouche *et al.*, 2008], que nous adoptons au réseau conceptuel $\mathcal{O}_{\mathcal{MGB}}$. Cette méthode est basée sur une analyse du document au niveau du mot. À l'issue de cette étape, nous aurons identifié les termes $t_i$ qui caractérisent le document $d$, noté par $T(d)$, soit :

$$T(d) = \{(t_1, f(t_1)), \ldots, (t_n, f(t_n))\}, \text{ tel que } t_i \in d \tag{3.9}$$

sachant que $f(t_i)$ désigne la fréquence d'occurrences du terme $t_i$ dans le document $d$.

#### Pondération des nouveaux concepts

Une fois les termes représentatifs d'un document extraits, il s'agit de leur affecter un poids qui détermine leur importance respective dans ce dernier. Dans les SRI classiques, plusieurs méthodes de pondération sont utilisées. Elles sont généralement des variantes de la pondération $tf \times idf$, exprimée par [Salton and Buckely, 1988] :

$$W_{(t,d)} = tf(t) \times idf(t, d) \tag{3.10}$$

où $tf$ représente la fréquence du terme $t$ et $idf$ est la fréquence inversée du document, soit :

$$idf(t, d) = \ln\left(\frac{N}{df(t)}\right) \tag{3.11}$$

sachant que $N$ est le nombre de documents dans la collection de documents et $df(t)$ est le nombre de documents de la collection contenant le terme $t$.

Cependant, les méthodes à base de $tf \times idf$ manquent d'efficacité et de pertinence quand il s'agit de pondérer des concepts multi-termes. Dans [Baziz *et al.*, 2005], les auteurs ont proposé une méthode de pondération statistique, appelée $cf \times idf$, qui tient compte de la taille du multi-terme, *i.e.,* du nombre de termes le composant, et de la mesure $tf \times idf$. Cette méthode de

pondération stipule que chaque terme extrait représente forcément un concept de l'ontologie WordNet, étant donné que la même ontologie est utilisée lors de l'identification des termes d'indexation. Ainsi, pour un multi-terme $t$ composé de $n$ termes, sa fréquence dans un document dépend du nombre d'occurrences du multi-terme lui-même, et de celui de tous ses sous-ensembles. Formellement, nous avons [Baziz *et al.*, 2005] :

$$cf(t) = f(t) + \sum_{t_i \in sub(t)} (\frac{|t_i|}{|t|} \times f(t_i)) \tag{3.12}$$

tel que $sub(t)$ est l'ensemble de tous les sous-ensembles de termes dérivés à partir du multi-terme $t$, $|t|$ représente la taille du multi-terme $t$ et $f(t)$ est la fréquence d'occurrences de $t$ dans le document $d$.

En se référant à la proposition décrite dans [Baziz *et al.*, 2005], nous définissons une nouvelle mesure de pondération d'un terme, qui prend en compte à la fois *la représentativité statistique* et *la représentativité sémantique* du terme dans le document. L'idée est que le poids d'un terme $t$ au sein d'un document $d$ se calcule sur la base de son poids $cf \times idf$ [Baziz *et al.*, 2005] et également en considérant les poids des concepts $C_i$ qui lui sont proches, *i.e.,* les concepts appartenant à son voisinage.

De ce fait, la *représentativité statistique*, notée $W_{Stat}$, est calculée sur la base de l'équation (3.12), comme suit [Baziz *et al.*, 2005] :

$$W_{Stat}(t, d) = cf(t) \times idf(t, d) \tag{3.13}$$

En considérant le réseau conceptuel proxémique $\mathcal{O}_{\mathcal{MGB}}$, nous définissons également la *représentativité sémantique* d'un terme $t$ dans un document $d$, que nous notons par $W_{Sem}(t, d)$. Celle-ci se base sur les différents liens existants dans $\mathcal{O}_{\mathcal{MGB}}$ entre chaque occurrence de $t$ et les concepts appartenant à son voisinage. Cette représentativité sémantique est calculée en utilisant la mesure de similarité sémantique $Sim_{\mathcal{O}_{\mathcal{MGB}}}$ définie dans les Propositions 2, 3 et 4 (*cf.* page 97) entre chaque occurrence $C_i$ du terme $t$ dans le réseau $\mathcal{O}_{\mathcal{MGB}}$ et les concepts de son voisinage $\mathcal{V}(C_i)$. Elle se calcule comme suit [Ben Ghezaiel *et al.*, 2011, Ben Ghezaiel *et al.*, 2012] :

$$W_{Sem}(t, d) = \sum_{C_i \in S_t} \sum_{C_j \in \mathcal{V}(C_i)} Sim_{\mathcal{O}_{\mathcal{MGB}}}(C_i, C_j) \times f(C_j) \tag{3.14}$$

tel que $S_t = \{C_1, \ldots, C_n\}$ est l'ensemble de tous les concepts liés au terme $t$ (occurrences de $t$).

L'idée sous-jacente est que la représentativité globale d'un terme $t$ dans un document $d$, notée $W_{Doc}$, est formulée par la combinaison entre la représentativité statistique et la représentativité sémantique. Cette représentativité évalue le poids d'un terme dans un document $d$ et se calcule comme suit [Ben Ghezaiel *et al.*, 2011, Ben Ghezaiel *et al.*, 2012] :

$$W_{Doc}(t, d) = W_{Stat}(t, d) + W_{Sem}(t, d) \tag{3.15}$$

Il en découle que l'index d'un document $d$, noté $Index(d)$, est généré en sélectionnant uniquement les termes et les multi-termes dont la représentativité globale, *i.e.,* $Rep_{Doc}$, est supérieure ou égale à un seuil minimal de représentativité.

### 3.5.2 Phase 2 : Désambiguïsation des concepts

L'objectif de la désambiguïsation est de retrouver le sens correct d'un terme dans son contexte d'énonciation. Dans ce cadre, nous proposons une approche de désambiguïsation basée sur le réseau conceptuel proxémique $\mathcal{O}_{\mathcal{MGB}}$. Celle-ci consiste à retrouver, pour chaque terme dans l'index d'un document $d$, *i.e.,* $Index(d) = \{t_1, \ldots, t_m\}$, tous les sens qui lui sont associés dans le réseau $\mathcal{O}_{\mathcal{MGB}}$, puis à le désambiguïser si nécessaire.

**Définition 22** *Chaque terme $t_i$ d'un document $d$, a un ensemble de sens noté par $S_i = \{C_1^i, \ldots, C_n^i\}$, représentant des concepts du réseau proxémique $\mathcal{O}_{\mathcal{MGB}}$. Ainsi, le terme $t_i$ est doté de $|S_i| = n$ sens correspondants à $n$ concepts différents dans le réseau $\mathcal{O}_{\mathcal{MGB}}$* [Ben Ghezaiel *et al.*, 2012].

En se référant à la Définition 22, nous postulons l'hypothèse que chaque terme d'index contribue à la définition de la sémantique d'un document $d$ avec uniquement un seul sens [Amirouche *et al.*, 2008] (même si un terme peut avoir différents sens dans un même document, nous retenons, dans notre approche, le sens le plus approprié). En se basant sur cette hypothèse, nous devons choisir pour chaque terme $t_i \in Index(d)$, son meilleur sens dans $d$.

Le principe de cette désambiguïsation consiste à supposer que, parmi les différents sens (qui correspondent à des concepts candidats) d'un terme donné, le sens le plus approprié est celui qui a le maximum de liens avec les autres concepts du même document que lui. En appliquant cette règle à tous les termes d'un document, il en résulte que les termes se désambiguïsent mutuellement et de manière globale, par rapport au contexte du document.

Dans la littérature, plusieurs méthodes et métriques existent pour la désambiguïsation des termes [Navigli, 2009]. Dans le cadre de notre recherche, nous nous sommes particulièrement intéressés à l'approche proposée dans [Baziz *et al.*, 2005]. Cette dernière est basée sur le calcul d'un score pour chaque sens associé à un concept. Ainsi, pour chaque terme $t_i$ appartenant à l'index d'un document $d$, le score de son $j^{ième}$ sens, noté $C\_Score(C_j^i)$, est calculé comme suit :

$$C\_Score(C_j^i) = \sum_{l \in [1..m],\ l \neq i} \sum_{k \in [1..n_l]} Dist(C_j^i, C_k^l) \tag{3.16}$$

Il importe de mentionner que dans l'équation (3.16), $m$ est le nombre de termes dans $Index(d)$, $n_l$ représente le nombre de sens de chaque terme $t_l$ et $Dist(C_j^i, C_k^l)$ est une mesure de proximité sémantique entre les concepts $C_j^i$ et $C_k^l$ définie initialement dans [Resnik, 1999]. Le concept qui maximise le score est alors retenu comme le meilleur sens du terme $t_i$.

Notre approche diffère de celle de Baziz *et al.* [Baziz *et al.*, 2005] dans la formule de calcul du score. En effet, nous partons de l'hypothèse que l'exploitation de la seule proximité sémantique entre concepts est insuffisante pour déterminer le meilleur sens d'un terme. Ceci est motivé, d'une part par le fait qu'elle ne tient pas compte de la représentativité sémantique des termes dans un document, et elle ne prend pas en considération le contexte local du terme dans le document, *i.e.,* la corrélation des sens des termes voisins, ainsi que dans la hiérarchie de concepts, d'autre part.

Pour cela, nous définissons la notion du *contexte local* d'un terme dans un document comme suit [Ben Ghezaiel *et al.*, 2011, Ben Ghezaiel *et al.*, 2012] :

**Définition 23** *Un contexte local d'un terme $t$ dans un document $d$, noté $Context_d(t)$, est formé par l'ensemble des termes $T(d)$ qui caractérisent le document $d$, appartenant à la même phrase que $t$.*

Dans le cadre de notre approche d'indexation conceptuelle, nous avons proposé un algorithme de désambiguïsation [Ben Ghezaiel *et al.*, 2011, Ben Ghezaiel *et al.*, 2012] qui consiste à sélectionner le sens le plus adéquat d'un concept $C_i$ en se basant sur l'unicité du sens d'un concept dans le document [Navigli, 2009] et sur les propositions suivantes [Ben Ghezaiel *et al.*, 2012] :

**Proposition 5** *Le meilleur sens pour un terme $t_i$ dans un document $d$ est corrélé au contexte local de $t_i$ dans $d$ (*cf. *Définition 23).*

**Proposition 6** *Le meilleur sens pour un terme $t_i$ dans un document $d$ est lié au contexte de chaque sens de $t_i$ dans le réseau $\mathcal{O}_{\mathcal{MGB}}$ (cf. équation (3.8), page 97).*

**Proposition 7** *Le meilleur sens pour un terme $t_i$ dans le document $d$ est corrélé avec les sens des termes fortement représentés dans $d$ (cf. équation (3.15), page 101).*

En se référant aux propositions ci-dessus, nous définissons d'abord le poids d'un concept (sens) $C_j^i \in S_i$ (*cf.* Définition 22) comme étant le poids des termes d'index reliés à $t_i$ dans le réseau proxémique conceptuel $\mathcal{O}_{\mathcal{MGB}}$. Ainsi, pour chaque terme $t_i$, le score de son $j^{ième}$ sens, noté par $C\_Score(C_j^i)$, est calculé comme suit [Ben Ghezaiel *et al.*, 2012, Latiri *et al.*, 2012a] :

$$C\_Score(C_j^i) = \sum_{C_v \in \mathcal{V}(C_j^i) \cup C_j^i} \sum_{t_l \in Context_d(t_i), l \neq i} Score_{Doc}(t_i, t_l) \times Dist(C_v, t_l) \tag{3.17}$$

sachant que :

$$Score_{Doc}(t_i, t_l) = W_{Doc}(t_i, d) \times W_{Doc}(t_l, d) \tag{3.18}$$

et

$$Dist(C_v, t_l) = \sum_{k \in [1..n_l]} Sim_{\mathcal{O}_{\mathcal{MGB}}}(C_v, C_k^l) \tag{3.19}$$

tel que $n_l$ représente le nombre de sens dans le réseau proxémique $\mathcal{O}_{\mathcal{MGB}}$ propre à chaque terme $t_l$ et $W_{Doc}(t_i, d)$, $W_{Doc}(t_l, d)$ sont les poids associés respectivement à $t_i$ et $t_l$ dans le document $d$, *i.e.,* leur représentativité globale respective dans le document $d$, estimée par l'équation (3.15) (*cf.* page 101).

Ainsi, le concept (sens) $C_i$ qui maximise le score $C\_Score(C_j^i)$ est retenu comme étant le sens le plus approprié au terme d'index $t_i$, soit :

$$C_i = \underset{j \in [1..n_i]}{\arg\max} \{C\_Score(C_j^i)\} \tag{3.20}$$

Par conséquent, en appliquant les équations (3.17), (3.18), (3.19) et (3.20), nous procédons ainsi à la désambiguïsation du concept $C_i$, qui est un concept représentatif dans le réseau conceptuel proxémique propre au document $d$.

### 3.5.3 Phase 3 : Construction du réseau proxémique d'un document *Doc*-$\mathcal{O}_{\mathcal{MGB}}$

La dernière phase de l'approche d'indexation conceptuelle que nous proposons est la construction du réseau sémantique d'un document, que nous appelons *Doc*-$\mathcal{O}_{\mathcal{MGB}}$. Les nœuds de ce réseau sont initialisés par les concepts extraits lors de l'étape de désambiguïsation, en retrouvant, pour

chaque concept, son meilleur sens. Ensuite, chaque concept de *Doc*-$\mathcal{O}_{\mathcal{MGB}}$ est décliné en plusieurs intentions et extensions grâce à la structure du réseau $\mathcal{O}_{\mathcal{MGB}}$. Ces dernières sont liées à d'autres concepts de la base générique de règles d'association $\mathcal{MGB}$.

Ainsi, autour de chaque concept de *Doc*-$\mathcal{O}_{\mathcal{MGB}}$ gravite un champ de proximité tridimensionnel, synthétisant les trois niveaux de sémantique définis dans la sous-section 3.4.3 (*cf.* page 98). Grâce à cette structure, nous passons d'un d'index simple composé de mono-termes à un index conceptuel représenté par un réseau conceptuel proxémique tridimensionnel, traduisant le contenu sémantique d'un document.

Il importe de préciser que notre approche d'indexation est indépendante de la spécificité de l'ontologie considérée et/ou de la collection de documents. Il est possible de la mettre en œuvre pour toute collection de documents dotée d'une ontologie du même domaine.

## 3.6 Évaluation de l'approche d'indexation conceptuelle

Nous proposons d'évaluer notre approche d'indexation conceptuelle à base du réseau proxémique $\mathcal{O}_{\mathcal{MGB}}$ dans le domaine la recherche d'information biomédicale. Pour ce faire, nous avons considéré comme ontologie de domaine $\mathcal{O}$, la structure poly-hiérarchique MeSH (Medical Subject Headings) [Díaz-Galiano *et al.*, 2008] [21]. L'objectif de notre évaluation expérimentale est double : (*i*) il s'agit premièrement d'étudier l'apport de l'enrichissement de l'ontologie par les nouveaux concepts issus des règles d'association de la base $\mathcal{MGB}$, découvertes à partir de la collection de documents ; (*ii*) deuxièmement, il s'agit de mesurer l'impact de l'approche proposée de désambiguïsation des sens des concepts sur la performance d'un SRI en terme des mesures du rappel et de précision.

Dans ce contexte, nous nous comparons aux travaux récents de Dinh *et al.* [Dinh and Tamine, 2011] où les auteurs ont proposé une méthode de désambiguïsation des termes utilisés dans la biomédecine et ils l'ont intégré ensuite dans un modèle d'indexation et d'interrogation sémantique.

Nous décrivons, dans ce qui suit, le cadre d'évaluation de notre approche ainsi que les résultats obtenus.

### 3.6.1 Cadre d'évaluation

#### Collection de test

Nous avons utilisé la collection OHSUMED, proposée dans le cadre de la tâche TREC9-Filtering [Robertson and Hull, 2000], qui est constituée de titres et/ou résumés de 270 journaux médicaux publiés entre 1987-1991 [Hersh *et al.*, 1994], extraits de la base MEDLINE [22]. Quelques caractéristiques statistiques de la collection OHSUMED sont données dans la TABLE 3.4. Nous avons mené les tests sur 63 requêtes, chacune étant fournie avec un ensemble de documents jugés pertinents par un groupe de médecins.

#### Medical Subject Headings (MeSH)

Nous avons utilisé l'ontologie de référence du domaine médical MeSH dans sa version 2012, qui est composé d'environ 26 582 descripteurs (MeSH Headings) répartis dans 16 catégories

21. MeSH : *Medical Subject Heading* contient le vocabulaire contrôlé de la NLM (National Library of Medicine) utilisé pour indexer les articles et faire des recherches dans les bases de données indexées par MeSH dont MEDLINE.

22. http ://www.ncbi.nlm.nih.gov/pubmed/

| | |
|---|---|
| Nombre de documents | 348 566 |
| Longueur moyenne d'un document | 100 termes |
| Nombre de requêtes | 63 |
| Longueur moyenne d'une requête | 12 termes (Titre&Description) |
| Nombre de documents jugés pertinents/requête | 50 |

TABLE 3.4 – Caractéristiques de la collection test OHSUMED.

thématiques. Chaque descripteur correspond à un concept dit *concept préféré* (main heading) pour l'indexation. Ce concept comprend un ou plusieurs termes. Dans le contexte de notre approche de désambiguïsation et d'indexation conceptuelle, nous avons également utilisé d'autres termes de MeSH appelés *qualificatifs* (Subheadings/Qualifiers). Ces qualificatifs sont souvent associés avec les descripteurs (concepts préférés) pour l'indexation.

**Phases d'évaluation**

Nous procédons à l'évaluation de notre approche d'indexation conceptuelle selon les étapes suivantes :

1. Générer la base générique minimale $\mathcal{MGB}$ à partir des documents de la collection OHSUMED.
2. Enrichir l'ontologie MeSH par les règles d'association non-redondantes de $\mathcal{MGB}$. Nous considérons ainsi les descripteurs (concepts préférés) et les qualificateurs de MeSH, sachant que la relation exprimée entre les concepts est celle de la subsomption "*is-a*", que nous estimons par la mesure de similarité de *Wu et Palmer* [Wu and Palmer, 1994] (*cf.* équation (3.4), page 96). La sélection des concepts candidats à l'enrichissement, à partir de la base $\mathcal{MGB}$, se base sur le calcul de la similarité sémantique entre un concept issu d'une hiérarchie de MeSH et son voisinage.
3. Identifier et pondérer les concepts les plus représentatifs des documents de la collection OHSUMED. En effet, en exploitant l'architecture poly-hiérarchique de l'ontologie enrichie MeSH et la mesure de similarité $Sim_{\mathcal{O}_{\mathcal{MGB}}}$, il est possible d'évaluer la représentativité globale d'un terme dans un document de la collection OHSUMED.
4. Désambiguïser les concepts retenus.
5. Enfin, construire le réseau conceptuel proxémique d'un document $Doc\text{-}\mathcal{O}_{\mathcal{MGB}}$ à partir de l'ontologie MeSH et de la base générique de règles $\mathcal{MGB}$. Ce réseau est intégré dans un processus d'indexation de documents afin d'évaluer la performance de notre méthode d'enrichissement d'ontologie et de désambiguïsation, ainsi que son impact sur la performance de l'indexation en RI.

**Mesures d'évaluation**

Nous avons utilisé les mesures de précision à P@10 et P@20 documents qui sont respectivement, la précision moyenne aux 10 et 20 premiers documents retournés et la MAP (Mean Average Precision), sur l'ensemble des 63 requêtes. Pour chaque requête, les 1 000 premiers documents sont renvoyés par le SRI expérimental et les précisions moyennes sont calculées pour mesurer la pertinence système.

**Scénarios d'évaluation**

Nous avons réalisé deux séries d'expérimentations : la première est basée sur l'indexation classique de la partie titre et résumé d'articles de OHSUMED en utilisant la configuration standard sous la plateforme TERRIER [23] et avec le schéma de pondération de référence OKAPI BM25 [Jones *et al.*, 2000], noté BM25. Cette configuration est utilisée comme base d'évaluation comparative (baseline). La seconde série d'expérimentations concerne notre méthode d'indexation conceptuelle, qui est évaluée suivant les scénarios suivants :

1. Le premier scénario décrit l'expansion de l'index d'un document moyennant les concepts de MeSH assignés manuellement par les experts humains, noté $I_{Expert}$.
2. Le second concerne l'expansion de l'index d'un document uniquement par les concepts préférés de l'ontologie MeSH, noté $I_{MeSH}$.
3. Le troisième décrit l'expansion de documents à base de termes additionnels, dérivés à partir des règles d'association valides de la base $\mathcal{MGB}$, noté $I_{\mathcal{MGB}}$. Notons que le seuil de support minimal *minsupp* ainsi que le seuil de confiance minimal $minconf$, sont respectivement de 0.05 et 0.30. Ces seuils sont définis expérimentalement en se basant sur l'étude la distribution *zipfienne* des termes dans la collection.
4. Le dernier scénario concerne l'expansion d'un document en utilisant les concepts identifiés à partir du réseau proxémique conceptuel *Doc-*$\mathcal{OMGB}$, noté $I_{\mathcal{O}_{\mathcal{MGB}}}$. Ce dernier est le résultat du processus d'enrichissement de l'ontologie MeSH par les règles d'association non redondantes de la base $\mathcal{MGB}$, dérivées à partir de la collection OHSUMED.

### 3.6.2 Résultats et discussion

Nous décrivons, dans ce qui suit, les résultats expérimentaux des différents scénarios d'indexation de documents proposés. Nous évaluons l'efficacité de la RI moyennant les concepts issus du réseau proxémique conceptuel ainsi que l'apport de la nouvelle approche de désambiguïsation.

| **Scénario** | MAP | P@10 | P@20 |
|---|---|---|---|
| Baseline (BM25) | 23.96 | 41.9 | 35.00 |
| $I_{Expert}$ | 29.55 (**+23.33**) | 45.08 (+7.59) | 39.92 (+14.06) |
| $I_{MeSH}$ | 24.73 (+3.21) | 41.27 (-1.50) | 35.87 (+2.49) |
| $I_{\mathcal{MGB}}$ | 24.96 (**+4.17**) | 42.77(+2.08) | 36.08 (+3.09) |
| $I_{\mathcal{O}_{\mathcal{MGB}}}$ | 28.17 (**+17.57**) | 44.33 (+5.80) | 38.17 (+10.86) |

TABLE 3.5 – Performance de la RI (P@10, P@20 et MAP) (Amélioration en %).

La TABLE 3.5 résume l'amélioration de la pertinence de la RI en terme de MAP, P@10 et P@20 des différents scénarios d'indexation à savoir $I_{Expert}$, $I_{MeSH}$ et les deux approches d'indexation conceptuelle à base de $\mathcal{MGB}$ et du réseau conceptuel $\mathcal{O}_{\mathcal{MGB}}$. Nous constatons que le meilleur scénario d'indexation est $I_{\mathcal{O}_{\mathcal{MGB}}}$, qui réalise une amélioration de la pertinence en terme de MAP de +17.57% tandis que le scénario d'indexation à base de $\mathcal{MGB}$ donne une amélioration moins significative (+4.17%), comparée au baseline BM25. Dans les deux cas de figure, ceci prouve l'apport en efficacité de la RI, obtenu en prenant en compte, d'un côté les

23. http ://ir.dcs.gla.ac.uk/terrier/

termes additionnels issus des règles d'association non redondantes et les concepts de l'ontologie enrichie, d'autre part.

Par ailleurs, les résultats montrent aussi que le scénario d'indexation conceptuel à base de l'ontologie MeSH (uniquement avec les concepts préférés de MeSH) aboutit à une amélioration faible de la pertinence en terme de MAP (+3.21%), comparée à celles données par les scénarios d'indexation $I_{Expert}$, $I_{\mathcal{MGB}}$ et $I_{\mathcal{O}_{\mathcal{MGB}}}$, qui dépasse nettement l'évaluation de base BM25. Bien que le gain en terme de pertinence système MAP réalisé par le scénario $I_{\mathcal{O}_{\mathcal{MGB}}}$ soit inférieur à celui donné par l'indexation $I_{Expert}$ (17.57% *vs.* 23.33%), qui représente le meilleur scénario puisqu'il intègre les annotations des experts du domaine, nous remarquons que le scénario $I_{\mathcal{O}_{\mathcal{MGB}}}$ donne une amélioration plus significative en terme de P@10 and P@20 que les autres scénarios, à savoir $I_{MeSH}$ et $I_{\mathcal{MGB}}$.

La FIGURE 3.2 illustre le gain assuré en terme d'efficacité de la RI à travers l'intégration des règles d'association de la base $\mathcal{MGB}$ et du réseau conceptuel proxémique $\mathcal{O}_{\mathcal{MGB}}$ dans le contexte de l'indexation conceptuelle de documents. Nous notons une faible amélioration de la précision moyenne à 11 points de rappel obtenue avec le scénario $I_{\mathcal{MGB}}$ par rapport au baseline BM25. Ceci est expliqué par le fait que OHSUMED est une collection médicale scientifique dont les termes ont de faibles distributions et co-occurrent marginalement ensemble. De plus, une large partie du vocabulaire est non utilisée puisqu'elle n'est pas correctement analysée par l'analyseur morpho-syntaxique, qui n'identifie pas les termes spécifiques et scientifiques de la collection OHSUMED. Cependant, nous avons remarqué, lors des expérimentations, que l'amélioration de la pertinence système est moins significative pour des seuils de support minimal *minsupp* assez élevés. En effet, l'extraction de règles d'association en considérant des *minsupp* élevés conduit à des associations entre termes qui s'avèrent triviales et qui sont très fréquentes dans la collection de documents.

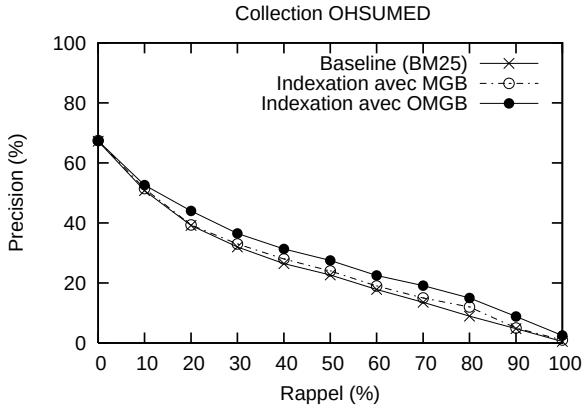


FIGURE 3.2 – Rappel/Précision (baseline *vs* index à base de $\mathcal{MGB}$ et $\mathcal{O}_{\mathcal{MGB}}$).

Afin de prouver la significativité statistique de notre approche d'indexation conceptuelle à base du réseau proxémique $\mathcal{O}_{\mathcal{MGB}}$, nous avons adopté le test de rangs signés Wilcoxon [Smucker *et al.*, 2007], appliqué au cas d'échantillons appariés entre les valeurs de MAP de notre approche et la base d'évaluation BM25. Les résultats des tests statistiques, pour un niveau de signification $\alpha = 1\%$, montrent que notre approche d'indexation à base de $\mathcal{O}_{\mathcal{MGB}}$ est statistiquement significative avec une *p-value* $< 0,01\%$ comparée à la baseline BM25.

## 3.7 Bilan des contributions

Dans ce chapitre, nous avons présenté et discuté de nos contributions dans le domaine de la RI, dont le noyau central est la base générique de règles d'association entre termes $\mathcal{MGB}$. La première partie du chapitre a mis en exergue une nouvelle approche d'expansion automatique de requêtes par cette base. L'évaluation menée sur différentes collections de test a montré le gain significatif réalisé en terme de pertinence système [Latiri *et al.*, 2012b]. La deuxième partie du chapitre a développé une idée d'enrichissement d'une ontologie de domaine par la base $\mathcal{MGB}$, identifiant des concepts additionnels à relier avec les concepts de l'ontologie originelle. Le placement de ces nouveaux concepts est réalisé à travers une nouvelle mesure de similarité ainsi que la notion de voisinage d'un concept, que nous avons définis. Le résultat du processus d'enrichissement est un réseau proxémique conceptuel qui favorise, pour chaque concept, un champ de proximité tridimensionnel, synthétisant trois niveaux de sémantique (inférentielle, différentielle et référentielle) [Ben Ghezaiel *et al.*, 2012]. Ce réseau proxémique a été intégré dans un processus de désambiguïsation et d'indexation conceptuelle des documents, dans le but de montrer l'apport d'une telle structure dans l'amélioration de l'efficacité de la tâche de la RI. Les évaluations menées montrent, en général, un gain en terme de pertinence système [Latiri *et al.*, 2012a]. Par ailleurs, ces contributions restent extensibles, dans le sens où les résultats peuvent être améliorés si nous intégrons les différentes pondérations statistiques et sémantiques proposées pour calculer le poids d'un concept dans un document, notamment dans un nouveau modèle de recherche d'information.

# Chapitre 4

# Règles d'association inter-langues pour la Traduction Automatique Statistique

**Sommaire**

## 4.1 Objectifs du chapitre

Dans la continuité de l'adaptation de l'extraction des motifs fréquents pour des applications connexes à l'ECT, nous abordons, dans ce chapitre, la problématique de l'extraction de séquences fréquentes dans le domaine de l'ECT en prenant en compte l'ordre d'apparition des mots dans une phrase, et ce à partir de corpus de textes parallèles. Notre objectif est de les exploiter dans la Traduction Automatique Statistique (TAS). Dans un premier temps, nous présentons une approche de découverte de séquences de termes fermées fréquentes (STFF) à partir d'un corpus parallèle [Latiri *et al.*, 2010a]. Nous définissons ensuite la notion de *règles d'association inter-langues* pour aboutir à la fin à un nouveau modèle de traduction à base de ces règles d'associations [Latiri *et al.*, 2010b, Latiri *et al.*, 2011].

## 4.2 Motivations

La traduction automatique à base de méthodes statistiques doit ses origines aux travaux de Brown *et al.* [Brown *et al.*, 1993]. Elle repose sur cinq composantes élémentaires, à savoir : (*i*) un corpus parallèle dont les phrases sont alignées ; (*ii*) un modèle statistique du langage ; (*iii*) le concept d'alignement de mots ; (*iv*) un modèle statistique de traduction ; et, (*v*) un algorithme de recherche de traductions appelé *décodeur*.

Au début des années 90, IBM a proposé cinq modèles qui ont contribué à dynamiser la recherche dans le domaine de la TAS. Depuis l'apparition de l'approche pionnière d'IBM [Brown *et al.*, 1993], la majorité des travaux proposés dans la TAS ont été fondés sur ces modèles. Ceci est étroitement lié à la force de l'approche et à la disponibilité des outils de TAS, tels que Giza++ pour la génération de la table de traduction [Och and Ney, 2000], Cmu [Rosenfeld, 1995] et Srilm [Stolcke, 2002] pour le développement du modèle de langage, Pharao [Koehn, 2004] et Moses [Koehn and Hoang, 2007] pour le décodage.

Notre contribution dans la TAS émane de notre conviction qu'il est possible d'étudier d'autres approches issues du domaine de l'ECT, qui soient capables de constituer une alternative aux méthodes proposées par IBM et d'améliorer la pertinence des modèles de traduction à base de séquences [Koehn *et al.*, 2007]. Notre objectif est d'étendre les approches d'extraction des motifs fréquents pour supporter des corpus parallèles de très grande taille et d'extraire des connaissances utiles et pertinentes pour la TAS.

L'originalité de notre proposition repose ainsi sur le couplage de deux types de motifs fréquents, à savoir les règles d'association entre termes [Agrawal and Skirant, 1994, Balcázar, 2010] et les séquences fermées fréquentes de termes [Chang, 2004] (*cf.* chapitre 1, section 1.5, page 60). Ce couplage donne naissance à la définition d'un nouveau concept que nous appelons *Règles d'Association Inter-Langues* (RAIL). Intuitivement, l'interprétation d'une règle d'association inter-langues dans le domaine de la TAS est que sa *conclusion*, représentant une unité linguistique dans une langue cible, est une traduction potentielle de sa *prémisse*, représentant une autre unité linguistique dans la langue source.

En effet, le principal avantage du modèle de traduction, à base de règles d'association inter-langues, est qu'il ne nécessite pas d'alignement [Marcu and Wong, 2002, Lavecchia *et al.*, 2008] contrairement à ce qui se fait généralement dans la communauté de la TAS, où il existe plusieurs travaux qui se sont intéressés à l'optimisation de cette étape d'alignement [Och and Ney, 2000, Singh and Husain, 2005, Riesa *et al.*, 2011].

Nous présentons dans la section suivante un bref aperçu sur la TAS, en rappelant des concepts clés que nous reprenons dans le cadre de notre contribution.

## 4.3 Autour de la Traduction Automatique Statistique

La TAS repose sur l'utilisation de corpus parallèles alignés. De tels corpus sont obtenus en alignant chaque segment d'un corpus source avec sa traduction dans le corpus cible. L'approche statistique consiste alors à dériver, par des méthodes statistiques, un modèle de traduction à partir de récurrences d'événements dans les corpus alignés. Ce modèle permettra ensuite de calculer la traduction la plus probable d'une phrase source.

Considérons une paire de langues $(\mathbf{F}, \mathbf{E})$. Afin de traduire une phrase source $f \in \mathbf{F}$, nous supposons que toutes les phrases cibles $e \in \mathbf{E}$ sont des traductions possibles de $f$. Le processus de traduction est vu ainsi comme un problème d'optimisation. Nous notons par $P(e|f)$, la probabilité conditionnelle pour que la phrase cible $e$ soit une traduction de la phrase source $f$ [Brown *et al.*, 1993]. Nous cherchons donc $\hat{e}$, une des phrases cibles, qui maximise $P(e|f)$. En utilisant la formule de Bayes, nous avons l'équation fondamentale de la traduction automatique statistique, soit :

$$\hat{e} = \underset{e}{argmax}\ P(e|f) = \underset{e}{argmax}\ P(f|e)P(e) \tag{4.1}$$

où $P(e)$ est estimée par *un modèle de langage* dont les paramètres sont appris sur un corpus d'apprentissage de la langue source $\mathbf{F}$ et $P(f|e)$ par *un modèle de traduction* dont les paramètres sont appris sur un corpus d'apprentissage parallèle aligné.

Ainsi, l'équation (4.1) fait appel à *un modèle de langage*, un *modèle d'alignement* caché et un *modèle de traduction*, que nous définissons dans ce qui suit.

### 4.3.1 Modèle de langage

Le modèle de langage est illustré par $P(e)$ dans l'équation (4.1). Il permet de déterminer la vraisemblance des hypothèses de traduction produites dans la langue cible. Il intervient directement dans le calcul effectué par le décodeur et attribut un score à chaque hypothèse. Ces scores sont ensuite combinés par le décodeur afin de maximiser une fonction objective pour trouver la meilleure hypothèse parmi toutes celles qui sont possibles [Haton *et al.*, 2006]. La plupart des systèmes de TAS utilisent un modèle de langage de type $n$-grammes d'ordre[24], où $n$ peut atteindre 5 [Brants *et al.*, 2007].

### 4.3.2 Alignement de $n$-grammes

L'utilisation du modèle du langage uniquement est insuffisante. La quasi-totalité des modèles de TAS introduisent une variable cachée $A(e, f)$, appelée *alignement* [Brown *et al.*, 1993], qui décrit les correspondances existantes entre les mots de la phrase source $f$ et ceux de la phrase cible $e$. Les alignements de groupes de mots à d'autres groupes de mots sont autorisés, *i.e.,* que l'alignement n'est pas une bijection entre les mots d'une phrase source $f$ et d'une phrase cible $e$. De même, l'alignement à un mot spécial appelé NULL est utilisé lorsqu'un ou plusieurs mots d'une phrase source $f$ n'ont pas de correspondance dans la phrase cible $e$. Le concept d'alignement de $n$-grammes est introduit implicitement dans l'équation (4.1).

### 4.3.3 Modèle de traduction à base de mots

Ce modèle permet de calculer la probabilité $P(f|e)$ dans l'équation (4.1) à l'aide de différents paramètres. Ces derniers sont estimés sur un corpus d'apprentissage parallèle composé de

24. Un n-grammes est une suite de $n$ mots.

paires de phrases, où chaque paire représente la phrase source et sa traduction dans la langue cible. Les cinq modèles de traduction d'IBM à base de mots sont dotés d'un algorithme pour leur apprentissage [Brown *et al.*, 1993]. Autrement dit, l'unité de traduction qui apparaît dans l'expression $P(f|e)$ est le mot. La performance et la complexité des modèles augmentent du premier au cinquième. Ils représentent ainsi des modèles de TAS mot-à-mot.

Dans la suite de cette section, la phrase source est notée $f = f_1 \dots f_I$ et la phrase cible par $e = e_1 \dots e_J$. Les modèles d'IBM [Brown *et al.*, 1993] proposent chacun une expression de la quantité $P(e|f)$.

En considérant l'ensemble des alignements possibles, la vraisemblance d'une traduction $(e, f)$ est définie par :

$$P(e|f) = \sum_a P(e, a|f) \tag{4.2}$$

où $a$ est un alignement possible entre $e$ et $f$. La somme s'effectue sur tous les alignements possibles. Les cinq modèles d'IBM procèdent de façon itérative pour estimer la probabilité $P(e|f)$ où chaque modèle s'appuie sur les paramètres estimés par le modèle qui le précède. Nous décrivons brièvement dans ce qui suit les cinq modèles d'IBM. Une étude plus détaillée de ces modèles est donnée dans [Lavecchia, 2010].

- **Le modèle IBM-1 :** C'est le plus simple des cinq. Il considère que tous les mots de la phrase source peuvent être alignés à tous les mots de la phrase cible avec la même probabilité. Ainsi, l'ordre des mots n'affecte pas le calcul de $P(f|e)$. Ce modèle repose donc sur une table de traduction de mots dans laquelle sera stockée les probabilités de traduction entre chaque mot source et chaque mot cible de la forme $t(e_j|f_i)$, *i.e.,* la probabilité de traduction de $f_i$ en $e_j$.
- **Le modèle IBM-2 :** Ce modèle suppose que la probabilité d'un alignement dépend de la position des mots dans la phrase. Il impose des restrictions sur l'alignement $A(e, f)$ entre les mots des phrases source $f$ et cible $e$. Ainsi, le modèle de l'alignement est de la forme $A = a_1 \dots a_j$, où pour tout $j \in [1, J]$, nous avons l'alignement $a_j \in [0, I]$. Le cas où $a_j = i > 0$ signifie que le mot cible $e_j$ est aligné au mot source $f_i$, tandis que le cas de figure où $a_j = 0$ signifie que le mot $e_j$ n'est pas aligné, ou qu'il est aligné au mot NULL $s_0$. De ce fait, un alignement de cette forme autorise l'alignement de plusieurs $e_j$ à un seul $f_i$, mais pas l'inverse, *i.e.,* un mot cible $e_j$ est aligné à zéro ou un mot source. Le modèle IBM-2 est donc asymétrique et dispose d'une loi d'alignement dite aussi *de distorsion* de la forme $P(a_j|j)$.
  Ces deux premiers modèles d'IBM donnent souvent des alignements de mots peu satisfaisants mais ils permettent l'entraînement de modèles plus complexes.
- **Les Modèles IBM-3, IBM-4 et IBM-5 :** Pour ces modèles, le processus de génération de $e$ et de l'alignement $a$ se déroule en trois étapes, à savoir :

  1. La première étape consiste à choisir, pour chaque mot source $f_i$ de la phrase source $f$, le nombre de mots qui lui seront connectés dans la phrase cible $e$. Ceci introduit la notion de *fertilité*, *i.e.,* le nombre de mots dans $e$ générés par un mot $f_i$. Notons que dans les modèles IBM-1 et IBM-2, la fertilité d'un mot est obtenue de manière implicite étant donné que l'alignement permet de lier plusieurs $e_j$ à un même $f_i$.
  2. La deuxième étape identifie le ou les mots $e_j$ connectés à chaque mot $f_i$ à l'aide des probabilités $t(e_j|f_i)$ de la table de traduction de mots, tel que $t(e_j|f_i)$ exprime la probabilité de traduction de $f_i$ en $e_j$.
  3. La troisième étape attribue aux mots identifiés dans $f$ une position. Elle fait intervenir un modèle de distorsion qui estime la probabilité pour que la position $j$ dans la phrase

$f$ soit connectée à la position $a_j$ dans la phrase $e$.

Notons que l'entraînement des paramètres des modèles d'IBM se fait par l'outil GIZA++ [Och and Ney, 2000]. Il est réalisé à partir d'un corpus parallèle, aligné par phrase. Aucune information a priori n'est nécessaire, pas même un lexique. GIZA++ entraîne successivement des modèles de complexité croissante, comme proposé par Brown *et al.* [Brown *et al.*, 1993].

### 4.3.4 Processus de décodage

Pour réaliser un processus complet de TAS, en plus du modèle de langage, du modèle d'alignement et du modèle de traduction, il est indispensable de faire appel à un décodeur. Ce dernier utilise tous ces modèles afin de trouver, dans l'espace de recherche, les meilleures hypothèses de traduction. D'une manière simplifiée, le décodeur utilise quatre aspects de la traduction, à savoir : (*i*) le *modèle de langage* $P(e)$ ; (*ii*) le *modèle de traduction* $P(f|e)$ ; (*iii*) le *modèle de distorsion* $d(f, e)$ ; et, (*iv)* un modèle de longueur de traduction, appelé *modèle de brièveté*, noté $\omega$.

Ainsi, le décodeur est chargé de produire la ou les $n$ meilleures traductions possibles, à l'aide des paramètres estimés par ces modèles impliqués dans le processus de traduction. Les décodeurs les plus connus sont PHARAOH [Koehn, 2004] et MOSES [Koehn *et al.*, 2007].

### 4.3.5 Évaluation d'un système de traduction

Plusieurs mesures automatiques existent dans la littérature pour évaluer un système de traduction automatique, telles que : WER, NIST, BLEU, etc. [Do, 2011]. Ces mesures utilisent comme paramètre principal la proximité de la traduction automatique par rapport à une ou plusieurs traductions humaines, dites *traductions de référence*. La mesure qui a été la plus utilisée par la communauté de la TAS est le score BLEU (BiLingual Evaluation Understudy), proposé par Papineni et *al.* en 2001 [Papineni *et al.*, 2001]. Le principe de cette mesure est de calculer le degré de similitude entre une traduction automatique et une ou plusieurs traductions de référence, en se basant notamment sur la précision $n$-grammes. De manière simple, il s'agit de compter le nombre d'unités linguistiques d'une phrase à évaluer, contenues dans une ou plusieurs phrases de référence. Une traduction est d'autant meilleure qu'elle partage un grand nombre de $n$-grammes avec une ou plusieurs des traductions de référence. Le score BLEU est défini comme suit [Papineni *et al.*, 2001] :

$$BLEU = BP \times e^{\sum_{n=1}^{N} w_n \log_2 (P_n)} \tag{4.3}$$

La mesure BLEU calcule ainsi la moyenne géométrique des précisions $n$-grammes $P_n$, obtenue avec des $n$-grammes d'ordre 1 jusqu'à $N$ et des poids $w_n$ positifs. $P_n$ est le nombre de $n$-grammes de la traduction automatique présents également dans une ou plusieurs traduction de référence, divisé par le nombre de $n$-grammes total de la traduction automatique. $BP$ est une pénalité de brièveté, calculée pour défavoriser les traductions automatiques courtes par rapport aux références.

Les précisions $n$-grammes sont généralement combinées jusqu'au 4-grammes avec des poids $w_n$ uniformes. Une traduction automatique se voit attribuer un score BLEU de 1 lorsqu'elle est identique à une traduction de référence. Autrement, elle aura un score de 0 si aucun de ses $n$-grammes n'est présent dans une référence.

### 4.3.6 Corpus parallèles

Un corpus parallèle est une très large base d'exemples de traductions qui permet à un système de TAS de construire le modèle statistique de traduction. Schématiquement, un corpus est un ensemble de paires de phrases $\{(f, e)\}$, où $f$ est une phrase de la langue source et $e$ une phrase de la langue cible considérée comme une traduction de $f$. À titre d'exemple, le corpus parallèle bilingue EUROPARL [25] [Rama and Borin, 2011] provient des actes du Parlement Européen entre Mars 1996 et Septembre 2009. Il faut préciser que l'ensemble du corpus est aligné phrase par phrase, *i.e.,* pour chaque phrase d'une langue source, nous avons une traduction dans la langue cible proposée par des experts humains. En effet, cet ensemble de phrases dites parallèles est indispensable pour déduire une distribution de probabilités pour le modèle de traduction $P(f|e)$. Les caractéristiques du corpus que nous utiliserons sont données dans la TABLE 4.1.

| | | **Français** | **Anglais** |
|---|---|---|---|
| **Apprentissage** | Phrases | $596 \times 10^3$ | |
| | Mots | $17,3 \times 10^6$ | $15,8 \times 10^6$ |
| | Singletons | $26,6 \times 10^3$ | $22,2 \times 10^3$ |
| | Vocabulaire | $77,5 \times 10^3$ | $60,3 \times 10^3$ |
| **Développement** | Phrases | 1444 | |
| | Mots | $15,0 \times 10^3$ | $14,0 \times 10^3$ |
| **Test** | Phrases | 500 | |
| | Mots | $5,2 \times 10^3$ | $4,9 \times 10^3$ |

TABLE 4.1 – Caractéristiques du corpus EUROPARL.

### 4.3.7 Modèles de traduction à base de séquences de mots

Les modèles d'IBM à base de mots décrits précédemment utilisent *le mot* comme unité de traduction. Ces modèles ne permettent pas d'avoir une bonne traduction que par une technique de traduction "mot-à-mot", puisqu'ils autorisent le fait qu'un mot de la langue source se traduit par plusieurs mots dans la langue cible ou encore que les mots soient réordonnés. À titre d'exemple, c'est le processus de réordonnement qui va produire la traduction correcte de *"Eiffel Tower"* en *"Tour Eiffel"*.

De plus, le décodeur qui réalise la traduction utilise toujours un modèle de langage. De ce fait, le choix de la traduction de chaque mot est étroitement lié à son impact sur l'ensemble de la phrase produite. Un autre problème des modèles à base de mots est l'asymétrie des alignements qu'ils imposent. Rappelons en effet, que plusieurs mots cibles peuvent être alignés à un même mot source mais que l'inverse n'est pas possible suivant les modèles proposés dans [Brown *et al.*, 1993]. À titre d'exemple, la traduction du mot "*sortira*" posera un problème, puisqu'il se traduit en anglais par "*will go out*" et que les modèles à base de mots ne permettent pas de traiter ce cas de figure.

Ces limitations des modèles à base de mots ont donné naissance, à partir de 1999, à plusieurs travaux en TAS qui considèrent comme unité de traduction *une séquence de mots* [Zens *et al.*, 2002, Koehn *et al.*, 2003, Hoang *et al.*, 2009, Hayashi *et al.*, 2010]. Ces modèles à base de séquences de mots permettent de modéliser le fait que plusieurs mots peuvent être alignés à plusieurs autres mots. Ils viennent ainsi corriger les limites des modèles à base de mots.

25. Disponible gratuitement sur le site : www.statmt.org/europarl/

En TAS, une séquence de mots regroupant $l$ mots, avec $l \geq 1$, est notée comme suit : $\tilde{f} = f_i, \ldots, f_{(i+l-1)}$. Le principe commun aux modèles de TAS à base de séquences de mots est décrit selon les étapes suivantes [Koehn *et al.*, 2003] :

1. La phrase source $f$ est d'abord segmentée en $K$ séquences, *i.e.*, $f = \tilde{f}_1 \ldots \tilde{f}_K$.
2. Chaque séquence de mots $\tilde{f}$ de la phrase source est ensuite traduite en une séquence de mots cible $\tilde{e}$.
3. Ces séquences de mots sont éventuellement réordonnées selon une permutation $\rho(.)$ de $[1..K]$, elles sont ensuite simplement accolées pour constituer la phrase cible finale $e = \tilde{e}_{\rho(1)} \ldots \tilde{e}_{\rho(K)}$.

En effet, utiliser des séquences de mots comme unité de traduction permet d'aligner $n$ mots source à $m$ mots cible et éviter ainsi les alignements parfois peu satisfaisants imposés par les modèles à base de mots. Ils sont en mesure de traduire directement, par exemple, des groupes nominaux ou des unités lexicales et syntaxiques observées sur le corpus d'apprentissage et ainsi parvenir à préserver certaines contraintes locales sur l'ordre des mots [Koehn *et al.*, 2003, Koehn, 2004].

Le modèle à patrons d'alignement [Och and Ney, 2000, Och and Ney, 2004] fut le premier modèle de traduction à base de séquences de mots. Depuis, il a inspiré de nombreux travaux qui ne diffèrent que sur les détails de leur apprentissage et sur leur modélisation des réordonnements. Par ailleurs, Koehn *et al.* ont publié les articles fondateurs de la traduction à base de séquences de mots [Koehn *et al.*, 2003, Koehn, 2004]. D'autres travaux ont abordé des modèles différents de la TAS à base de groupes de mots, tels que le modèle de traduction hiérarchique [Hayashi *et al.*, 2010] et le modèle statistique syntaxique [Zhang *et al.*, 2009]. Compte tenu du fait que les modèles de TAS hiérarchiques, syntaxiques et ceux à base de séquences, sont similaires dans leur phase d'apprentissage, Hoang *et al.* ont proposé dans [Hoang *et al.*, 2009] une extension du décodeur Moses vers une plate-forme unifiée capable de mettre en œuvre ces trois modèles populaires dans leurs différentes phases du processus de traduction.

Une autre approche originale basée sur le concept de *triggers inter-langues* a été proposée dans [Lavecchia *et al.*, 2007, Lavecchia *et al.*, 2008] qui ne nécessite aucun alignement de mots au sein des corpus parallèles. Les triggers inter-langues permettent de mettre en évidence des unités fortement corrélés en se basant sur l'Information Mutuelle (IM). L'idée derrière ce concept est que si une séquence de mots source est fortement corrélée à une séquence de mots cible en terme d'IM, alors la présence de la première dans une phrase source déclenchera la présence de la seconde dans sa traduction et vice versa. Les triggers inter-langues sont extraits à partir d'un corpus d'apprentissage parallèle dans le but de trouver les traductions possibles de séquences de mots et constituer ainsi un nouveau modèle de traduction à base de ces triggers.

Pour conclure, étant donné un corpus d'apprentissage parallèle, le processus de traduction automatique à base de séquences de mots peut être résumé en deux principales phases, à savoir [Duan *et al.*, 2010] :

1. **Phase d'extraction de séquences :** Cette phase permet de découvrir, à partir d'un corpus parallèle, toutes les paires de traductions pertinentes qui sont conformes avec les contraintes du modèle de traduction choisi.
2. **Phase de paramétrage :** Durant cette phase, une série de probabilités est attribuée à chaque paire de traduction. Ces différentes probabilités sont indispensables pour le décodeur afin de choisir la meilleure paire de traduction.

Notre contribution dans la TAS s'adresse aux deux phases du processus de traduction, citées ci dessus. Nous proposons d'extraire en premier lieu des séquences de mots fréquentes à partir d'un corpus parallèle par une approche de fouille de séquences issue du domaine du data mining [Chang, 2004]. Ensuite, la paire de traduction sera décrite par un nouveau paradigme que nous appelons *Règles d'Association Inter-langues* [Latiri *et al.*, 2010b, Latiri *et al.*, 2011] et dont la mesure de confiance sera utilisée pour dériver la probabilité de la paire de traduction.

### 4.3.8 Vers un modèle de traduction à base de règles d'association inter-langues

Pour un système de traduction à base de séquences, le modèle de traduction est la source de connaissance principale qui établit une correspondance entre les deux langues source **F** et cible **E**. Son rôle, pour chaque phrase source, est de guider la construction d'un ensemble d'hypothèses de traduction en langue cible. L'unité de traduction est la séquence, qui correspond à une suite de mots contigus. Ainsi, l'association entre une séquence source et une traduction possible en langue cible forme une paire de traductions. Notons qu'il est possible qu'une séquence source admette plusieurs traductions alternatives, donnant lieu à plusieurs paires de traductions partageant la même séquence source. Afin de faire un bon usage de ces paires de traductions, il est nécessaire de leur associer des mesures, comme par exemple statistiques, pour mesurer la confiance en l'association ainsi réalisée [Zens *et al.*, 2002].

En se référant aux travaux de Lavecchia *et al.* [Lavecchia *et al.*, 2007, Lavecchia *et al.*, 2008] qui ont introduit le concept des *triggers inter-langues* en TAS, l'idée de notre contribution a été de repenser le problème moyennant des techniques d'extraction de motifs fréquents issues du domaine de l'ECT. La finalité de notre approche est de pouvoir estimer les probabilités de traduction entre les paires d'unités linguistiques sources et cibles, sans avoir recours à l'alignement qui demeure une tâche fastidieuse faisant appel à des algorithmes complexes [Ortiz-Martinez, 2011].

Pour ce faire, en se positionnant par rapport aux modèles de TAS à base de séquences, nous considérons que la paire de traduction exprimée et modélisée différemment dans chacun des modèles de l'état de l'art, peut être formalisée par un nouveau motif dit *Règle d'Association Inter-langues* [Latiri *et al.*, 2010b], tel que la prémisse de la règle est exprimée dans la langue source **F** alors que sa conclusion l'est dans la langue cible **E**. De ce fait, ces règles d'association seront découvertes à partir d'un corpus parallèle par un parcours de l'espace de recherche qui s'effectue au niveau de la phrase. Une interprétation intuitive d'une règle inter-langues est que la conclusion de cette dernière est une traduction potentielle de sa prémisse [Latiri *et al.*, 2010b].

L'approche proposée se déroule en deux étapes : (*i*) extraire les séquences de termes à partir des corpus source et cible du corpus parallèle et ce moyennant une approche de fouille de séquences ; et, (*ii*) dériver les règles d'association valides entre ces séquences par rapport à un seuil de confiance minimal $minconf$.

En effet, la confiance d'une règle d'association est considérée comme une mesure de co-occurrence qui se calcule simplement. Pour dériver les règles d'association inter-langues, nous supposons que deux séquences source et cible co-occurrent si elles apparaissent dans une même paire de phrases du corpus parallèle. De ce fait, notre méthode ne requiert qu'un alignement au niveau des phrases et non au niveau des mots au sein du corpus parallèle. Les valeurs de confiance des règles inter-langues valides permettent de calculer directement les probabilités de traduction. Ainsi, l'utilisation de règles d'association inter-langues, pour estimer une table de traduction, rend notre approche moins complexe mais tout aussi efficace que les approches existantes. Nous la comparons avec la méthode de l'état de l'art [Koehn *et al.*, 2003] ainsi qu'avec celle à base

des triggers inter-langues [Lavecchia *et al.*, 2008].

Il importe de noter que l'originalité de notre contribution tient dans le couplage de deux techniques d'extraction de motifs fréquents en ECT, à savoir la fouille de séquences fréquentes et la découverte de règles d'association. Notre objectif est de définir un nouveau modèle de TAS à base de séquence de mots.

Nous décrivons dans la section qui suit notre approche d'extraction des séquences fermées fréquentes à partir d'un corpus parallèle pour la TAS.

## 4.4 Extraction des séquences fermées fréquentes à partir d'un corpus parallèle

Dans le chapitre 1 (*cf.* section 1.5, page 60), nous avons introduit le cadre formel et les approches existantes de fouille de séquences. Nous avons aussi motivé notre intérêt pour l'extraction *des motifs séquentiels fermés fréquents* [Yan *et al.*, 2003, Wang and Han, 2004, Chang, 2004] dans le cadre de l'ECT.

En effet, nous avons opté pour adapter à notre contexte d'extraction textuel $\mathfrak{M} = (\mathcal{C}, \mathcal{T}, \mathcal{I})$, l'algorithme d'extraction de séquences fermées fréquentes BFSM, proposé dans [Chang, 2004] et qui est basé sur une méthode de recherche "*en largeur d'abord*". L'algorithme BFSM tient compte de l'ordre d'apparition des termes dans une séquence et introduit la notion de *l'information de position.*

### 4.4.1 Notre approche pour la TAS

L'idée clé de notre approche d'extraction de séquences de termes à partir d'un corpus parallèle aligné au niveau de la phrase est l'utilisation du principe de *l'extension de séquence* [Chang, 2004]. En effet, pour étendre une $k$-séquence $S_k$ à une $(k+1)$-séquence $S_{k+1}$, nous procédons à une extension de séquence qui consiste à lui ajouter un terme comme étant un nouvel élément de la séquence. Ainsi, la séquence $S_{k+1}$ est le résultat de la jointure de la séquence $S_k$ avec une séquence $S_\alpha$ appartenant à la liste des 2-séquences comme le décrit la procédure EXTENSION-SÉQUENCE illustrée dans l'Algorithme 2.

**Procédure** Extension-séquence
**Entrée:** une $k$-séquence $S_k$ et une 2-séquence $S_\alpha$
**Sortie:** $(k+1)$-séquence
**Si** $(dernier\text{-}terme(S_k) = premier\text{-}terme(S_\alpha)) \wedge id\_ph(S_k) = id\_ph(S_\alpha) \wedge (pos\_seq(S_\alpha) = pos\_seq(S_k)+1)$ **alors**
$S_{k+1} = \langle (S_k \oplus (S_\alpha \setminus premier\text{-}terme(S_\alpha))) \rangle \wedge |\{id\_ph_{S_{k+1}}\}| \geq minsupp.$
**Fin Si**

**Algorithme 2:** PROCÉDURE EXTENSION-SÉQUENCE.

En considérant un contexte d'extraction textuel $\mathfrak{M} = (\mathcal{C}, \mathcal{T}, \mathcal{I})$ et en tenant compte de l'ordre d'apparition d'un terme dans une phrase, notre approche de génération de l'ensemble des séquences de termes fermées fréquentes, noté par $\mathcal{STFF}$, se déroule en quatre étapes inspirées de l'algorithme BFSM [Chang, 2004], à savoir :

1. L'extraction des 1-séquences fréquentes.
2. La génération des 2-séquences fréquentes.
3. La génération des séquences fréquentes de termes de taille supérieure à 2.

4. L'élagage de l'espace de recherche pour ne retenir que les séquences fermées fréquentes.

Ces différentes étapes sont décrites ci-dessous.

**Étape 1 : Extraction des 1-séquences fréquentes (les termes)** Lors de la première étape, l'algorithme parcourt le corpus une seule fois pour extraire les termes, *i.e.,* les 1-séquences, et enregistre leurs informations de position. Nous rappelons que l'information de position d'une séquence $S$, notée par $\mathcal{POS}_S$, se compose d'un ensemble de paires de $(id\_ph, pos\_seq)$, tel que le *id_ph* est le numéro de la phrase dans le corpus où apparaît $S$ et *pos_seq* est la position de $S$ dans la phrase (*cf.* chapitre 1, section 1.5, page 60).

Les approches d'extraction de séquences basées sur l'algorithme Apriori sont très consommatrices en accès disques. En effet, chaque phase de comptage, aussi optimisée soit-elle, déclenche une lecture de toute la base de séquences. Une solution consiste alors à stocker les informations relatives aux séquences en mémoire vive, sous la forme de *listes d'occurrences.* Dans ce cas, la génération des candidats n'est qu'une opération de jointure de listes. Ainsi, le comptage de séquences fréquentes consiste à comparer la cardinalité des jointures effectuées avec la valeur *minsupp.*

**Étape 2 : L'extraction des 2-séquences fréquentes (les bigrammes)** La génération des 2-séquences se fait par jointure des 1-séquences fréquentes, sans avoir à parcourir le corpus pour compter leurs supports, puisque leurs informations de positions peuvent être obtenues à partir de celles des 1-séquences fréquentes. Pour fusionner les données de positions $\mathcal{POS}_i$ et $\mathcal{POS}_j$ de deux 1-séquences $S_i$ et $S_j$, l'algorithme procède à une *extension de séquence* comme le décrit l'Algorithme 2.

**Étape 3 : La génération des séquences fréquentes de taille supérieure à 2 (les $n$-grammes)** En se basant sur le principe de l'algorithme Gsp [Srikant and Agrawal, 1996], la génération des $(k+1)$-séquences se fait en étendant les $k$-séquences par les 2-séquences fréquentes, dont le premier terme est le même que le dernier terme de la $k$-séquence. Nous procédons également à une extension de séquences telle qu'elle est décrite dans l'Algorithme 2. Le principe de la jointure est identique à celui utilisé pour la génération des 2-séquences fréquentes. En considérant une approche en "*largeur d'abord*", nous utilisons des 2-séquences fréquentes pour générer des séquences fréquentes de longueur supérieure à 2.

**Étape 4 : L'élagage de l'espace de recherche pour dériver les séquences de termes fermées fréquentes** Pour obtenir un ensemble compact de séquences de termes fermées fréquentes et éliminer la redondance, notre approche fait appel à une technique d'élagage de l'espace de recherche, utilisée dans l'approche originelle Bfsm de [Chang, 2004]. Cette technique est basée sur la condition de la *super-séquence arrière* illustrée par la Proposition 8 ci dessous, qui est utilisée à chaque itération après avoir trouvé les séquences fréquentes. Si une séquence $S_{(k+n)}$ est une super-séquence arrière d'une séquence $S_k$, alors cette dernière est considérée comme redondante et elle est élaguée de l'ensemble des $\mathcal{STFF}$.

**Proposition 8** *Soient deux séquences $S_k$ et $S_{(k+n)}$. Si $S_{(k+n)}$ est une super-séquence de $S_k$ et elles ont les mêmes $k$ premiers termes et le même support, alors $S_{(k+n)}$ est dite une* ***super-séquence arrière*** *de $S_k$ ; autrement dit la séquence $S_k$ est redondante par rapport à $S_{(k+n)}$* [Chang, 2004].

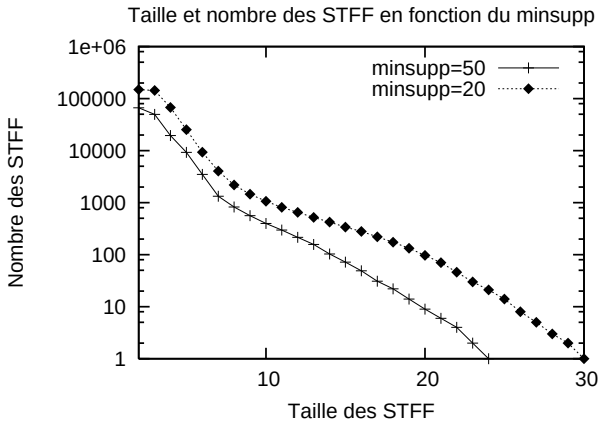


FIGURE 4.1 – Variation du nombre de séquences de termes fermées fréquentes (STFFs) en fonction de leurs tailles respectives pour deux valeurs de *minsupp*.

### 4.4.2 Évaluation empirique de l'extraction des séquences de termes fermées fréquentes

Nous avons mené une évaluation sur le corpus parallèle d'apprentissage d'EUROPARL décrit dans la TABLE 4.1. Ce type de corpus représente un vrai défi pour la communauté de l'ECT. À cet égard, il est indispensable d'adapter les algorithmes d'extraction de motifs fréquents aux contextes d'extraction denses [Latiri *et al.*, 2012b]. En effet, lors de nos différentes expérimentations, nous avons été confrontés à ce problème où l'objectif a été de dépasser les limites qui dépendent principalement du temps de traitement et de la complexité du calcul.

Nous avons mis en place plusieurs scénarios d'évaluation empirique afin de trouver le meilleur seuil de support minimal *minsupp* permettant d'extraire un ensemble de séquences de termes non-redondantes, *i.e.,* des séquences fermées fréquentes, contenant le minimum de bruit. Par bruit, nous entendons les séquences marginales dans le corpus qui dégradent en conséquence la qualité de la traduction. Notons que, outre la mesure du support, aucun autre filtrage de type linguistique ou sémantique n'a été effectué sur l'ensemble des séquences.

La FIGURE 4.1 illustre le nombre de séquences de termes fermées fréquentes découvertes en fonction de leurs tailles, pour des valeurs *minsupp* de 20 et de 50 phrases sur le corpus parallèle d'apprentissage EUROPARL. Nous constatons que plus le support minimum *minsupp* décroît, plus le nombre de séquences de termes fermées fréquentes augmente. Ceci est expliqué par le fait qu'un faible seuil de support permet de retenir, lors de l'exploration des séquences, celles qui sont trop fréquentes dans le corpus, et n'exclut pas par ailleurs celles qui sont rares. Nous avons également constaté que la longueur maximale des séquences de termes fermées fréquentes (STFFs) augmente dès qu'on diminue la valeur de *minsupp*. Ceci est justifié par le nombre élevé des séquences fréquentes qui sont générées en amont et qui sont candidates à la jointure et à l'extension des $k$-séquences.

Il importe de signaler que les méthodes statistiques de traduction automatique à base de séquences de mots ne retiennent qu'au maximum les 5-grammes [Och and Ney, 2004]. Dans notre approche, nous avons retenu toutes les séquences quelque soit leur taille afin de comparer notre méthode avec celles de l'état de l'art.

## 4.5 Règles d'association inter-langues

À l'issue de la phase d'extraction de séquences de termes fermées fréquentes, nous proposons de les utiliser dans la traduction automatique à base de séquences, et ce en définissant la notion de *règles d'association inter-langues*, qui représente notre nouvelle formalisation d'une paire de traduction source et cible [Latiri *et al.*, 2010a, Latiri *et al.*, 2010b].

### 4.5.1 Définition d'une règle d'association inter-langues

Le concept de *règle d'association inter-langues* est une extension de la règle d'association telle qu'elle a été définie dans [Agrawal and Skirant, 1994], pour le domaine de la TAS.

**Définition 24** *Une règle d'association inter-langues, notée par RAIL, est une implication de la forme : $R : S_f \Rightarrow S_e$ telles que $S_f$ et $S_e$ sont deux séquences de termes fermées fréquentes en langue source et cible, de tailles respectives n et m mots* [Latiri *et al.*, 2010b].

En TAS, cette règle d'association signifie que la séquence en langue cible $S_e$ est une traduction candidate de la séquence en langue source $S_f$. Une règle d'association inter-langues est également appréciée par les deux métriques de *support* et de *confiance* [Agrawal and Skirant, 1994] définies dans le chapitre 1 (*cf.* section 1.4, page 58).

Le support de la règle d'association $R : S_f \Rightarrow S_e$ exprime, dans notre contexte de recherche, la fréquence avec laquelle deux séquences $S_f$ et $S_e$ co-occurrent ensemble dans le corpus parallèle. Autrement dit, la cardinalité de l'ensemble de phrases du corpus contenant en même temps les deux séquences $S_f$ et $S_e$. La confiance de $R$ exprime la probabilité conditionnelle pour qu'une phrase contienne la séquence $S_e$, sachant qu'elle contient la séquence $S_f$.

Une règle d'association inter-langues est dite *valide* si sa confiance est supérieure ou égale au seuil minimal de confiance *minconf*.

Nous décrivons, dans ce qui suit, l'approche qui permet de dériver les règles d'association inter-langues à partir de l'ensemble $\mathcal{STFF}$ [Latiri *et al.*, 2010a, Latiri *et al.*, 2011]

### 4.5.2 Dérivation de règles d'association inter-langues

En considérant une paire de langues source **F** et cible **E**, nous avons mis en place un algorithme qui considère en entrée les deux ensembles de séquences fermées fréquentes groupées par taille en langue source $\mathcal{STFF}_f$ et en langue cible $\mathcal{STFF}_e$, ainsi que le seuil minimal de confiance *minconf*. Pour chaque $k$-séquence fermée fréquente de la langue source $S \in \mathcal{STFF}_f$, l'algorithme dérive toutes les règles d'association de la forme : $S \Rightarrow S^{'}$, tel que $S^{'}$ est une séquence fermée fréquente en langue cible **E** appartenant à l'ensemble $\mathcal{STFF}_e$. L'algorithme vérifie, pour chaque séquence $S^{'}$ en langue cible **E**, si elle admet les mêmes identifiants de phrases *id_ph* que celles en langue source **F**, tel que :

$$\frac{Supp(S \cup S^{'})}{Supp(S)} \geq minconf \tag{4.4}$$

Dans ce cas, la séquence en langue cible $S^{'}$ est insérée dans la liste des conclusions valides représentant les traductions potentielles de $S$. L'algorithme s'arrête lorsqu'il n y a plus de séquences en langue source à traiter.

**Exemple 24** *La* Table *4.2 illustre des exemples de règles d'association inter-langues du français vers l'anglais, générées à partir du corpus d'apprentissage* Europarl. *Les règles notées*

*par RAIL-1-1 désignent dans ce cas des paires de traduction mot-à-mot alors que celles notées par RAIL-n-m représentent des paires de traduction entre des séquences de termes fermées fréquentes sources et cibles.*

| **RAIL-1-1** | **RAIL-*n*-*m*** |
|---|---|
| coopération ⇒ cooperation, $Conf = 0.86$ | ⟨toujours possible⟩ ⇒ ⟨possible to⟩, $Conf = 0.22$ |
| coopération ⇒ development, $Conf = 0.14$ | ⟨toujours possible⟩ ⇒ ⟨always possible⟩, $Conf = 0.28$ |
| coopération ⇒ countries, $Conf = 0.11$ | ⟨toujours possible⟩ ⇒ ⟨it is not always⟩, $Conf = 0.11$ |

TABLE 4.2 – Exemples de règles d'association inter-langues du français vers l'anglais extraites du corpus d'apprentissage EUROPARL.

La génération de l'ensemble de règles d'association inter-langues induit ainsi la mise en place d'un nouveau modèle de traduction que nous définissons dans ce qui suit.

### 4.5.3 Modèle de traduction à base de règles d'association inter-langues

La définition du modèle de traduction à base de règles d'association inter-langues implique nécessairement la construction de la table de traduction à partir des RAILs qui se fait en deux étapes :

1. Dériver les associations inter-langues ayant un seul mot au niveau de la prémisse et de la conclusion. Ces associations permettent ensuite d'estimer la table de paires de traduction mot-à-mot, notée par RAIL-1-1. Cette étape donne lieu à un modèle de traduction statistique à base de mots [Latiri *et al.*, 2010b].
2. Dériver les associations inter-langues à partir des séquences de termes fermées fréquentes pour aboutir à une table de traduction contenant des paires de traduction de séquences, notée RAIL-*n*-*m*. En la fusionnant avec la table RAIL-1-1, nous générons ainsi la nouvelle table de traduction, correspondante à notre nouveau modèle de TAS à base de séquences [Latiri *et al.*, 2011].

#### Modèle de traduction à base de mots : les RAIL-1-1

Au niveau de ce modèle, le vocabulaire utilisé, dont les statistiques sont décrites dans la TABLE 4.3, est l'ensemble de mots fréquents par rapport à un seuil prédéfini *minsupp*. Les associations inter-langues entre ces mots sont ensuite dérivées en utilisant l'algorithme GEN-MGB [Latiri *et al.*, 2012b], qui génère une base générique minimale d'associations $\mathcal{MGB}$ (introduite dans le chapitre 2, page 67). Il est utile de noter qu'afin d'extraire les termsets fermés fréquents et leur générateurs minimaux, nous avons essayé d'utiliser plusieurs algorithmes dédiés à cette tâche, tels que PRINCE [Hamrouni *et al.*, 2005], CLOSE [Pasquier *et al.*, 1999], CHARM [Zaki and Hsiao, 2002], TITANIC [Stumme *et al.*, 2002] et CARD [Latiri *et al.*, 2005a]. Toutefois, tous ces algorithmes ont échoué à traiter un contexte d'extraction textuel aussi volumineux et dense que le corpus d'apprentissage EUROPARL contenant 596. 381 phrases et 137. 898 termes différents.

Pour dépasser cette limite, nous avons adapté l'algorithme GC-GROWTH [Haiquan *et al.*, 2005] pour l'extraction des termsets fermés fréquents et leurs générateurs minimaux à partir du corpus parallèle, en faisant varier le seuil de support minimal *minsupp*. Tel que décrit dans la

| ***minsupp*** | **Vocabulaire Français** | **Vocabulaire Anglais** |
|---|---|---|
| 20 | $17,2 \times 10^3$ | $11.4 \times 10^3$ |
| 30 | $14,2 \times 10^3$ | $9,6 \times 10^3$ |

TABLE 4.3 – Caractéristiques des vocabulaires Français et Anglais utilisés par les règles d'association inter-langues.

TABLE 4.4, nous avons obtenu un nombre très important de termsets fermés fréquents ainsi que leurs générateurs correspondants [Latiri *et al.*, 2010b].

Néanmoins, pour des valeurs faibles du seuil de support minimal *minsupp* et même avec les optimisations apportées à l'algorithme GEN-MGB et à l'algorithme GC-GROWTH [Haiquan *et al.*, 2005], le nombre très élevé de termsets fermés et de générateurs minimaux, constituent un réel obstacle vers une extraction efficace de règles d'association inter-langues. En effet, plus le seuil du support minimal est faible, plus la densité du contexte d'extraction devient élevée. Nous nous sommes ainsi limités à un seuil de support minimal égal à 20 phrases. Ce seuil est considéré comme très faible par rapport à ce qui est habituellement utilisé par la communauté de datamining.

| ***minsupp*** | **# TFF** | **# Gen** |
|---|---|---|
| 30 phrases | 3, 50 | 2, 70 |
| 25 phrases | 4, 70 | 3, 64 |
| 20 phrases | 5, 20 | 6, 70 |

TABLE 4.4 – Nombre de termsets fermés fréquents (TFF) et leurs générateurs minimaux respectifs (Gen), exprimés en *millions*, en fonction du seuil de support minimal *minsupp*.

Ainsi, les traductions potentielles d'un mot $f$ en langue source **F** qui apparaît dans la prémisse d'une association inter-langues sont obtenues en sélectionnant tous les mots de la langue cible $e_1, e_2, \ldots, e_n$ qui sont présents dans la conclusion de la même règle d'association. L'axiome de décomposition relatif à la base générique $\mathcal{MGB}$ [Latiri *et al.*, 2010b] est appliqué afin de dériver les associations contenant un seul mot au niveau de la prémisse et de la conclusion, *i.e.,* les RAIL-1-1. Le principe de la décomposition conditionnelle est que chaque association de la forme $r : f \Rightarrow e_1, \ldots, e_n \in \mathcal{MGB}$, les différentes règles d'association inter-langues, de la forme : $\forall\ i \in [1..n]$, $r : f \Rightarrow\ e_i$, sont dérivées en tant qu'associations valides. Dans le cas où le termset $\{f\ e_i\}$ n'appartient pas à l'ensemble des termsets fermés fréquents $\mathcal{TFF}$, alors la valeur de support qui lui est attribuée est celle du plus petit termset fermé fréquent de l'ensemble $\mathcal{TFF}$ contenant le termset $\{f\ e_i\}$.

Formellement, une entrée dans la table de traduction RAIL-1-1 est définie comme suit :

$$\begin{array}{l} \forall\ j \in [1..n],\ f \Rightarrow e_1, \ldots, e_n \in \mathcal{MGB} \\ r_j : f \Rightarrow e_j \in \text{RAIL-1-1}\ \wedge\ Conf(r_j) \geq minconf \end{array} \tag{4.5}$$

La conversion de la valeur de la confiance de l'association inter-langues en probabilité se fait comme suit :

$$\forall\ f, e_i \in PT(f),\ P_{RAIL}(e_i|f) = \frac{Conf(r : f \Rightarrow e_i)}{\sum_{e \in PT(f)} Conf(r : f \Rightarrow e)} \tag{4.6}$$

sachant que $PT(f)$ représente les traductions potentielles du mot source $f$.

**Modèle de traduction à base de séquences : les RAIL-*n*-*m***

L'idée est de considérer que les traductions potentielles d'une séquence $S_f$ en langue source **F**, qui apparaît comme prémisse de $n$ règles d'association valides, sont obtenues en sélectionnant les séquences fermées fréquentes en langue cible $S_{e_1}, S_{e_2}, \ldots, S_{e_n}$ qui sont les conclusions de ces mêmes règles d'association inter-langues.

Formellement, une entrée dans la table de traduction RAIL-*n*-*m* est définie comme suit [Latiri *et al.*, 2010a] :

$$\forall\ j \in [1..n],\ r_j : S_f \Rightarrow S_{e_j} \in \text{RAIL-n-m} \wedge Conf(r_j) \geq minconf \tag{4.7}$$

sachant que $n$ est le nombre de règles d'association valides ayant $S_f$ comme prémisse.

La normalisation des confiances en probabilités se fait de la même manière, que celle décrite dans l'équation (4.6).

## 4.6 Évaluation des règles d'association inter-langues

### 4.6.1 Stratégies d'évaluation et résultats

Lors de la génération des règles d'association inter-langues, nous avons fixé en amont des seuils de support minimal *minsupp* très faibles, qui paradoxalement s'avèrent très élevés si nous les comparons aux seuils de fréquences minimaux utilisés dans les approches statistiques de traduction automatique [Lavecchia *et al.*, 2007]. Ceci pénalise nos résultats dans la mesure où certains termes du vocabulaire sont élagués lors du processus de fouille et ne figurent donc pas dans les tables de traduction générées.

Pour pallier à cette limite, nous avons fixé des seuils de confiance très faibles pour favoriser la dérivation d'un nombre élevé de règles d'association inter-langues, autrement dit un nombre élevé de traductions potentielles pour chaque unité linguistique.

Afin d'évaluer la pertinence de l'approche de TAS proposée, les tables de traduction, *i.e.,* RAIL-1-1 et RAIL-*n*-*m*, dérivées à partir du corpus d'apprentissage Europarl (*cf.* Table 4.1, page 114) sont intégrées dans un processus complet de TAS. Nous considérons, comme paire de langues source et cible, le français et l'anglais. Nous avons comparé nos résultats d'une part à la méthode de référence de l'état de l'art [Koehn *et al.*, 2003], et à celle basée sur les triggers inter-langues [Lavecchia *et al.*, 2007, Lavecchia *et al.*, 2008], d'autre part. Pour ce faire, deux étapes expérimentales s'imposent :

1. **Optimisation du décodeur :** Dans le cadre de nos recherches dans le domaine de la TAS, nous avons démarré nos évaluations avec le décodeur Pharaoh [Koehn, 2004] et par souci de cohérence des résultats, nous avons conclu nos évaluations avec ce décodeur [26]. Le processus de traduction automatique effectué par ce décodeur implique quatre modèles, à savoir : (1) une table de traduction ($tm$) qui permet de trouver les traductions de chaque

26. Il importe de mentionner que nos travaux de recherche qui sont en cours de progression en TAS sont évalués avec le décodeur de référence actuel Moses.

composant de la phrase à traduire ; (2) un modèle de langage ($lm$) qui contrôle la vraisemblance de la traduction produite ; (3) un modèle de distorsion ($d$) qui permet d'ordonner différemment les mots de la phrase à traduire ; et, (4) une pénalité ($w$) qui prend en compte la différence de taille entre la phrase à traduire et la traduction proposée. Chacun de ces modèles a un poids dans le calcul des hypothèses de traduction. La qualité des traductions automatiques produites dépend fortement de la valeur de ces poids. Les différents poids des modèles ont été optimisés sur le corpus de développement EUROPARL (*cf.* TABLE 4.1, page 114). Les paramètres optimaux pour la table de traduction RAIL-1-1 sont décrits dans la TABLE 4.5.

2. **Validation sur le corpus de test :** Après l'optimisation des différents paramètres des tables de traduction basées sur les règles d'association inter-langues, mais aussi ceux du décodeur PHARAOH, une dernière étape de validation sur un corpus de test est nécessaire pour tirer les premières conclusions concernant notre contribution. La TABLE 4.5 montre les performances en terme de score BLEU [Papineni *et al.*, 2001] données par le décodeur PHARAOH sur le corpus de test, constitué de 500 phrases. Ces scores BLEU sont réalisés, respectivement, avec le modèle de traduction basé sur les règles d'association inter-langues mot-à-mot (RAIL-1-1), le modèle IBM-3 et le modèle de traduction à base des triggers inter-langues (Trig-1-1). Nous avons également considéré comme modèle de langage, le modèle IBM-3 entraîné avec GIZA++ [Och and Ney, 2000].

| **Modèle** | **tm** | **lm** | **d** | **w** | **BLEU** |
|---|---|---|---|---|---|
| RAIL-1-1 | 1 | 0.8 | 0.2 | -1 | 22.07 |
| Modèle IBM 3 | 0.6 | 0.7 | 0.4 | -1 | 29.57 |
| Trig-1-1 | 0.9 | 0.8 | 0.4 | -3 | 30.97 |

TABLE 4.5 – Comparison du modèle RAIL-1-1 au modèle IBM-3 et au modèle à base des triggers inter-langues.

Le score BLEU de 22.07 obtenu par le modèle de traduction RAIL-1-1 est considéré comme faible. Ceci peut être expliqué par la taille du vocabulaire utilisé, comme le montre la TABLE 4.3. En comparaison avec le vocabulaire utilisé par la table de traduction à base des triggers inter-langues, notre vocabulaire ne peut pas dépasser 17 200 mots pour un seuil de support minimal *minsupp* égal à 20 [Latiri *et al.*, 2010b]. Pour des limitations de calcul, il devient contraignant d'extraire des règles d'association inter-langues pour des seuils de support minimal *minsupp* inférieurs à 20. Ainsi, un seuil de *minsupp* égal à 20 phrases écarte du vocabulaire les mots ayant des supports plus faibles, qui par conséquent ne vont pas apparaître dans le règles d'association inter-langues. Ceci justifie l'écart entre le score obtenu par le modèle RAIL-1-1 (**22.07**) et celui du modèle IBM-3 (**29.57**) (*cf.* TABLE 4.5).

Par ailleurs, comme l'illustre la TABLE 4.6, les évaluations du modèle de traduction RAIL-1-1 ont montré que les règles exactes, *i.e.,* les règles ayant une confiance égale à 1, expriment des traductions correctes des mots en français vers l'anglais. À titre d'exemple, la paire de traduction $Tempêtes \Rightarrow Storms$, figurant dans la TABLE 4.6, exprime une traduction exacte avec une probabilité égale à 1.0. Nous avons également noté que l'algorithme GEN-MGB génère des règles d'association approximatives avec une forte confiance. Leurs conclusions représentent généralement des traductions possibles des termes de leurs prémisses. Toutefois, nous remarquons que certaines associations inter-langues sont non significatives et introduisent du bruit dans la table de traduction. Ces associations peuvent apparaître avec des confiances fortes ou faibles,

puisque le filtrage est basé uniquement sur les métriques statistiques, à savoir le support et la confiance.

| **Terme en français** | **Traduction potentielle en anglais** | $P_{RAIL}(e_i\|f_i)$ |
|---|---|---|
| Coopération | cooperation | 0.13 |
| | development | 0.02 |
| | countries | 0.01 |
| Pêche | fisheries | 0.08 |
| | fishing | 0.06 |
| | policy | 0.02 |
| Difficulté | difficulty | 0.16 |
| | difficulties | 0.06 |
| | problem | 0.03 |
| Compétences | powers | 0.05 |
| | competences | 0.01 |
| | competence | 0.01 |
| Alimentaire | food | 0.15 |
| | safety | 0.06 |
| | aid | 0.02 |
| Tempêtes | storms | 1.00 |

TABLE 4.6 – Exemples de paires de traduction entre mots (RAIL-1-1).

Nous résumons, dans la TABLE 4.7, les évaluations menées avec les règles inter-langues à base de séquences de termes fermées fréquentes, *i.e.,* en utilisant le modèle de traduction RAIL-*n*-*m*, sur le corpus de test d'EUROPARL. Nous notons que plus le seuil de *minsupp* est élevé, plus le score BLEU diminue. Ceci est justifié par le nombre de traductions candidates qui sont éliminées en élevant le seuil de support minimal.

| ***minsupp*** | ***minconf*** | $\#\ \mathcal{STFF}_f$ | $\#\ \mathcal{STFF}_e$ | **Score BLEU** |
|---|---|---|---|---|
| 20 | 0.1 | 219 090 | 187 207 | **34.18** |
| 30 | 0.1 | 140 082 | 120 317 | 33.79 |
| 100 | 0.1 | 39 024 | 33 765 | 32.10 |

TABLE 4.7 – Évaluation du modèle de traduction RAIL-*n*-*m* en fonction du seuil *minsupp*.

Dans le contexte de la TAS à base de séquences, le modèle à base des triggers inter-langues, qui a été étendu vers les séquences [Lavecchia *et al.*, 2008], a donné sur le même corpus de test un score BLEU égal à **34.41**. Par ailleurs, le modèle à base des triggers considère des séquences de taille maximale égale à 5 mots ; or dans le cadre de notre approche, nous n'avons imposé aucune limite de taille pour les séquences fermées fréquentes générées. Ces dernières peuvent même atteindre une taille de 25 mots. Nous avons également remarqué, de manière expérimentale qu'au delà d'une certaine taille, *i.e.,* par exemple des séquences de 14 mots, le score BLEU cesse d'augmenter comme le montre la FIGURE 4.2. Ceci est expliqué par le fait que ces séquences sont très peu fréquentes dans le corpus de développement, et elles constituent des séquences rares

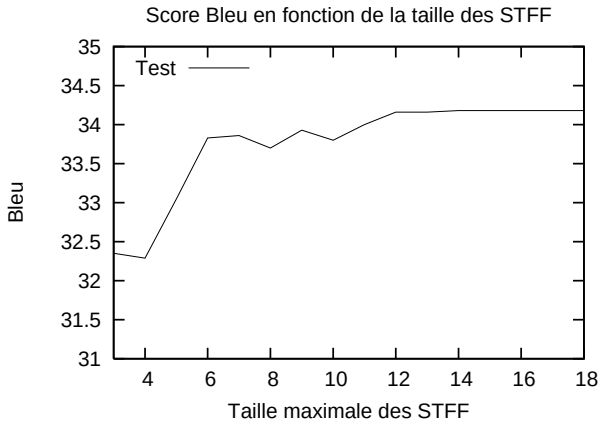


FIGURE 4.2 – Score BLEU en fonction de la taille de séquences de termes fermées fréquentes.

avec un très faible support. Ainsi, leur support, dans le corpus de test, est encore plus faible et donc elles n'interviennent pas dans le calcul du score BLEU.

La TABLE 4.7 montre que le modèle de traduction RAIL-*n*-*m* réalise un score BLEU de **34.18**. Ainsi, en ajoutant les séquences fermées fréquentes pertinentes, nous obtenons une amélioration de plus de 12 points du score BLEU par rapport à celui donné par le modèle de traduction mot-à-mot, *i.e.,* RAIL-1-1, avec un score BLEU de **22.07**.

### 4.6.2 Couplage des règles d'association avec les triggers inter-langues

Les évaluations du modèle de traduction RAIL-1-1 ont donné un score BLEU relativement faible (**22.07**) contrairement à l'approche à base des triggers, qui donne un score BLEU meilleur (**30.97**) que celui obtenu avec le modèle IBM-3 (**29.57**). Par ailleurs, les deux approches ont donné des scores BLEU comparables dans le contexte de la TAS à base de séquences. Afin de mettre en évidence les avantages de chacune de ces approches, nous avons jugé utile de construire une nouvelle table de traduction, notée par Trig-RAIL-*n*-*m*, qui est composée par le couplage de deux tables de traductions, à savoir, la table Trig-1-1 représentant les paires de traduction mot-à-mot, et la table RAIL-*n*-*m* représentant les paires de traduction à base de séquences fermées fréquentes de termes. Le résultat de l'évaluation de la nouvelle table de traduction est donné dans la TABLE 4.8. Nous constatons une nette amélioration du score BLEU (**35.52**). Ce résultat montre aussi que nous nous approchons du résultat de la méthode de référence de l'état de l'art de Koehn *et al.* [Koehn *et al.*, 2003] qui donne un score BLEU de (**37.15**).

Avec ce résultat, nous confirmons la pertinence des séquences fermées fréquentes découvertes et de leurs traductions dans un processus de TAS, sans alignment au préalable. Cette pertinence est étroitement liée au fait que les séquences sont élaguées en amont lors du processus d'extraction pour ne retenir que celles qui sont non-redondantes et valides par rapport aux seuils de $minsupp$ et $minconf$. Cette sélection minimise ainsi le bruit dans les tables de traduction générées à partir de ces séquences [Latiri *et al.*, 2011].

Les évaluations empiriques nous ont permis de constater que la variation du nombre et de la taille des séquences dans les tables de traduction influe sur l'amélioration du score BLEU. La FIGURE 4.3 et la FIGURE 4.4 illustrent l'évolution du nombre de séquences en fonction du nombre de mots par séquence. Nous remarquons que la courbe relative au modèle de référence

| **Modèle** | **w** | **tm** | **lm** | **d** | **BLEU** |
|---|---|---|---|---|---|
| Koehn *et al.* | 0 | 0.8 | 0.6 | 0.3 | 37.15 |
| RAIL-*n*-*m* | -1 | 0.9 | 0.8 | 0.3 | 34.18 |
| Trig-*n*-*m* | 0 | 0.7 | 0.9 | 0.4 | 34.41 |
| Trig-RAIL-*n*-*m* | -1 | 0.9 | 0.8 | 0.3 | 35.52 |

TABLE 4.8 – Évaluation des différentes tables de traduction à base de séquences sur le corpus de test.

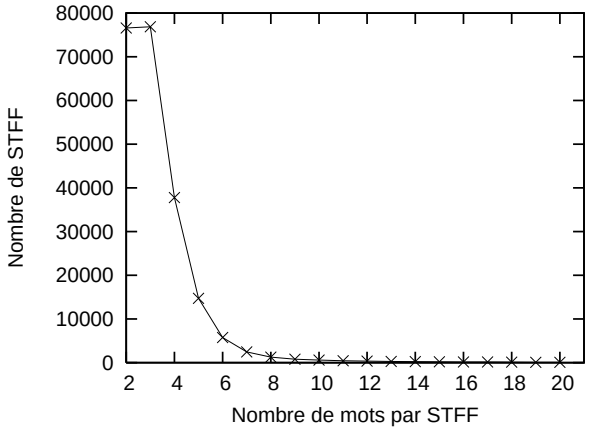


FIGURE 4.3 – Évolution du nombre de séquences dans la table de traduction à base des RAILs.

.

de Koehn et *al.* [Koehn *et al.*, 2003] a une forme cubique [27], alors que la courbe relative à notre modèle de traduction RAIL-*n*-*m* correspond à une fonction puissance décroissante [28]. Nous notons également que la courbe relative au modèle de Koehn et *al.* (*cf.* FIGURE 4.4) augmente fortement jusqu'à une taille égale à 5 et diminue pour des séquences plus longues, tandis que pour notre modèle, plus la longueur des séquences augmente et plus la courbe diminue et tend à devenir asymptotique sur l'axe des abscisses (*cf.* FIGURE 4.3).

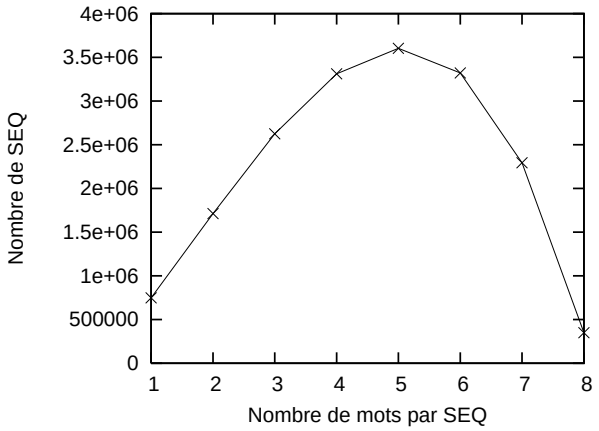


FIGURE 4.4 – Évolution du nombre de séquences dans la table de traduction de Koehn *et al.*

Par conséquent, l'écart entre notre résultat et celui du modèle de référence de Koehn *et al.* [Koehn *et al.*, 2003] pourrait être expliqué par le réel impact des séquences de taille quatre et cinq. Alors que le nombre de séquences de 5 mots dans la table de Koehn *et al.* est de environ 2,8 millions mots, notre table de traduction englobe seulement 10 500 séquences de taille 5. Rappelons aussi que notre approche de fouille de séquences fermées fréquentes est beaucoup plus stricte que celle utilisée par Koehn *et al.*, dérivant beaucoup moins de séquences étant donné qu'elle élague les séquences redondantes. De ce fait, nos séquences apparaissent plus rarement dans le corpus de test, et leur impact positif s'avère donc plus faible.

## 4.7 Bilan des contributions

La contribution présentée dans ce chapitre introduit une définition originale d'un modèle de traduction à base de séquences, présentant une autre alternative aux modèles d'IBM pour l'apprentissage des tables de traduction de mots ou de séquences de mots. La première originalité de notre approche tient au fait que nous sélectionnons les séquences de mots bilingues par le biais d'une technique d'extraction de séquences fermées fréquentes issue du domaine de datamining [Chang, 2004], que nous avons adaptée à l'ECT. Ceci nous a permis de constituer un ensemble de séquences sources et cibles pertinentes. La deuxième originalité est que nous utilisons les règles d'association inter-langues (RAILs) pour déterminer les traductions des séquences sources sans faire appel à l'alignement des mots au sein du corpus d'apprentissage.

Le modèle de traduction à base de mots RAIL-1-1 ainsi que le modèle à base de séquences RAIL-*n*-*m*, ont permis d'estimer l'ensemble de mots et de séquences sources associées à leurs

27. Dans la table de Koehn et *al.*, nous avons $y = -28953X^3 + 148491X^2 + 720014X - 90643$
28. Notre méthode : $y = 1.36E6X^{-3.6}$

traductions potentielles. Afin d'évaluer ces deux modèles de TAS, nous avons utilisé le décodeur PHARAOH pour mener une série de tests sur le corpus parallèle EUROPARL en choisissant comme langues source et cible, respectivement, le français et l'anglais. Chaque stratégie d'évaluation est comparée à l'approche de référence de Koehn et *al.* [Koehn *et al.*, 2003] qui consiste à extraire la table de traduction de séquences à partir de l'alignement bi-directionnel de mots, et également à l'approche basée sur les triggers inter-langues [Lavecchia *et al.*, 2008], représentant la genèse de notre approche à base de règles d'association inter-langues.

Dans un premier temps, nous avons adapté l'algorithme BFSM pour extraire les séquences fermées fréquentes à partir du corpus d'apprentissage et nous avons généré dans un deuxième temps les règles d'association inter-langues mot-à-mot, *i.e.,* RAIL-1-1, en utilisant l'algorithme GEN-MGB [Latiri *et al.*, 2010b], pour les étendre ensuite moyennant les séquences fermées fréquentes aux RAIL-*n-m* [Latiri *et al.*, 2010a, Latiri *et al.*, 2011]. Par ailleurs, la densité du contexte d'extraction a contraint le processus de fouille, dans le sens où les limitations de calcul nous ont amené à baisser les seuils de support minimal *minsupp* sans pouvoir dépasser un *minsupp* égal à 20 phrases. Ceci a pénalisé nos résultats dans la mesure où certains termes du vocabulaire sont élagués lors du processus de fouille et ne figurent pas dans les tables de traduction générées. De ce fait, nous avons baissé les seuils de confiance minimale $minconf$ pour favoriser la dérivation d'un nombre élevé de règles d'association inter-langues, autrement dit un nombre élevé de traductions potentielles pour chaque terme et chaque séquence fermée.

Toutefois, l'absence de termes du vocabulaire dû à la limite du seuil de support minimal a fait que l'utilisation des règles d'association inter-langues ne s'est pas avéré concluante dans le cadre de la traduction mot-à-mot, *i.e.,* RAIL-1-1 [Latiri *et al.*, 2010b]. La qualité des traductions a en effet baissé par rapport à celle des traductions produites avec la table de traduction du modèle d'IBM-3. En revanche, l'introduction des RAIL-*n-m*, dans notre table de traduction RAIL-1-1, a engendré un gain conséquent en terme de score BLEU.

Suite au couplage de notre table de traduction RAIL-*n-m* avec la table de traduction Trig-1-1, les traductions produites par le décodeur PHARAOH ont atteint un score BLEU de **35,42** sur le corpus de test d'EUROPARL, soit un gain de **1,24** points par rapport au fait d'utiliser la table de traduction à base des RAILs exclusivement. Ce gain est expliqué d'un côté par la meilleure qualité des paires de traduction dans le modèle Trig-1-1 comparé au modèle RAIL-1-1, et, par la pertinence des séquences fermées fréquentes non-redondantes du modèle de TAS RAIL-*n-m*, d'un autre côté.

Il importe de noter que, les traductions produites avec notre modèle de traduction RAIL-*n-m* ont une qualité inférieure en terme de score BLEU, comparée avec celles obtenues avec le modèle de référence de Koehn *et al.* [Koehn *et al.*, 2003] et qui sont construites à partir de l'alignement des mots. Bien qu'ayant obtenu un score BLEU inférieur, notre approche ainsi que celle à base des triggers inter-langues [Lavecchia *et al.*, 2008] présentent certains avantages par rapport à cette dernière. Tout d'abord, la découverte des séquences fermées fréquentes, qui se fait en amont sur le corpus d'apprentissage, nous permet de contrôler la taille de la table ainsi que la pertinence des couples de traduction qu'elle contient. Dans [Lavecchia, 2010], l'auteur a noté, dans la table de Koehn *et al.* construite à partir de l'alignement des mots, que 50, 24% des entrées avaient une probabilité de traduction égale à 1 et que si elles étaient écartées durant la phase de décodage, la qualité des traductions produites par le décodeur PHARAO restait inchangée. En effet, dans son approche, Koehn *et al.* extrait autant de paires de séquences que possible pourvu qu'elles soient consistantes avec l'alignement des mots, avec le risque de sélectionner des couples non pertinents. Ceci conduit à des tables de traduction de taille démesurée de plusieurs millions d'entrées. À titre indicatif, la table de Koehn *et al.* en compte plus de 24 millions alors que notre table de traduction RAIL-*n-m* comporte 8 millions 300. 000 entrées. Un autre

avantage de notre modèle, tout comme le modèle à base de triggers inter-langues, tient au fait que nous n'utilisons pas l'alignement de mots pour sélectionner les couples de séquences de la table de traduction, contrairement à la méthode de Koehn *et al.*. Bien que cette dernière extrait les couples de séquences par de simples heuristiques, elle nécessite toutefois l'alignement bi-directionnel des mots du corpus d'apprentissage. Chacun des alignements Anglais-Français et Français-Anglais est calculé automatiquement par un processus itératif complexe, qui peut prendre, dans le cas d'un grand corpus comme EUROPARL des heures, voire des jours de calcul. De plus, les alignements résultants peuvent être entachés d'erreurs conduisant à de mauvais couples de traduction. Dans notre contribution, une fois les séquences fermées fréquentes sélectionnées, nous identifions leurs traductions potentielles à l'aide des règles d'association inter-langues. La sélection des associations s'opère par un processus simple en une seule itération et notre table de traduction basée sur les RAILs se construit beaucoup plus rapidement qu'une table mise en place à partir de l'alignement des mots.

Nos contributions futures avec des approches d'ECT consisteront à réduire cet écart et ce en repensant l'algorithmique d'extraction des séquences fréquentes. Il serait intéressant pour la TAS de relaxer les contraintes de sélection et d'élagage des séquences lors du processus de fouille, afin de favoriser l'augmentation du nombre de paires de séquences dans la table de traduction.

# Conclusion

Le premier constat que nous pouvons faire à partir de nos recherches menées dans le domaine de l'ECT, est qu'il n'existe pas de méthodes et d'outils génériques pour la découverte de connaissances utiles à partir de textes. Nous avons montré que ces connaissances peuvent se présenter sous diverses formes et peuvent être générées par différentes techniques, qui sont étroitement liées à la nature de l'application dans laquelle ces connaissances seront intégrées. Nous avons aussi remarqué que, indépendamment de la technique abordée pour la fouille de textes, des défis se posent pour la représentation des connaissances découvertes et leur évaluation, ainsi que pour leur utilisation pour la RI ou la TAS.

Les résultats que nous avons obtenu avec l'ensemble des approches proposées, tant en RI qu'en TAS, restent encourageants dans le sens où nous avons pu dépasser certaines limites de calcul des approches existantes et nous avons montré la possibilité de déployer des connaissances issues de l'ECT dans deux applications ayant des spécificités différentes, *i.e.,* la RI et la TAS. En se positionnant par rapport aux travaux de la littérature récente, nous avons réussi avec certains scénarios, à dépasser les baselines de certaines approches pionnières en RI et à nous rapprocher des baselines de référence en TAS.

Toutefois, l'ensemble de nos contributions restent ouvertes à diverses perspectives qui sont liées essentiellement à la nature des données textuelles considérées lors du processus de fouille. En premier lieu, nous avons constaté que, malgré les résultats obtenus, nos algorithmes d'extraction de règles d'association entre termes restent coûteux du point de vue temps et espace. Ceci se justifie par la grande taille des corpus textuels que nous manipulons en RI et en TAS. De ce fait, l'optimisation et la mise à l'échelle de ces algorithmes sont indispensables pour mieux exploiter les métriques de support et de confiance. Cette optimisation peut être envisagée dans le cadre de plateformes parallèles et/ou distribuées, et notamment les architectures multi-cœurs qui offrent des possibilités d'utilisation du multithreading.

Par ailleurs, l'évaluation de la qualité des règles d'association inter-langues dans le domaine de TAS s'est faite uniquement par la seule mesure du score BLEU. Or, ces règles ont plus d'intérêt et d'apport si nous considérons le domaine de la Recherche d'Information Multilingue (RIM) [Peters *et al.*, 2012].

Ainsi, les perspectives de recherche que nous souhaitons développer représentent une extension de nos contributions dans le domaine de l'ECT, en considérant le caractère multilingue des corpus et des ressources externes existantes telles que les ontologies.

Nous proposons deux principales orientations de recherche futures que nous détaillons dans la partie suivante du présent rapport.

# Troisième partie

# Projet de Recherche

# Orientations et problématiques de recherche futures

**Sommaire**



## 1 Objectifs du chapitre

Ce chapitre synthétise mon projet de recherche qui s'inscrit à la confluence de deux domaines actifs de recherche : la Recherche d'Information Multilingue (RIM) et l'Analyse des Réseaux Sociaux (ARS). Nous mettons en exergue un ensemble de problématiques pertinentes pour le déploiement des motifs découverts à partir des corpus multilingues. Dans ce contexte, nous proposons de nouvelles approches liées à l'indexation multilingue, l'appariement multilingue et l'enrichissement d'ontologies multilingues.

Comme continuité à nos contributions proposées dans le domaine de fouille de graphes [Douar *et al.*, 2011c, Douar *et al.*, 2011b, Douar *et al.*, 2011a], nous donnons également quelques réflexions relatives à la recherche de communautés et la prédiction de liens dans les réseaux sociaux. La FIGURE 1 illustre les deux principales orientations de recherche. La première met en évidence le déploiement possible des règles d'association inter-langues définies pour la TAS [Latiri *et al.*, 2010b, Latiri *et al.*, 2011], dans le domaine de la RI multilingue, à travers la fouille des corpus comparables [Talvensaari *et al.*, 2007, Sadat, 2010, Hazem and Morin, 2012a]. La deuxième orientation s'articule autour de l'utilisation des fondements de l'Analyse Formelle de Concepts

pour la découverte des fermés de cliques maximales fréquents. Notre motivation derrière l'extraction de ce type de motifs fréquents est de les utiliser pour l'analyse des réseaux sociaux. L'ensemble des réflexions de recherche exposées dans ce chapitre vont s'intégrer à court terme dans des propositions de sujets de thèse de doctorat et de mastères.

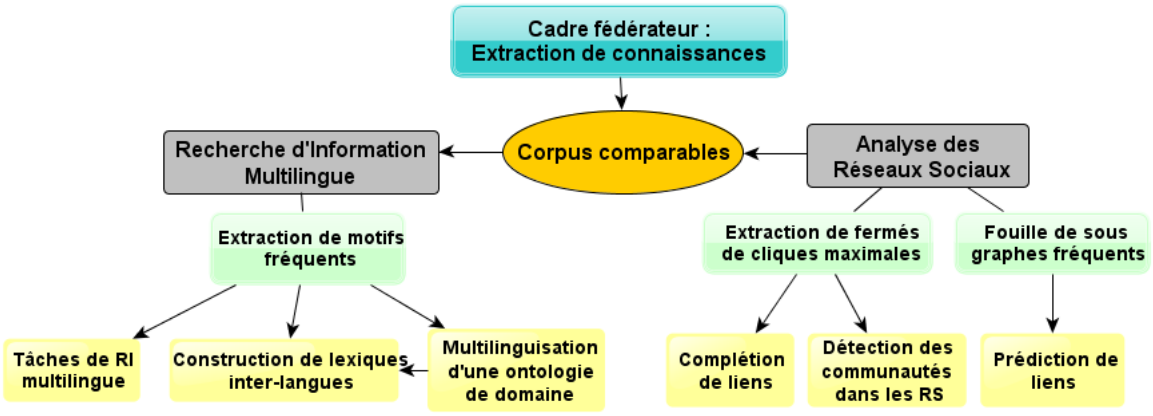


FIGURE 1 – Positionnement du projet de recherche.

Dans le cadre de la RIM, nous partons du fait que les données fournies par les corpus multilingues permettent d'inférer des relations de traduction entre des unités linguistiques. Nous assumons que l'intégration de différentes ressources de connaissances permet de renforcer globalement les méthodes d'extraction de lexiques bilingues.

Notre objectif, à travers l'ensemble de réflexions de recherche proposées dans le domaine de la RIM, est de tenter de déployer les techniques d'ECT pour la recherche et l'identification de lexiques bilingues à partir des corpus comparables (non parallèles).

## 2 Corpus parallèles *vs* corpus comparables

L'extraction de lexiques bilingues à partir de corpus multilingues s'est largement concentrée sur les corpus parallèles, notamment pour améliorer la pertinence de la TAS. Les approches proposées dans la littérature s'appuient principalement sur l'hypothèse de correspondance entre une phrase dans un document source et sa traduction dans le document cible [Renders *et al.*, 2002]. Autrement dit, l'organisation du document cible sera similaire à l'organisation du document source (*i.e.,* ordre des phrases). Cette hypothèse nous a permis d'adapter (*cf.* chapitre 4, page 109), des techniques d'ECT telles que l'extraction des séquences fréquentes et les règles d'association inter-langues pour la TAS [Latiri *et al.*, 2010b, Latiri *et al.*, 2011].

Bien que les corpus parallèles constituent une ressource incontestable aussi bien pour la construction de systèmes de TAS que pour la génération de dictionnaires bilingues, leur emploi présente plusieurs limites [Hazem and Morin, 2012b]. La première limite concerne la rareté de ces corpus, qui d'autant plus, nécessitent une traduction humaine coûteuse et peu disponible entre couples de langues n'impliquant pas toujours l'anglais. En outre, il est difficile de trouver des corpus parallèles en quantité suffisante, même entre langues largement utilisées. La seconde limite vient du fait qu'ils contiennent des textes qui sont le résultat d'une traduction, et risquent ainsi d'être de mauvais représentants d'une langue [Hazem and Morin, 2012a].

Les corpus comparables visent à pallier ces deux limites, en allant puiser dans les corpus monolingues dans deux langues source et cible, composés de textes non parallèles, mais reliés au même domaine. Ces types de corpus constituent par essence des ressources abondantes puisqu'ils sont composés de documents partageant différentes caractéristiques tel que le domaine. Notons également que l'usage des corpus comparables élimine le biais de traduction étant donné que les textes sont des originaux [Li and Gaussier, 2010, Hazem *et al.*, 2011].

Il importe de souligner que la majorité des travaux d'extraction de lexiques bilingues à partir de corpus comparables s'inscrivent dans le cadre d'une sémantique distributionnelle [Renders *et al.*, 2002], et décrivent le sens d'une unité linguistique (mot ou séquence de mots) par sa distribution sur un ensemble de contextes. La mise en relation entre deux unités de langues différentes s'établit sur un plan sémantique ; et le corpus bilingue est vu ici comme un objet d'acquisition de connaissances et de mise à jour de ressources lexicales pré-existantes tels que les thésaurus et les ontologies.

Toutefois, les lexiques bilingues extraits à partir de corpus comparables sont d'une qualité bien inférieure à ce qui peut être obtenu à partir de corpus parallèles [Goeuriot *et al.*, 2009, Bo *et al.*, 2011].

La principale difficulté des approches liées à l'exploitation de corpus comparables par rapport aux corpus parallèles pour l'extraction de lexiques bilingues est l'absence d'éléments de référence entre les documents des langues source et cible composant les corpus comparables. Face à cette difficulté, les différentes approches liées à l'exploitation de corpus comparables reposent sur la simple observation qu'un mot et sa traduction ont tendance à apparaître dans les mêmes environnements lexicaux. La mise en oeuvre de cette observation repose sur l'identification *d'affinités du premier ordre* (*i.e.*, identifier les mots qui sont susceptibles d'être trouvés dans le voisinage immédiat d'un mot donné) ou *d'affinités du second ordre* (*i.e.*, identifier les mots qui partagent les mêmes environnements lexicaux sans nécessairement apparaître ensemble [Grefenstette, 1994]).

Nous distinguons deux approches pionnières associées à l'identification des lexiques bilingues à partir de corpus comparables, à savoir :

1. **L'approche standard** [Fung and McKeown, 1997] **:** Elle est basée sur une analyse du contexte lexical des mots et repose sur la simple observation qu'un mot et sa traduction tendent à apparaître dans les mêmes contextes lexicaux. La mise en oeuvre de cette observation repose sur l'identification d'affinités du premier ordre [Grefenstette, 1994]. Ces affinités peuvent être représentées sous la forme d'un vecteur de contexte, où chaque élément du vecteur représente un mot qui apparaît dans différentes fenêtres contextuelles [Goeuriot *et al.*, 2009]. Le principal inconvénient de l'approche standard est que ses performances dépendent de la couverture du dictionnaire bilingue par rapport au corpus comparable. En effet, en traduisant un maximum d'entrées du vecteur de contexte du mot à traduire, on maximise les chances de retrouver sa traduction. Bien que la couverture du dictionnaire puisse être étendue en utilisant des dictionnaires spécialisés ou des thésaurus multilingues [Renders *et al.*, 2002, Déjean *et al.*, 2002], la traduction des éléments du vecteur de contexte reste au centre de cette approche.
2. **L'approche par similarité interlangue :** Dans le but de rendre l'approche standard moins dépendante des dictionnaires spécialisés ou de ressources multilingues externes, Déjean et Gaussier ont proposé une extension de l'approche standard connue sous le nom d'approche par similarité interlangue [Renders *et al.*, 2002, Bo *et al.*, 2011]. Cette approche se base sur l'idée que les mots ayant le même sens, partagent les mêmes environnements lexicaux. Elle repose sur l'identification d'affinités du second ordre [Grefenstette, 1994]. Les

mots partageant des affinités du second ordre n'ont pas besoin d'apparaître ensemble, mais leurs environnements sont semblables. Dans cette approche, le dictionnaire bilingue établit un lien entre les langues du corpus comparable. L'approche par similarité interlangue a l'avantage d'éviter les traductions directes des éléments des vecteurs de contextes.

Nous orientons ainsi nos premières réflexions de recherche vers la RIM dans le domaine médical. Nous considérons une collection fournie dans le cadre du projet MUCHMORE (http ://muchmore/dfki.de). Nous avons utilisé un corpus composé de 845 résumés d'articles scientifiques provenant de la base de données MEDLINE (http ://www4.ncbi.nlm.nih.gov/PubMed) rédigés en anglais et en allemand et portant sur le domaine de la chirurgie, ce qui correspond environ à 100 000 mots pour chaque langue. Les résumés anglais sont en général des traductions des résumés allemands. Dans de nombreux cas toutefois, la version anglaise est une reformulation complète du résumé allemand, ce qui rend difficile un alignement au niveau des phrases. Ce corpus est considéré comme étant un corpus parallèle bruité. Associée à ce corpus, nous utilisons comme ressource bilingue l'ontologie anglaise du domaine médical MeSH (http ://www.nlm.nih.gov/mesh/MBrowser.html), et sa version allemande, DMD, fournie par le Deutsches Institut fuer Medezinische Dokumentation und Information (http ://www.dimdi.de). Le nombre d'entrées alignées est d'environ 15000.

## 3 Orientation 1 : Extraction de lexiques bilingues pour la RI multilingue

Le bilan des différentes contributions proposées dans les domaines de la Recherche d'Information et de la Traduction Automatique Statistique, nous ont permis de positionner une partie de nos travaux de recherche futurs dans le domaine de la Recherche d'Information Multilingue [Peters *et al.*, 2012]. Ce positionnement est justifié par les points suivants : (*i*) la disponibilité des corpus comparables liés à des domaines spécifiques ; et, (*ii*) le multilinguisme des ressources externes disponibles, telles que les ontologies multilingues [Nie *et al.*, 2012].

En effet, l'abondance des documents, notamment sur le Web, dans de nombreuses langues a rendu nécessaire l'existence d'une Recherche d'Information Multilingue (RIM), appelée aussi Recherche d'Information Interlangue [Savoy, 2002, Nie, 2010]. La RIM consiste ainsi à formuler une requête dans une langue source et à rechercher des documents pertinents dans des langues cibles. Le défi consiste donc à trouver une correspondance entre une requête formulée dans une langue donnée et les documents associés dans une ou plusieurs autres langues.

La revue de la littérature montre que la RIM aborde quatre configurations possibles du modèle de RI dans un contexte multilingue, à savoir [Hull and Grefenstette, 1996] :

1. La première configuration est considérée comme la plus simple à modéliser. Elle effectue une recherche qui prend en compte seulement la langue de la requête. Les documents concernés appartiennent à un corpus bilingue ou à une collection multilingue de documents. Cette configuration s'apparente en effet à une recherche d'information monolingue puisque le corpus est découpé en collections documentaires monolingues, indépendantes les unes des autres. Les documents de chacune des collections ne peuvent être restitués que par une requête formulée dans leur langue.
2. La seconde configuration interroge une collection monolingue de documents avec des requêtes exprimées dans des langues différentes de celle du corpus. Ce genre de SRI porte le nom Système de Recherche d'Information Multilingue (SRIM) car la représentation de l'information tente de passer outre les barrières langagières.

3. La troisième configuration de la RIM interroge une collection multilingue de documents dans plusieurs langues [Peters *et al.*, 2012]. Il s'agit d'une extension de la configuration précédente, où à partir d'une requête dans une langue donnée, on peut retrouver des documents exprimés dans chacune des langues de la collection.
4. La dernière configuration effectue une recherche d'information sur des documents multilingues où des parties d'un document sont formulées dans des langues différentes. Par exemple, à partir d'une requête dans une langue donnée, il est possible de retrouver des documents multilingues dont le résumé est exprimé en anglais et le corpus de textes en français.

Il importe de signaler que la plupart des travaux de recherche actuels sur la RIM tentent de relever les défis posés par l'interrogation multilingue dans un SRIM. Partant d'une requête exprimée dans une langue source, il s'agit de définir un modèle de RI capable de restituer des documents formulés dans chacune des langues cibles de la collection multilingue. Cette problématique a donné naissance à divers travaux qui traitent de la traduction de requêtes [Gao *et al.*, 2006, Wu and He, 2010, Dolamic and Savoy, 2010, Herbert *et al.*, 2011]. Dans ce contexte, plusieurs travaux ont émergé pour l'extraction de lexiques bilingues à partir de corpus comparables afin d'offrir une alternative crédible à l'exploitation de corpus parallèles [Hazem *et al.*, 2011].

Pour intégrer le multilinguisme en RI, trois principales approches ont été proposées. Ces propositions diffèrent principalement par le choix de la langue de l'espace d'indexation, *i.e.,* le choix de la langue utilisée pour construire les représentations des documents et des requêtes dans un SRIM. Les alternatives possibles sont [Oard, 1998] :

- *La langue du corpus de la collection* : les requêtes doivent être traduites dans la langue du corpus avant d'effectuer l'indexation [Dolamic and Savoy, 2010].
- *La langue de la requête* : le corpus doit être traduit dans la langue des requêtes avant d'effectuer l'indexation [McCarley and Roukos, 1998, Roussey *et al.*, 2010].
- *Un langage pivot* : les documents et les requêtes doivent être traduits dans la langue de l'index, qui est différente de leurs langues d'origine respectives [Hahn *et al.*, 2004].

Ainsi, plusieurs types d'approches existent pour modéliser la RIM, à savoir [Oard, 1998] :

1. **Approches basées sur la traduction automatique :** Ces approches s'appuient sur la traduction automatique d'une requête ou des documents de la collection. La traduction automatique de la requête est l'issue de recherche la plus explorée [Nie *et al.*, 1999, Wu and He, 2010, Herbert *et al.*, 2011]. Toutefois, elle souffre d'un manque de précision comparée à celle basée sur la traduction d'une collection de documents [Oard, 1998, Hutchins, 2005]. Cette dernière utilise un contexte d'information nettement plus important, diminuant ainsi les risques de mauvaise traduction.
2. **Approches basées sur des corpus d'apprentissage parallèles ou comparables :** Ces approches dites statistiques [Gao *et al.*, 2006, Wu and He, 2010] s'appuient, pour la traduction des requêtes, sur un thésaurus ou une ontologie créés à partir d'un corpus parallèle aligné au niveau de la phrase [Nie *et al.*, 1999, Wang and Oard, 2005] ou d'un corpus comparable [Talvensaari *et al.*, 2007, Hazem *et al.*, 2011, Bo *et al.*, 2011] pour trouver des co-occurrences de termes.
3. **Approches basées sur des lexiques bilingues :** Ce type d'approches se base principalement sur l'expansion de requêtes [Baziz *et al.*, 2003]. Elle consiste à étendre la requête à l'aide de dictionnaires monolingues qui permettent la reformulation dans une même langue (synonymes, antonymes, etc.), et de lexiques bilingues qui permettent la reformulation de la requête dans des langues différentes [Aljlayl and Frieder, 2001, Markó *et al.*, 2005, Levow *et al.*, 2005].

4. **Approches à base d'un language pivot (l'interlingua) :** Ces approches extraient la sémantique des textes de la langue source grâce à un langage pivot [Hahn *et al.*, 2004]. Il s'agit d'un langage unifié permettant de représenter la sémantique des différentes langues. Il est généralement basé sur le concept des graphes représentant les phrases [Hahn *et al.*, 2004, Habash *et al.*, 2006, Toumouh *et al.*, 2011, Lesmo *et al.*, 2011].
5. **Approches à base de ressources externes :** Différents types de ressources externes sont utilisés pour trouver les traductions d'un terme dans le contexte de la RIM, tels que les thésaurus [Wang and Oard, 2005] et les ontologies [Potthast *et al.*, 2008]. Ces ressources permettent de mettre en évidence des associations et des relations possibles entre termes. Dans un thesaurus multilingue, la relation d'équivalence inclue les traductions et les synonymes des termes choisis comme représentant d'un concept [Knoth *et al.*, 2010]. En effet, au lieu d'effectuer une traduction mot à mot en considérant les termes comme indépendants, il faut d'abord chercher à les relier par des relations d'association pour obtenir la traduction la plus appropriée d'un concept multi-termes. Le problème soulevé aussi par les thésaurus ainsi que par les ontologies réside dans leur construction laborieuse, leur mise à jour et leur maintenance qui sont relativement coûteuses. De plus, il est difficile d'établir exactement des équivalences entre des concepts de langues différentes, surtout lorsque plus de trois langues sont en jeu.

Nous nous focalisons dans le cadre de notre projet de recherche sur les approches basées sur les corpus parallèles ou comparables, les lexiques bilingues ainsi que les ressources externes (ontologies multilingues).

Il importe de signaler que par rapport à notre contribution décrite dans le chapitre 4 (*cf.* page 109), l'évaluation de la qualité des règles d'association inter-langues s'est faite uniquement dans le domaine de la TAS en considérant des corpus parallèles et par la seule mesure du score BLEU. Dans le cadre de notre première orientation de recherche, nous suggérons d'étudier l'apport de ces règles inter-langues dans le domaine de la RI multilingue, et ce en fouillant des corpus parallèles bruités et des corpus comparables. Pour cela, nous proposons, dans ce qui suit, trois réflexions de recherche en RIM qui font appel à notre modèle de traduction décrit dans le chapitre 4 [Latiri *et al.*, 2010b, Latiri *et al.*, 2011]

## 3.1 Axe 1 : Fouille des corpus comparables pour la traduction d'une requête

En recherche d'information multilingue, l'idée la plus simple pour traduire une requête est d'utiliser conjointement deux outils, à savoir : (*i*) un système de traduction automatique pour traduire la requête dans la langue de la collection de documents ; et, (*ii*) un SRI monolingue pour l'interrogation [Gao *et al.*, 2006, Herbert *et al.*, 2011]. Toutefois, étant donné que la requête est généralement courte, les traductions résultantes sont souvent incorrectes, entraînant ainsi une perte de précision dans la recherche documentaire. D'autant plus, le traducteur ne fournissant qu'une unique traduction par terme de la requête, il diminue le rappel en ne mentionnant pas les synonymes de cette traduction [Dolamic and Savoy, 2010]. Cependant, ces lacunes ont permis de mettre en évidence les différentes étapes de traduction qui sont nécessaires pour améliorer les résultats. Ces étapes, décrites ci-dessous, peuvent se dérouler dans un ordre différent :

- Trouver les traductions potentielles des termes de la requête.
- Filtrer les traductions pour lever les ambiguïtés moyennant des métriques statistiques ou une approche de désambiguïsation.
- Étendre la requête pour élargir le contexte d'utilisation des termes.
- Pondérer les différentes propositions de traduction.

Dans le cadre de cet axe de recherche, nous proposons d'aborder chacune de ces étapes en considérant, comme ressources pour la traduction des requêtes : (*i*) les corpus comparables ; et, (*ii*) les ontologies multilingues.

Nous soutenons l'idée que l'extraction de lexiques bilingues à partir de corpus comparables peut être approchée comme un problème de recherche d'information. Dans cette représentation, la requête serait alors les mots à traduire et les documents retournés par le SRIM sont les candidats à la traduction de ce mot. Et de la même manière que les documents retournés sont ordonnés suivant leur adéquation avec la requête, les traductions candidates sont classées en fonction de leur pertinence par rapport au mot à traduire.

Ainsi, en partant de :
- une requête dans une langue source $l_S$, notée $Req_S$,
- une collection multilingue composée de corpus comparables abordant un domaine spécifique, notée $\mathcal{C}_\mathcal{M} = \{C_{l_1}, \ldots, C_{l_n}\}$,
- un corpus $C_{l_i}$ de la collection $\mathcal{C}_\mathcal{M}$ exprimé dans la langue cible $l_i$,
- une ontologie de domaine multilingue $\mathcal{O}_\mathcal{M}$,

la réflexion de recherche que nous proposons pour la traduction d'une requête peut être conduite en trois étapes, comme suit :

1. **Dériver à partir des corpus comparables un lexique bilingue pour trouver les traductions potentielles des termes d'une requête :** L'idée est de considérer des paires de corpus comparables $(l_S, l_C)$. Nous proposons de définir un nouveau processus d'extraction des règles d'association inter-langues afin de générer des corrélations entre des unités linguistiques, respectivement, dans les langues sources et cibles. Nous suggérons de représenter les documents des corpus comparables en langues source $l_S$ et cible $l_C$, à travers une structure laticielle (treillis de concepts formels) qui va mettre en relation les intensions et les extensions de chaque ensemble de termes fortement corrélés dans les corpus comparables et définir leur voisinage contextuel moyennant l'ontologie multilingue $\mathcal{O}_\mathcal{M}$. Ces structures laticielles seront par la suite projetées sur l'ontologie $\mathcal{O}_\mathcal{M}$ afin de sélectionner les concepts ayant les meilleurs sens respectifs aux termes des concepts formels. À partir de ces concepts désambiguïsés, il est possible de dériver des règles d'association inter-langues et inter-concepts. Nous soutenons l'idée que les termes de la requête à traduire, exprimée en langue source $l_S$, figurent dans les prémisses des règles inter-langues dérivées, sous forme de concepts désambiguïsés. De ce fait, leurs traductions potentielles sont représentées par les conclusions de ces règles d'association inter-langues. Il est possible également de générer les règles d'association inter-langues dans un sens bi-directionnel. Autrement dit, il s'agit de dériver les règles dont les prémisses sont dans la langue cible $l_C$ et les conclusions dans la langue source $l_S$. Ceci permet d'étendre les termes de l'index d'un document de la langue cible qui figurent dans les prémisses des règles d'association inter-langues par les concepts des conclusions dans la langue source et qui font partie de la requête $Req_S$. Ainsi, le résultat de cette étape est un ensemble de lexiques bilingues, construits à partir de corpus comparables et d'une ontologie multilingue.
2. **Étendre une requête :** La lacune majeure des requêtes d'un SRI est leur manque de contexte, souvent créateur d'ambiguïté. Ainsi, une autre manière d'améliorer les résultats est d'étendre la requête avec de nouveaux termes qui vont permettre de clarifier le concept caché derrière les termes [Baziz *et al.*, 2003]. Nous suggérons d'adapter pour la RIM, l'approche d'expansion automatique de requêtes par la base générique de règles d'association $\mathcal{MGB}$ que nous avons proposée dans [Latiri *et al.*, 2012b], et ce moyennant les règles d'association inter-concepts et inter-langues qui sont dérivées à partir de corpus comparables.

3. **Pondérer les différentes propositions de traduction :** Un système de recherche d'information multilingue n'a pas besoin de résoudre toutes les ambiguïtés du sens d'un terme pour traduire une requête, car le besoin d'information, même dans le cas de recherche monolingue, n'est pas formulé de manière précise dans une requête. Par conséquent, et dans un cas de doute, il est préférable de garder plusieurs traductions d'un terme sous peine de prendre le risque d'éliminer la bonne traduction. Par contre, il ne faut pas que le nombre de traductions d'un terme influence son poids dans la formulation de la requête. La solution que nous proposons est de diminuer au fur et à mesure le poids des multiples traductions d'un terme de la requête en fonction : (*i*) de la diminution de la valeur de la confiance de la règle d'association inter-langues valide, à partir de laquelle la traduction du terme en question est retenue ; et, (*ii*) d'une mesure de similarité qui lie le concept-sens du corpus comparable à son voisinage dans l'ontologie multilingue $\mathcal{O}_{\mathcal{M}}$.

La deuxième problématique de recherche que nous projetons d'aborder, moyennant les lexiques bilingues, concerne l'indexation dans le cadre de la RIM.

## 3.2 Axe 2 : Expansion d'un index multilingue par les lexiques bilingues

L'étape la plus importante de l'indexation est le choix des descripteurs représentant le contenu d'un document. Dans un contexte multilingue, cette étape devient plus complexe car elle passe obligatoirement par une étape de traduction pour représenter un document et une requête dans le même espace d'indexation. Dans la littérature, l'indexation multilingue est abordée principalement par l'utilisation de ressources externes, particulièrement les thésaurus et les ontologies [Potthast *et al.*, 2008, Roussey *et al.*, 2010, Knoth *et al.*, 2010, Ahmed *et al.*, 2011]. Dans cet axe de recherche, nous nous intéressons aux approches statistiques d'indexation multilingue impliquant des corpus comparables et des ontologies multilingues [McCarley and Roukos, 1998, Nie *et al.*, 2012].

En effet, il est possible par exemple de se poser la question de savoir comment retrouver un document écrit dans une langue cible $l_C$ à l'aide d'une requête exprimée dans une langue source $l_S$, sans passer par une étape de traduction ? Pour répondre à cette question, nous suggérons de déployer les règles d'association inter-langues. L'idée que nous proposons introduit un nouveau modèle de représentation des documents de la collection multilingue, basé sur les corrélations inter-langues extraites à partir d'un corpus comparable et d'une ontologie multilingue.

Nous partons d'une requête $Req_S$ formulée dans une langue source $l_S$ et d'une collection multilingue $\mathcal{C}_{\mathcal{M}} = \{C_{l_1}, \ldots, C_{l_n}\}$, avec $C_{l_i}$ un corpus de la collection $\mathcal{C}_{\mathcal{M}}$ exprimé dans la langue cible $l_i$. Pour chaque document $d$ de la collection $C_{l_i}$, son index dans la langue cible, noté $Index_{l_i}(d)$ est généré par un processus classique d'indexation. Cet index englobe les descripteurs d'un document dans la langue cible $l_i$ avec la pondération relative au schéma de pondération choisi lors de l'indexation. Ensuite, le processus d'extraction de lexiques bilingues est lancé sur les corpus comparables ($l_i$,$l_S$), afin de générer des corrélations entre des unités linguistiques de la langue cible $l_i$ dans laquelle est représenté l'index d'un document de la collection et celles de langue source $l_S$. Dans ce contexte, les corrélations inter-langues sont générées dans le sens contraire que celui considéré pour l'expansion de requêtes, *i.e.,* de la langue cible vers la langue source. L'index $Index_{li}(d)$ est par la suite étendu en utilisant le lexique bilingue, où chaque terme $t_i$ de l'index est étendu par le terme ou la séquence de termes qui apparaît dans une corrélation inter-langues contenant le terme $t_i$. Notons que le terme $t_i$ peut avoir autant de traductions que de corrélations possibles dans le lexique. La sélection et la pondération de ces traductions se feront en se basant sur une mesure de similarité qui les rapproche de leur voisinage contextuel, en les projetant sur l'ontologie du domaine associée au corpus comparable.

Cette nouvelle représentation de l'index des documents d'une collection multilingue permet ainsi une interrogation monolingue, avec une requête $Req_S$ exprimée en langue source et un corpus de documents en langue cible, dont les index des documents sont traduits moyennant les lexiques bilingues.

Toutefois, bien que cette orientation de recherche soit prometteuse par l'utilisation de nouvelles connaissances issues d'un processus d'ECT, nous relevons la limite relative à la rigidité de l'approche statistique et à la faible prise en compte du contexte du document. De ce fait, nous décrivons, dans la section qui suit, une autre proposition pour l'indexation conceptuelle multilingue, inspirée de l'approche d'indexation relative au contexte monolingue qui est décrite dans le chapitre 3 (*cf.* page 85).

## 3.3 Axe 3 : Vers la multilinguisation d'ontologies et l'indexation conceptuelle multilingue

Avec l'avènement du Web sémantique, des problèmes comme l'accessibilité dans d'autres langues des conceptualisations d'ontologies, décrites avec des terminologies monolingues, sont abordés en RIM et en TA [Knoth *et al.*, 2010, Ahmed *et al.*, 2011]. De ce fait, le bénéfice d'un accès multilingue, dans le contexte d'utilisation des ontologies, devient de plus en plus évident et indispensable pour certaines applications. Les méthodes visant à la multilinguisation d'ontologies se basent généralement sur des approches d'enrichissement. Ces approches permettent l'ajout de données multilingues à une ontologie donnée, pour en favoriser l'utilisation par des applications telles que la RI ou la TA, sans interférer avec la conceptualisation initiale de l'ontologie [Rouquet and Nguyen, 2009].

Dans la littérature, plusieurs approches ont été proposées pour la multilinguisation des ontologies. Ces approches peuvent être regroupées en trois familles :

1. **Approches par traduction :** Elles suggèrent la traduction vers la langue cible $l_C$. C'est le cas par exemple des méthodes décrites dans [Declerck *et al.*, 2006, Espinoza *et al.*, 2008] où les auteurs ont proposé de traduire les étiquettes exprimées en langue source $l_S$ de l'ontologie, directement vers d'autres langues cibles $l_C$. Les traductions possibles sont déterminées en consultant des ressources linguistiques (dictionnaires bilingues, bases lexicales, etc.). Ensuite, la liste des traductions est classée selon leur qualité en utilisant les voisinages dans la structure ontologique. Il importe de noter que les méthodes de désambiguïsation utilisant les ontologies sont intéressantes, mais doivent être appliquées de nouveau pour chaque langue cible $l_C$.
2. **Approches à base de correspondance avec des connaissances linguistiques externes :** Dans [Buitelaar *et al.*, 2006], les auteurs ont défini une trame de "sous-ontologie" linguistique, permettant de stocker la traduction d'un concept accompagnée de données morpho-syntaxiques. Il faut, pour chaque concept de l'ontologie source, instancier la trame dans la langue cible $l_C$ et la greffer au concept. Cette approche a pour inconvénient de rendre l'ontologie source bien plus complexe. De plus, l'instanciation et la greffe des sous-ontologies linguistiques doivent être faites pour chaque langue cible $l_C$ ; or à ce jour, aucune méthode automatisée n'a été proposée. En outre, une autre idée a été proposée dans [Peters *et al.*, 2007] où les auteurs ont développé un modèle qui intègre les informations terminologiques et linguistiques pour tenir compte de la diversité linguistique des différentes communautés. Ce modèle peut être utilisé en combinaison avec un modèle ontologique afin de relier les lexiques multilingues aux concepts ontologiques. L'objectif principal d'un tel modèle est de fournir de l'information multilingue aux ontologies [Montiel-Ponsoda *et al.*, 2011].

3. **Approches par correspondance avec les Wordnets :** D'autres approches basées sur les bases lexicales WordNet ont été proposées, tels que les travaux de Niles *et al.* [Niles and Pease, 2003] qui ont suggéré de relier l'ontologie à des WordNets. Leur idée est de générer une correspondance entre l'ontologie SUMO (Suggested Upper Merged Ontology) [29] et WordNet. Cette correspondance a pour objectif de rendre l'ontologie accessible à des humains et utilisable automatiquement par des applications traitant des textes. Elle comprend ainsi des relations de synonymie, d'hyperonymie et d'instanciation entre les concepts de l'ontologie et les *synsets* des WordNets. Bien que cette méthode ne permet pas la multilinguisation, elle est cependant intéressante puisqu'elle autorise l'ajout de données monolingues à une ontologie.

   Par ailleurs, dans [Vossen *et al.*, 2008], les auteurs ont proposé un environnent appelé KYOTO (Knowledge-Yielding Ontologies for Transition-Based Organization) pour le développement collaboratif d'une ontologie inter-langues et de sa correspondance avec des WordNets (Sept langues sont intégrées dans KYOTO). Ici encore, l'approche ne permet pas de traiter séparément les aspects de conceptualisation et de multilinguisation, puisque des experts de chaque langue doivent considérer les entrées des WordNets, pour proposer des liens vers les concepts existants ou vers de nouveaux concepts consensuels à insérer dans l'ontologie.

4. **Approches par correspondance avec un langage pivot :** Dans [Rouquet and Nguyen, 2009], les auteurs ont proposé une approche avec un langage pivot en vue de permettre l'accès multilingue à une ontologie, précédemment développée. L'objectif est de construire un lexique non ambigu pour l'ontologie, dans un langage pivot approprié, et d'utiliser ce lexique inter-langues comme portail vers les langues naturelles. Pour atteindre cet objectif, le langage pivot doit disposer d'un espace lexical autonome et non ambigu qui est mis en correspondance avec les étiquettes de l'ontologie par des affectations lexicales. Il doit également permettre la construction de syntagmes pour traiter les concepts portant des étiquettes "composées". Cette approche présente plusieurs avantages. D'une part, elle allège la tâche de désambiguïsation, qui est nécessaire pour le calcul du lexique pivot, mais pas pour l'ajout de nouvelles langues. Une fois le lexique pivot construit, l'ajout de nouvelles langues peut se faire par une simple acquisition de dictionnaires reliant la langue cible au langage pivot. D'autre part, la construction de ces ressources ne requiert pas un expert du domaine de l'ontologie et se résume en une tâche linguistique. De plus, la méthode est respectueuse de la conceptualisation proposée dans l'ontologie, car les affectations lexicales sont clairement distinguées des relations initialement présentes dans l'ontologie. La méthode a été adoptée dans le projet OMNIA pour permettre l'accès multilingue à une ontologie en utilisant cette approche par correspondance avec un langage pivot [Falaise *et al.*, 2010].

À l'issue de cette revue de la littérature, nous constatons que la problématique de multilinguisation d'ontologies couvre une tâche de conceptualisation qui consiste à maintenir la cohérence d'une ontologie multilingue, ainsi qu'une tâche linguistique pour la construction d'un lexique bilingue.

L'idée que nous suggérons dans cet axe de recherche est la suivante : partant d'une ontologie de domaine monolingue et d'un corpus parallèle du même domaine, il s'agit de proposer une approche de multilinguisation d'une ontologie en se basant sur l'extraction d'un lexique bilingue,

29. Il s'agit d'une "ontologie supérieure" qui contient des concepts génériques, abstraits, aptes à être spécialisés dans un grand nombre de domaines. Elle offre une structure et un ensemble de concepts généraux afin de permettre la construction d'ontologies de domaine.

illustrant des associations inter-langues entre les syntagmes nominaux (SNs) extraites à partir d'un corpus parallèle. Notre proposition est motivée par le fait qu'il a été prouvé, dans le contexte de la RI monolingue [Haddad, 2003], que l'intégration des SNs dans l'indexation, en plus des mono-termes, permet l'amélioration des performances d'un SRI. De même, dans [Ho *et al.*, 2006], les auteurs ont utilisé une approche d'ECT hybride qui fusionne les associations entre les paires de termes monolingues extraites statistiquement avec les relations sémantiques dérivées linguistiquement. Notons que l'extraction des SNs est réalisée en utilisant des patrons morpho-syntaxiques selon des règles grammaticales. Les auteurs ont montré expérimentalement que l'utilisation des SNs permet d'obtenir un gain de pertinence remarquable en RI.

Ainsi, nous proposons d'aborder la problématique en trois étapes, à savoir :

1. Définir un processus de fouille d'un corpus parallèle afin de deriver des règles d'association inter-langues entre les SNs en langue source $l_S$ et en langue cible $l_C$. L'interprétation de ces règles sera la même que celle adoptée dans le chapitre 4, où la conclusion de la règle est une traduction potentielle de la prémisse. Notre objectif est de construire un lexique bilingue de SNs. L'originalité vient du fait que l'approche de génération des ces règles d'association aura une dimension linguistique de part l'utilisation des patrons morpho-syntaxiques pour la sélection des SNs, en plus de sa dimension statistique héritée de la technique d'extraction de règles d'association [Agrawal and Skirant, 1994]. Il importe de noter que le lexique bilingue englobera également les règles d'association inter-langues entre termes, où les termes sélectionnés seront des substantifs communs.
2. Enrichir l'ontologie monolingue de domaine par le lexique bilingue de SNs construit lors de la première étape dans un objectif de multilinguisation de l'ontologie. Pour ce faire, nous suggérons d'étendre, dans un contexte multilingue, l'approche d'enrichissement d'ontologie par des règles d'association non-redondantes entre termes (*cf.* chapitre 3, page 85). Il est ainsi nécessaire, au niveau de cette étape, de définir une nouvelle mesure de similarité qui permettra de pondérer les relations entre les concepts dans l'ontologie multilingue.
3. Le résultat de l'enrichissement d'une ontologie monolingue par un lexique bilingue de SNs offre un modèle de représentation de connaissances multilingues d'un domaine donné. Partant de ce modèle, nous proposons de faire évoluer l'approche d'indexation conceptuelle, décrite dans le chapitre 3, pour la recherche d'information multilingue. Dans ce contexte, l'ontologie de domaine doit être accessible dans la langue source $l_S$ d'une requête ou dans la langue cible $l_C$ d'un document. Par ailleurs, une méthode de désambiguïsation des nouveaux concepts est aussi à définir en tenant compte de la dimension linguistique de l'approche.

Dans ce contexte, nous projetons d'explorer les ressources linguistiques disponibles à travers la plate-forme du centre commun de recherche de la Commission Européenne [30]. Nous allons exploité deux types de ressources, à savoir :

1. Le corpus parallèle multilingue JRC-Acquis [Steinberger *et al.*, 2006], qui est composé de la législation de l'UE des années 1950 jusqu'à présent. Ce corpus est disponible pour 231 paires de langues obtenues à partir de 22 langues officielles de l'UE. Le corpus JRC-Acquis est aligné au niveau de paragraphes spécialisés du domaine juridique et administratif. Le corpus parallèle sera utilisé pour extraire les règles inter-langues et inter-syntagmes.
2. Le thésaurus multilingue Eurovoc [31] [Fiser and Sagot, 2008, Steinberger *et al.*, 2011], spécialement construit à l'origine pour le traitement de l'information documentaire des ins-

30. http ://langtech.jrc.ec.europa.eu/Eurovoc.html.
31. http ://eurovoc.europa.eu/.

titutions de l'Union Européenne. Il couvre 21 domaines. Il s'agit des principales activités communautaires et politiques de l'UE : institutions, vie politique, relations internationales, droit, questions sociales, emploi et travail, environnement, agriculture, transport.... La structure d'Eurovoc comporte une classification hiérarchique et peut comprendre jusqu'à quatre niveaux de termes spécifiques.

Nous pensons que l'idée d'intégrer une ontologie multilingue dans un modèle d'indexation conceptuelle multilingue en RI peut permettre :

- La réalisation de recherches précises ou larges suivant les besoins de l'utilisateur.
- La génération automatique des traductions des structures sémantiques et d'une requête ainsi que la traduction partielle des documents.
- La mise à jour du vocabulaire du domaine par l'enrichissement de l'ontologie multilingue.
- L'aide à l'interrogation d'une collection de documents multilingues en naviguant dans l'ontologie du domaine exprimée dans la langue source $l_S$ de l'utilisateur.

# 4 Orientation 2 : Ouverture vers le domaine de l'Analyse des Réseaux Sociaux

En plus de l'ensemble des contributions dans le domaine de l'ECT à base d'AFC, nous avons proposé une nouvelle approche de fouille de graphes appelée AC-Miner [Douar *et al.*, 2011b, Douar *et al.*, 2011a]. Cette dernière tire son originalité du domaine de la programmation par contraintes (CSP) et plus précisément de sa technique de filtrage local, qui est la technique de consistance d'arc (Arc Consistency) [Liquiere, 2007]. Nous avons ainsi introduit la notion de biais de projection afin de proposer un opérateur similaire à l'isomorphisme de sous-graphes, mais ayant une complexité polynomiale et des contraintes relaxées [Douar *et al.*, 2011c].

Dans le cadre du présent projet de recherche, nous proposons l'utilisation de notre approche de fouille de graphes [Douar *et al.*, 2011a] ainsi que la base générique minimale de règles d'association $\mathcal{MGB}$ [Latiri *et al.*, 2012b] dans le domaine de l'Analyse de Réseaux Sociaux (ARS) [Al Hasan and Zaki, 2011]. Plus précisément, nous proposons de mener nos travaux selon deux axes de recherche : (*i*) l'extraction de fermés de cliques maximales et la base générique minimale de règles d'association entre cliques maximales pour la complétion de liens et la détection de communautés dans les réseaux sociaux ; et, (*ii*) l'utilisation de l'approche de fouille de graphes pour la prédiction de liens dans les réseaux sociaux dynamiques. Nous décrivons, dans ce qui suit, ces deux axes.

## 4.1 Axe 1 : Extraction de fermés de cliques maximales pour la complétion de liens et la détection de communautés dans les réseaux sociaux

L'étude des graphes de terrain tels que les réseaux sociaux a montré qu'ils partagent un certain nombre de propriétés [Scott, 2011]. En particulier, bien que leur densité soit en général très faible, leur degré de cohésion est souvent élevé et ils se comportent donc localement comme des *cliques* [Al-Naymat, 2008]. Cette particularité peut être expliquée par l'existence de communautés [Yan and Gregory, 2009]. Il s'agit de groupes de nœuds fortement connectés à l'intérieur, *i.e.,* chacun des nœuds est directement relié à tous les autres nœuds du groupe, mais avec peu de liens vers l'extérieur. Une formalisation du concept de *communautés* a été proposée dans Newman et Girvan [Newman and Girvan, 2004]. Ainsi, une décomposition en communautés est une partition de l'ensemble des nœuds qui maximise une fonction de qualité appelée *modularité* [Newman and Girvan, 2004]. Le principe de la modularité est qu'un bon

partitionnement d'un graphe implique un nombre d'arêtes intra-communautaires important et un nombre d'arêtes inter-communautaires faible [Newman and Girvan, 2004]. Trouver cette partition optimale est un problème NP-complet et de nombreuses heuristiques ont été proposées pour le résoudre [Fortunato, 2010].

Dans la littérature, depuis l'article fondateur de Girvan et Newman [Girvan and Newman, 2002], les approches de détection de communautés ont fait l'objet de nombreux travaux [Flake *et al.*, 2002, Newman and Girvan, 2004, Friggeri *et al.*, 2011]. La plupart d'entre eux consistent à déterminer une partition des sommets du graphe optimisant un certain critère de qualité d'un partitionnement, défini à partir de la structure du graphe. Un aperçu des critères utilisés est donné dans [Brandes *et al.*, 2007].

L'énumération de l'ensemble des cliques maximales d'un graphe est un problème ayant des applications dans de nombreux domaines tels que l'IA, la théorie de graphes et la fouille de données. La multiplicité des domaines d'application pour lesquels ce problème intervient a conduit à l'émergence de nombreux algorithmes [Makino and Uno, 2004, Pei *et al.*, 2005, Li *et al.*, 2007, Al-Naymat, 2008, Liu and Wong, 2008].

La problématique que nous abordons dans cet axe de recherche se situe au croisement de deux thématiques de recherche, à savoir : l'énumération d'un ensemble réduit de cliques maximales non redondantes et leur utilisation pour la détection des communautés dans les réseaux sociaux.

La contribution attendue à travers cet axe est double :

1. Dans un premier temps, nous proposons d'étudier le problème d'énumération des cliques maximales à partir d'un graphe $\mathcal{G}$ et de proposer un algorithme d'extraction, qui tient compte de la taille des grands graphes de terrains tel que les réseaux sociaux. Cette phase d'extraction posera le problème du nombre très élevé de cliques maximales énumérées. Ceci nous a inspiré l'idée de formaliser le problème d'extraction d'un noyau compact de cliques maximales non redondantes.

2. Dans un deuxième temps, nous proposons de formaliser l'ensemble des cliques maximales énumérées à partir d'un grand graphe $\mathcal{G}$ sous forme d'un *méta-contexte formel*, où les attributs sont, dans ce cas de figure, des cliques maximales [Gaume *et al.*, 2010]. Nous allons étudier par la suite l'algorithmique appropriée pour l'extraction des *fermés de cliques maximales*, qui sera une première formalisation afin de réduire le nombre de cliques maximales générées. L'idée est de se baser sur les fondements mathématiques de l'AFC [Ganter and Wille, 1999] et la sémantique des bases génériques de règles associatives [Balcázar, 2010, Latiri *et al.*, 2012b] pour introduire une nouvelle approche d'agrégation de cliques [Yan and Gregory, 2009] et aboutir à des *fermés de quasi-cliques maximales*.

Notre principale motivation émane du fait que la détection des communautés à partir d'un réseau social ne peut être réduite au problème de détection de cliques, et ce, pour plusieurs raisons. Tout d'abord, la plupart des communautés ne sont pas représentées par des cliques (l'exigence que chaque paire de sommets soit connectée est trop stricte dans la pratique). De plus, les cliques, même si elles sont maximales, peuvent être nombreuses (nombre exponentiel lié à la taille du réseau) et avec un fort taux de chevauchement. Cela ne correspond pas à la notion intuitive de communautés, surtout si les communautés sont définies pour être disjointes.

Une phase de formalisation est nécessaire pour une nouvelle définition d'un méta-contexte formel d'extraction, de la sémantique d'un fermé de cliques et de la relation entre l'extension et l'intention du dit fermé. Vient ensuite, l'étude de l'aspect algorithmique et heuristique de l'extraction des fermés de cliques maximales et de leur réduction, afin d'inférer des quasi-cliques plus représentatives des connexions existantes entre les nœuds d'un grand graphe $\mathcal{G}$.

À titre d'exemple, considérons la formalisation des cliques maximales à travers un méta-contexte formel, illustré dans la TABLE 1. Dans cet exemple, les attributs du contexte sont des cliques maximales alors que les objets représentent les différents nœuds figurants dans ces cliques.

| $\mathcal{R}$ | $\langle SMBC \rangle$ | $\langle SMCJ \rangle$ | $\langle SMBW \rangle$ |
|---|---|---|---|
| $M$ | × | × | × |
| $C$ | × | × | |
| $B$ | × | | × |
| $W$ | | | × |
| $J$ | | × | |
| $S$ | × | × | × |

TABLE 1 – Exemple d'un méta-contexte formel de cliques maximales.

En utilisant un algorithme d'extraction de motifs fermés, adapté au nouveau méta-contexte formel, nous dérivons les fermés de cliques maximales (de taille $\geq 2$) qui sont illustrés dans la TABLE 2.

| **Intention** | **Extension** |
|---|---|
| $\langle SMB\mathbf{C} \rangle, \langle SMB\mathbf{W} \rangle$ | $B, M, S$ |
| $\langle SM\mathbf{B}C \rangle, \langle SMC\mathbf{J} \rangle$ | $C, M, S$ |
| $\langle SM\mathbf{BC} \rangle, \langle SM\mathbf{BW} \rangle, \langle SM\mathbf{CJ} \rangle,$ | $M, S$ |

TABLE 2 – Les fermés de cliques maximales.

Nous constatons l'existence d'un lien de corrélation fort entre l'intention et l'extension d'un fermé de cliques maximales. L'extension traduit en effet la clique minimale partagée par les différents éléments d'un fermé de cliques. Si nous dérivons une base générique de règles d'association à partir des fermés de cliques maximales, nous retrouvons la règle d'association exacte, *i.e.,* de confiance égale à 1.0, suivante (*cf.* équation (1)) :

$$\langle SM\mathbf{BC} \rangle, \langle SM\mathbf{CJ} \rangle \Rightarrow \langle SM\mathbf{BW} \rangle \quad (1)$$

Notre objectif est d'exploiter cette association pour justifier formellement la fusion de chaque clique de la prémisse avec la clique de la conclusion, afin de générer une nouvelle clique à 5 nœuds et rajouter deux liens pertinents à savoir : le premier entre $C$ et $W$ et le deuxième entre $B$ et $J$. Par ailleurs, le lien entre $J$ et $W$ ne peut être établi puisqu'il ne figure dans aucun des fermés de cliques.

Il importe de souligner que l'idée proposée est dédiée aux graphes statiques. Ainsi, nous pensons qu'il serait intéressant d'étudier la possibilité d'utiliser notre approche de fouille de graphes [Douar *et al.*, 2011a] pour la prédiction de liens dans un réseau temporel dynamique.

## 4.2 Axe 2 : Fouille de graphes pour la prédiction de liens dans les réseaux sociaux

L'émergence récente de nouvelles applications utilisant des réseaux sociaux qui évoluent dans le temps a mis en évidence la nécessité de développer de nouveaux outils. Ces outils permettant

de modéliser et d'analyser les propriétés de réseaux dynamiques. Bien que plusieurs travaux de recherche aient été proposés dans ce sens [Scott, 2011], il existe un besoin réel de développer des méthodes efficaces pour l'analyse de réseaux dynamiques, modélisant des entités complexes reliées par des relations complexes.

Dans ce contexte, nous proposons d'étudier l'aspect dynamique des réseaux sociaux et ce en abordant la problématique de prédiction de liens dans un réseau temporel de réseaux sociaux [Chan *et al.*, 2009]. Informellement, la prédiction de liens consiste à prédire la formation d'un lien entre deux nœuds jamais connectés auparavant. Ceci peut être exprimé par la définition suivante [Chan *et al.*, 2009] :

**Définition 25** *Un réseau social est dit dynamique s'il peut être représenté par une séquence de graphes statiques* $G = \langle G_1, \ldots, G_T \rangle$ *où* $G_i$ *est le graphe du réseau social à l'instant* $i$*. Le problème de prédiction de liens consiste à prédire l'apparition de nouveaux liens dans le graphe* $G_{i+1}$ *à partir de l'analyse de la séquence de graphes statiques* $G$*.*

À court terme, il s'agit de déployer nos approches de fouille de graphes [Douar *et al.*, 2011b, Douar *et al.*, 2011a] pour analyser et prédire l'évolution des réseaux sociaux complexes et de leurs relations. L'objectif de cette recherche pourra être atteint par la réalisation des deux tâches suivantes :

1. **Découverte de patrons fréquents corrélés dans les réseaux dynamiques :** Les travaux les plus représentatifs de cette tâche sont ceux de Borgwardt *et al.* [Borgwardt *et al.*, 2006], proposés pour trouver des séquences de sous-graphes qui apparaissent fréquemment dans des états consécutifs d'un réseau social. Dans [Bringmann *et al.*, 2010], les auteurs ont utilisé une méthode similaire pour trouver des règles d'évolution, correspondantes à des paires de sous-graphes successifs se retrouvant fréquemment dans les séquences de sous-graphes. Par ailleurs, un nombre restreint de travaux se sont intéressés à la découverte de patrons corrélés. Nous citons la proposition de Chan *et al.* [Chan *et al.*, 2009] qui permet de trouver des régions d'un réseau dynamique contenant des liens dont la présence ou l'absence est temporellement corrélée. Alors que ce travail considère la corrélation entre des liens du réseau, nous proposons de nous intéresser à la corrélation entre sous-graphes fréquents. En effet, Ozaki et Ohkawa [Ozaki and Ohkawa, 2009] ont utilisé la corrélation entre les sous-parties d'une séquence de graphes pour filtrer les séquences ayant un grand nombre de sous-parties non corrélées. Cependant, contrairement à ce que nous suggérons comme proposition, ce travail n'utilise pas la corrélation pour identifier des liens de causalité dans un réseau social.
2. **Prédiction de liens dans les réseaux dynamiques :** Il s'agit ici de développer une méthode prédisant globalement l'état futur d'un réseau dynamique, en caractérisant son évolution par la prédiction de nouveaux liens, et ce à partir des sous-graphes fréquents découverts. Dans la littérature, des travaux ont été proposés pour la prédiction de liens futurs dans un réseau social à partir d'une séquence d'états antérieurs de ce réseau. Ainsi, O'Madadhain et al. [O'Madadhain *et al.*, 2005] prédisent la co-participation d'entités dans un événement futur à l'aide d'un modèle de classification appris avec les attributs des entités et les propriétés d'événements antérieurs. Par ailleurs, dans [Huang and Lin, 2009], les auteurs proposent une approche auto-régressive utilisant la séquence binaire de la présence ou l'absence d'un lien entre deux entités pour prédire la présence de ce lien au temps suivant. Toutefois, ces deux travaux ne tiennent pas compte de la dépendance entre les liens lors de la prédiction. Dans le cadre notre projet de recherche, nous nous intéressons particulièrement aux approches de prédiction de liens dans les réseaux sociaux qui se basent

sur la fouille de graphes [Tang and Liu, 2010, Al Hasan and Zaki, 2011, Silva *et al.*, 2012] permettant d'extraire les régularités d'un réseau temporel. Nous pensons que ces régularités serviront par la suite à prédire l'évolution de ce réseau dans le temps. La capacité de prédiction de la formation de liens dans le réseau temporel $G$, nous permettra de simuler l'évolution de celui-ci dans le temps.

Les principales contributions attendues, pour cet axe de recherche, sont doubles. D'une part, nous proposons l'utilisation d'approches de fouille de graphes basées sur la technique de consistance d'arc [Douar *et al.*, 2011c] pour identifier des phénomènes de corrélation et de récurrence dans les réseaux sociaux. D'autre part, nous envisageons de développer une nouvelle approche prédisant, pour l'ensemble du réseau, la formation de nouveaux liens.

# Table des figures

# Liste des tableaux

# Bibliographie


[Agrawal and Skirant, 1994] R. Agrawal and R. Skirant. Fast algorithms for mining association rules. In *Proceedings of the* $20^{th}$ *International Conference on Very Large Databases, VLDB 1994*, pages 478–499, Santiago, Chile, September 1994.

[Ahmed *et al.*, 2011] F. Ahmed, A. Nürnberger, and M. Nitsche. Supporting arabic cross-lingual retrieval using contextual information. In *Proceedings of the Second Information Retrieval Facility Conference, IRFC 2011*, volume 6653 of *LNCS*, pages 30–45, Vienna, Austria, June 2011. Springer-Verlag.

[Al Hasan and Zaki, 2011] M. Al Hasan and M. J. Zaki. A survey of link prediction in social networks. In *Social Network Data Analytics*, pages 243–275. Springer-Verlag, 2011.

[Al-Naymat, 2008] G. Al-Naymat. Enumeration of maximal clique for mining spatial co-location patterns. In *Proceedings of the* $6^{th}$ *IEEE/ACS International Conference on Computer Systems and Applications, AICCSA'08*, pages 126–133, Doha, Qatar, 2008. IEEE Computer Society.

[Aljlayl and Frieder, 2001] M. Aljlayl and O. Frieder. Effective Arabic-English Cross-Language Information Retrieval via Machine-Readable Dictionaries and Machine Translation. In *Proceedings of the ACM International Conference on Information and Knowledge Management, CIKM 2001*, pages 295–302, Atlanta, Georgia, USA, November 2001. ACM.

[Amirouche *et al.*, 2008] F. Boubekeur Amirouche, M. Boughanem, and L. Tamine. Exploiting association rules and ontology for semantic document indexing. In *Proceedings of the* $12^{th}$ *International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems, IPMU'08*, pages 464–472, Malaga, Espagne, June 2008.

[Andreasen *et al.*, 2009] T. Andreasen, H. Bulskov, P. Jensen, and T. Lassen. Conceptual indexing of text using ontologies and lexical resources. In *Proccedings of the* $8^{th}$ *International Conference on Flexible Query Answering Systems, FQAS 2009*, volume 5822 of *LNCS*, pages 323–332, Roskilde, Denmark, October 2009. Springer-Verlag.

[Armstrong, 1974] W. W. Armstrong. Dependency structures of database relationships. In *Proceedings of IFIP Congress, Geneva, Switzerland*, pages 580–583, September 1974.

[Ashrafi *et al.*, 2007] M. Z. Ashrafi, D. Taniar, and K. Smith. Redundant association rules reduction techniques. *International Journal Business Intelligence and Data Mining*, 1(2) :29–63, 2007.

[Ayres *et al.*, 2002] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu. Sequential pattern mining using a bitmap representation. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 429–435, Edmonton, Alberta, Canada, July 2002. ACM.

[Bachimont, 2000] B. Bachimont. Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. *Ingénierie des Connaissances : Evolutions récentes et nouveaux défis*, 1 :1–16, 2000.

[Balcázar, 2010] J. L. Balcázar. Redundancy, deduction schemes, and minimum-size bases for association rules. *Logical Methods in Computer Science*, 6(2) :1–33, 2010.

[Bastide *et al.*, 2000a] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal. Mining minimal non-redundant association rules using frequent closed itemsets. In *Proceedings of the* $1^{st}$ *International Conference on Computational Logic*, volume 1861 of *LNAI*, pages 972–986, London, UK, July 2000. Springer-Verlag.

[Bastide *et al.*, 2000b] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal. Mining frequent patterns with counting inference. *In ACM-SIGKDD Explorations*, 2(2) :66–75, December 2000.

[Baziz *et al.*, 2003] M. Baziz, M. Boughanem, and N. Nassr. La recherche d'information multilingue : désambiguïsation et expansion de requêtes basées sur wordnet. In *Proceedings of the International Symposium On Programming and Systems, ISPS'03*, pages 175–186, Alger, Algérie, mai 2003.

[Baziz *et al.*, 2004] M. Baziz, M. Boughanem, and N. Aussenac-Gilles. The use of ontology for semantic representation of documents. In *Proceedings of the Semantic Web and Information Retrieval Workshop at SIGIR 2004*, pages 38–45, Sheffield, UK, July 2004. ACM Press.

[Baziz *et al.*, 2005] M. Baziz, M. Boughanem, N. Aussenac-Gilles, and C. Chrisment. Semantic cores for representing documents in IR. In *Proceedings of the 2005 ACM Symposium on Applied Computing*, SAC'05, pages 1011–1017, New York, USA, 2005. ACM Press.

[Béchet *et al.*, 2009] N. Béchet, M. Roche, and J. Chauché. Towards the selection of induced syntactic relations. In *Proceedings of the* $31^{th}$ *European Conference on IR Research, ECIR 2009*, volume 5478 of *LNCS*, pages 786–790, Toulouse, France, 2009. Springer-Verlag.

[Ben Ahmed, 2007] M. Ben Ahmed. *Cognition entre Philosophie, Science et Technologie*. Centre de Publication Universitaire Tunisien, 2007.

[Ben Ghezaiel *et al.*, 2010] L. Ben Ghezaiel, C. Latiri, M. Ben Ahmed, and N. Gouider-Khouja. Enrichissement d'ontologie par une base générique minimale de règles associatives - application aux maladies neurologies : les dystonies. In *Actes de la COnférence en Recherche d'Infomations et Applications, CORIA 2010*, pages 289–300, Sousse, Tunisie, 2010.

[Ben Ghezaiel *et al.*, 2011] L. Ben Ghezaiel, C. Latiri, M. Ben Ahmed, and N. Gouider-Khouja. Un réseau proxémique pour la recherche d'information : Application aux maladies des dystonies. In *Actes du* $1^{er}$ *Symposium sur l'Ingénierie de l'Information Médicale, SIIM 2011*, pages 25–40, Toulouse, France, Juin 2011. IRIT Press.

[Ben Ghezaiel *et al.*, 2012] L. Ben Ghezaiel, C. Latiri, and M. Ben Ahmed. Conceptual indexing documents in ir based on ontology enrichment. In *Proccedings of the* $16^{th}$ *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, In Advances in Knowledge-Based and Intelligent Information and Engineering Systems M. Graña et al. (Eds.), KES 2012*, pages 1920–1931, San Sebastian, Spain, September 2012. IOS Press.

[Ben Yahia *et al.*, 2006] S. Ben Yahia, T. Hamrouni, and E. Mephu Nguifo. Frequent closed itemset based algorithms : A thorough structural and analytical survey. *ACM-SIGKDD Explorations*, 8(1) :93–104, June 2006.

[Ben Yahia *et al.*, 2009] S. Ben Yahia, G. Gasmi, and E. Mephu Nguifo. A new generic basis of factual and implicative association rules. *Intelligent Data Analysis*, 13(4) :633–656, 2009.

[Bendaoud *et al.*, 2008] R. Bendaoud, A. Napoli, and Y. Toussaint. Formal concept analysis : A unified framework for building and refining ontologies. In *Proceedings of* $16^{th}$ *International*

*Conference on the Knowledge Engineering : Practice and Patterns, EKAW 2008*, volume 5268 of *LNCS*, pages 156–171, Acitrezza, Italy, 2008. Springer-Verlag.

[Benz *et al.*, 2010] D. Benz, A. Hotho, and G. Stumme. Semantics made by you and me : Self-emerging ontologies can capture the diversity of shared knowledge. In *Proceedings of the* $2^{nd}$ *Web Science Conference, WebSci'10*, Raleigh, NC, USA, April 2010.

[Berry, 2008] M. W. Berry. *Survey of Text Mining II : Clustering, Classification, and Retrieval*. Springer-Verlag, 2008.

[Bhogal *et al.*, 2007] J. Bhogal, A. Macfarlane, and P. Smith. A review of ontology based query expansion. *Information Processing and Management*, 43(4) :866 – 886, 2007.

[Bo *et al.*, 2011] L. Bo, E. Gaussier, E. Morin, and A. Hazem. Degré de comparabilité, extraction lexicale bilingue et recherche d'information interlingue. In *Proccedings de la* $18^{ème}$ *Conférence sur le Traitement Automatique des Langues Naturelles, TALN 2011*, pages 211–222, Montpellier, France, Juin 2011.

[Bodner and Song, 1996] R. C. Bodner and F. Song. Knowledge-based approaches to query expansion in information retrieval. In *Proceedings of the 11th Biennial Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence, AI 1996, LNCS, volume 1081*, pages 146–158, Toronto, Ontario, Canada, May 1996. Springer-Verlag.

[Borgwardt *et al.*, 2006] K. M. Borgwardt, H. Kriegel, and P. Wackersreuther. Pattern mining in frequent dynamic subgraphs. In *Proceedings of the* $6^{th}$ *IEEE International Conference on Data Mining, ICDM 2006*, pages 818–822, Hong Kong, China, December 2006. IEEE Computer Society.

[Brandes *et al.*, 2007] U. Brandes, D. Delling, M. Höfer, M. Gaertler, R. Görke, Z. Nikoloski, and D. Wagner. On Finding Graph Clusterings with Maximum Modularity. In *Proceedings of the* $33^{rd}$ *International Workshop on Graph-Theoretic Concepts in Computer Science, WG'07*, LNCS. Springer, 2007.

[Brants *et al.*, 2007] T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean. Large language models in machine translation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL'07*, pages 858–867, Prague, Czech Republic, June 2007.

[Bringmann *et al.*, 2010] B. Bringmann, M. Berlingerio, F. Bonchi, and A. Gionis. Learning and predicting the evolution of social networks. *IEEE Intelligent Systems*, 25(4) :26–35, 2010.

[Brown *et al.*, 1993] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation : parameter estimation. *Computational Linguistics*, 19(2) :263–311, 1993.

[Buckley *et al.*, 1994] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic Query Expansion Using SMART : TREC-3. In *Proceedings of the* $3^{rd}$ *Text REtrieval Conference*, 1994.

[Buitelaar *et al.*, 2006] P. Buitelaar, M. Sintek, and M. Kiesel. A multilingual/multimedia lexicon model for ontologies. In *Proceedings of* $3^{rd}$ *European Semantic Web Conference, ESWC 2006*, volume 4011 of *LNCS*, pages 502–513, Budva, Montenegro, June 2006. Springer-Verlag.

[Carpineto and Romano, 1996] C. Carpineto and G. Romano. Information retrieval through hybrid navigation of lattice representations. *International Journal of Human-Computer Studies*, 45(5) :553 – 578, 1996.

[Carpineto and Romano, 2000] C. Carpineto and G. Romano. Order-theoretical ranking. *Journal of The American Society for Information Sciences*, 51 :587–601, 2000.

[Carpineto *et al.*, 1998] C. Carpineto, G. Romano, and F. U. Bordoni. Effective reformulation of boolean queries with concept lattices. In *Proceedings of the* $3^{rd}$ *International Conference on Flexible Query-Answering Systems*, pages 83–94. Springer-Verlag, 1998.

[Carpineto *et al.*, 2004] C. Carpineto, G. Romano, and F. U. Bordoni. Exploiting the potential of concept lattices for information retrieval with CREDO. *Journal of Universal Computer Science*, 10 :985–1013, 2004.

[Castells *et al.*, 2007] P. Castells, M. Fernandez, and D. Vallet. An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 19 :261–272, 2007.

[Chan *et al.*, 2009] J. Chan, J. Bailey, and C. Leckie. Using graph partitioning to discover regions of correlated spatio-temporal change in evolving graphs. *Intell. Data Anal.*, 13(5) :755–793, 2009.

[Chang, 2004] K. Y. Chang. Efficient sequential pattern mining by breadth-first approach. Master degree, National Taiwan University, 2004.

[Charlet *et al.*, 2005] J. Charlet, P. Laublet, and C. Reynaud. Web sémantique. *I3 (Information, Interaction, Intelligence)*, 4(1), 2005.

[Cimiano *et al.*, 2005] P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. *J. Artif. Intell. Res.*, 24 :305–339, 2005.

[Cole II *et al.*, 2003] R. J. Cole II, P. W. Eklund, and G. Stumme. Document retrieval for e-mail search and discovery using formal concept analysis. *Applied Artificial Intelligence*, 17(3) :257–280, 2003.

[Croft and Yufeng, 1994] B. Croft and J. Yufeng. An association thesaurus for information retrieval. In *Proceedings of the* $4^{th}$ *International Conference on Computer-Assisted Information Retrieval, RIAO'94, New York (USA)*, pages 146–161. CID Press, October 1994.

[Declerck *et al.*, 2006] T. Declerck, A. Gómez Pérez, O. Vela, Z. Gantner, D. Manzano, and D. Saarbrücken. Multilingual lexical semantic resources for ontology translation. In *Proceedings of the* $5^{th}$ *International Conference on Language Resources and Evaluation, LREC'06*, pages 1492–1495, Genoa, Italy, May 2006.

[Déjean *et al.*, 2002] H. Déjean, E. Gaussier, and F. Sadat. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the* $19^{th}$ *international conference on Computational linguistics - Volume 1, COLING'02*, pages 1–7, Taipei, Taiwan, 2002. Association for Computational Linguistics.

[Di-Jorio *et al.*, 2008] L. Di-Jorio, S. Bringay, C. Fiot, A. Laurent, and M. Teisseire. Sequential patterns for maintaining ontologies over time. In *Proceedings of the International Conference On the Move to Meaningful Internet Systems, OTM 2008*, volume 5332 of *LNCS*, pages 1385–1403, Monterrey, Mexico, November 2008. Springer-Verlag.

[Díaz-Galiano *et al.*, 2008] M. C. Díaz-Galiano, M. Á. García-Cumbreras, M. T. Martín-Valdivia, A. Montejo-Ráez, and L. A. Ure na López. Integrating MeSH Ontology to Improve Medical Information Retrieval. In *Proceedings of the* $8^{th}$ *Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Advances in Multilingual and Multimodal Information Retrieval*, volume 5152 of *LNCS*, pages 601–606, Budapest, Hungary, September 2008. Springer-Verlag.

[Dinh and Tamine, 2011] D. Dinh and L. Tamine. Combining global and local semantic contexts for improving biomedical information retrieval. In *Proceedings of the* $33^{rd}$ *European Conference on IR Research, ECIR 2011*, volume 6611 of *LNCS*, pages 375–386, Dublin, Ireland, April 2011. Springer-Verlag.

[Djouadi, 2011] Y. Djouadi. Extended galois derivation operators for information retrieval based on fuzzy formal concept lattice. In *Scalable Uncertainty Management*, volume 6929 of *LNCS*, pages 346–358. Springer-Verlag, 2011.

[Do, 2011] T. Do. *Extraction de corpus parallèle pour la traduction automatique depuis et vers une langue peu dotée*. PhD thesis, Université Joseph Fourier, Grenoble I, 2011.

[Dolamic and Savoy, 2010] L. Dolamic and J. Savoy. Retrieval effectiveness of machine translated queries. *Journal of the American Society for Information Science and Technology*, 61(11) :2266–2273, 2010.

[Dong and Pei, 2007] G. Dong and J. Pei. *Sequence Data Mining*. Springer-Verlag, 2007.

[Douar *et al.*, 2011a] B. Douar, M. Liquiere, C. Latiri, and Y. Slimani. FGMAC : Frequent subgraph mining with Arc Consistency. In *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2011, part of the IEEE Symposium Series on Computational Intelligence*, pages 112–119, Paris, France, April 2011. IEEE Computer Society.

[Douar *et al.*, 2011b] B. Douar, M. Liquiere, C. Latiri, and Y. Slimani. Graph-based relational learning with a polynomial time projection algorithm. In *Proceedings of the* $21^{st}$ *International Conference on Inductive Logic Programming, ILP 2011*, volume 7207 of *LNAI*, pages 96–112, Windsor Great Park, United Kingdom, July/August 2011. Springer-Verlag.

[Douar *et al.*, 2011c] B. Douar, M. Liquiere, C. Latiri, and Y. Slimani. Nouvelle approche de fouille de graphes AC-réduits fréquents. In *Actes des onzièmes journées francphones en Extraction et gestion des connaissances, EGC'2011*, volume RNTI-E-20 of *Revue des Nouvelles Technologies de l'Information*, pages 473–478, Brest, France, janvier 2011. Hermann-Éditions.

[Duan *et al.*, 2010] N. Duan, H. Sun, and M. Zhou. Translation model generalization using probability averaging for machine translation. In *Proceedings of the* $23^{rd}$ *International Conference on Computational Linguistics, COLING 2010*, pages 304–312, Beijing, China, August 2010.

[Ducrou and Eklund, 2007] J. Ducrou and P. W. Eklund. Searchsleuth : The conceptual neighbourhood of an web query. In *Proceedings of the Fifth International Conference on Concept Lattices and Their Applications, CLA 2007*, volume 331, Montpellier, France, October 2007. CEUR-WS.org.

[Espinoza *et al.*, 2008] M. Espinoza, A. Gómez-Pérez, and E. Mena. Enriching an ontology with multilingual information. In *Proceedings of* $5^{th}$ *European Semantic Web Conference, ESWC 2008*, volume 5021 of *LNCS*, pages 333–347, Tenerife, Spain, June 2008. Springer-Verlag.

[Faatz and Steinmetz, 2002] A. Faatz and R. Steinmetz. Ontology enrichment with texts from the www. In *Proceedings of the* $2^{nd}$ *ECML/PKDD-Workshop on Semantic Web Mining*, pages 20–34, Helsinki, Finland, August 2002.

[Falaise *et al.*, 2010] A. Falaise, D. Rouquet, D. Schwab, H. Blanchon, and C. Boitet. Ontology driven content extraction using interlingual annotation of texts in the OMNIA project. In *Proceedings of the workshop CLIA, at the* $23^{rd}$ *International Conference on Computational Linguistics, COLING 2010*, pages 13–22, Bejin, China, August 2010.

[Falk and Gardent, 2011] I. Falk and C. Gardent. Combining formal concept analysis and translation to assign frames and thematic role sets to french verbs. In *Proceedings of the* $8^{th}$ *International Conference on Concept Lattices and Their Applications, CLA 2011*, pages 223–238, Nancy, France, October 2011.

[Feldman *et al.*, 1996] R. Feldman, I. Dagan, and W. Kloegsen. Efficient algorithm for mining and manipultationg associations in texts. In *Proceedings of EMCRS96*, pages 949–954, Vienna, Austria, April 1996.

[Feldman *et al.*, 1998] R. Feldman, I.Dagan, and H. Hirsh. Mining Text Using Keyword Distributions. *J. Intell. Inf. Syst.*, 10(3) :281–300, 1998.

[Fiser and Sagot, 2008] D. Fiser and B. Sagot. Combining multiple resources to build reliable wordnets. In *Proceedings of the* $11^{th}$ *International Conference on Text, Speech and Dialogue, TSD 2008*, volume 5246 of *LNCS*, pages 61–68. Springer-Verlag, September 2008.

[Flake *et al.*, 2002] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee. Self-organization and identification of web communities. *IEEE Computer*, 35 :66–71, 2002.

[Fortunato, 2010] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5) :75 – 174, 2010.

[Friberg, 2007] K. Friberg. Query expansion using domain information in compounds. In *Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 1–4, New York, USA, April 2007. The Association for Computational Linguistics.

[Friggeri *et al.*, 2011] A. Friggeri, G. Chelius, and E. Fleury. Communautés : Arrêtons de ne compter que les arêtes. In *Proceedings des* $13^{èmes}$ *Rencontres Francophones sur les Aspects Algorithmiques de Télécommunications, AlgoTel*, Cap Estérel, France, 2011.

[Fung and McKeown, 1997] P. Fung and K. McKeown. A technical word- and term-translation aid using noisy parallel corpora across language groups. *Machine Translation*, 12(1-2) :53–87, 1997.

[Ganter and Stumme, 2003] B. Ganter and G. Stumme. Creation and merging of ontology top-levels. In *Proceedings of the* $11^{th}$ *International Conference on Conceptual Structures for Knowledge Creation and Communication, ICCS 2003*, volume 2746 of *LNCS*, pages 131–145, Dresden, Germany, July 2003. Springer-Verlag.

[Ganter and Wille, 1999] B. Ganter and R. Wille. *Formal Concept Analysis*. Springer-Verlag, 1999.

[Gao *et al.*, 2006] J. Gao, J. Y. Nie, and M. Zhou. Statistical query translation models for cross-language information retrieval. *ACM Trans. Asian Lang. Inf. Process.*, 5(4) :323–359, 2006.

[Gaume *et al.*, 2010] B. Gaume, E. Navarro, and H. Prade. A parallel between extended formal concept analysis and bipartite graphs analysis. In *Computational Intelligence for Knowledge-Based Systems Design*, volume 6178 of *LNCS*, pages 270–280. Springer, June 2010.

[Girvan and Newman, 2002] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12) :7821–7826, June 2002.

[Goeuriot *et al.*, 2009] L. Goeuriot, E. Morin, and B. Daille. Reconnaissance de critères de comparabilité dans un corpus multilingue spécialisé. In *Proceedings of the* $6^{th}$ *French Information Retrieval Conference, CORIA 2009*, pages 33–47, Presqu'île de Giens, France, May 2009. LSIS-USTV.

[Gonzalo *et al.*, 1998] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran. Indexing with wordnet synsets can improve text retrieval. In *Proceedings the COLING/ACL '98 Workshop on Usage of WordNet for Natural Language Processing*, 1998.

[Grahne and Zhu, 2003] G. Grahne and J. Zhu. Efficiently using prefix-trees in mining frequent itemsets. In *Proceedings of the ICDM 2003 Workshop on Frequent Itemset Mining Implementations, FIMI'03, Frequent Itemset Mining Implementations*, volume 90, Melbourne, Florida, USA), December 2003.

[Grefenstette, 1992] G. Grefenstette. Use of semantic context to produce term association lists for text retrieval. In *Proceedings of the* $15^{th}$ *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'92*, pages 89–97, Copenhagen, Denmark, June 1992. ACM Press.

[Grefenstette, 1994] G. Grefenstette. Corpus-derived first, second and third-order word affinities. In *Proceedings of the* $6^{th}$ *Congress of the European Association for Lexicography, EURALEX'94*, pages 279–290, Amsterdam, The Netherlands, 1994.

[Gruber, 1993] T. R. Gruber. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5 :199–220, June 1993.

[Guillet and Hamilton, 2007] F. Guillet and H. J. Hamilton. *Quality Measures in Data mining : Association Rule Interestingness Measures : Experimental and Theoretical Studies*, volume 43 of *Studies in Computational Intelligence*, pages 51–76. Springer-Verlag, 2007.

[Habash *et al.*, 2006] N. Habash, C. Mah, S. Imran, R. Calistri-Yeh, and P. Sheridan. Design, Construction and Validation of an Arabic-English Conceptual Interlingua for Cross-lingual Information Retrieval. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 06*, pages 107–112, Genoa, Italy, May 2006.

[Haddad *et al.*, 2000] H. Haddad, J. P. Chevallet, and M. F. Bruandet. Relations between terms discovered by association rules. In *Proceedings of the Workshop on Machine Learning and Textual Information Access in conjunction with the* $4^{th}$ *European Conference on Principles and Practices of Knowledge Discovery in Databases, PKDD 2000*, Lyon, France, September 2000.

[Haddad, 2003] H. Haddad. French noun phrase indexing and mining for an information retrieval system. In *Proccedings of the* $10^{th}$ *International Symposium on String Processing and Information Retrieval, SPIRE 2003*, volume 2857 of *LNCS*, pages 277–286, Manaus, Brazil, October 2003. Springer-Verlag.

[Hahn *et al.*, 2004] U. Hahn, K. G. Markó, M. Poprat, S. Schulz, J.Wermter, and P. Nohama. Crossing languages in text retrieval via an interlingua. In *Proceedings of* $7^{th}$ *International Conference on Computer-Assisted Information Retrieval, RIAO 2004*, pages 100–115, Avignon, France, April 2004. CID.

[Haiquan *et al.*, 2005] L. Haiquan, L. Jinyan, L. Wong, M. Feng, and Y. P. Tan. Relative risk and odds ration : A data mining perspective. In *Proceedings of the* $24^{th}$ *ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, PODS 2005*, pages 368–377, Baltimore, Maryland, USA, June 2005. ACM Press.

[Hamrouni *et al.*, 2005] T. Hamrouni, S. Ben Yahia, and Y. Slimani. Prince : An algorithm for generating rule bases without closure computations. In *Proceedings of the* $7^th$ *International Conference on Data Warehousing and Knowledge Discovery, DaWaK'05*, volume 3589 of *LNCS*, pages 346–355, Copenhagen, Denmark, August 2005. Springer-Verlag.

[Han *et al.*, 2000] J. Han, J. Pei, B. Mortazavi-asl, Q. Chen, U. Dayal, and M. Hsu. FreeSpan : Frequent pattern-projected sequential pattern mining. In *Proceedings of the* $6^{th}$ *International Conference on Knowledge Discovery and Data Mining*, pages 355–359, Boston, USA, August 2000.

[Haton *et al.*, 2006] J. P. Haton, C. Cerisara, D. Fohr, Y. Laprie, and K. Smaïli. *Reconnaissance automatique de la parole : du signal à son interprétation.* DUNOD, 2006.

[Hayashi *et al.*, 2010] K. Hayashi, H. Tsukada, K. Sudoh, K. Duh, and S. Yamamoto. Hierarchical phrase-based machine translation with word-based reordering model. In *Proceedings of the*

$23^{rd}$ *International Conference on Computational Linguistics, COLING 2010*, pages 439–446, Beijing, China, August 2010.

[Hazem and Morin, 2012a] A. Hazem and E. Morin. Qalign : A new method for bilingual lexicon extraction from comparable corpora. In *Proceedings of the* $13^{th}$ *International Conference, CICLing 2012*, volume 7182 of *LNCS*, pages 83–96. Springer-Verlag, March 2012.

[Hazem and Morin, 2012b] A. Hazem and E. Morin. Qalign : A new method for bilingual lexicon extraction from comparable corpora. In *Proceedings of the* $13^{th}$ *International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2012*, volume 7182 of *LNCS*, pages 83–96, New Delhi, India, March 2012. Springer-Verlag.

[Hazem *et al.*, 2011] A. Hazem, E. Morin, and S. P. Saldarriaga. Métarecherche pour l'extraction lexicale bilingue à partir de corpus comparables. In *Proceedings de la* $18^{th}$ *Conférence sur le Traitement Automatique des Langues Naturelles, TALN 2011*, pages 283–293, Juin-Juillet 2011.

[Herbert *et al.*, 2011] B. Herbert, G. Szarvas, and I. Gurevych. Combining query translation techniques to improve cross-language information retrieval. In *Proceedings of* $33^{rd}$ *European Conference on IR Research, ECIR 2011*, volume 6611 of *LNCS*, pages 712–715, Dublin, Ireland, April 2011. Springer-Verlag.

[Hernandez *et al.*, 2007] Nathalie Hernandez, Josiane Mothe, Claude Chrisment, and Daniel Egret. Modeling context through domain ontologies. *Information Retrieval*, 10(2) :143–172, 2007.

[Hersh *et al.*, 1994] W. R. Hersh, C. Buckley, T. J. Leone, and D. H. Hickam. OHSUMED : An Interactive Retrieval Evaluation and new large Test Collection for Research. In *Proceedings of the* $17^{th}$ *Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, Dublin, Ireland, July 1994. ACM Press.

[Ho *et al.*, 2006] B. Ho, D. T. B. Thuy, J. P. Chevallet, and M. F. Bruandet. A structured indexing model based on noun phrases. In *Proceedings of the* $4^{th}$ *International Confernce on Computer Sciences : Research, Innovation and Vision for the Future*, pages 81–89, Ho Chi Minh City, Vietnam, February 2006. IEEE.

[Hoang *et al.*, 2009] H. Hoang, P. Koehn, and A. Lopez. A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation, IWSLT'09*, Tokyo, Japan, December 2009.

[Hoser *et al.*, 2006] B. Hoser, A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Semantic network analysis of ontologies. In York Sure and John Domingue, editors, *The Semantic Web : Research and Applications*, volume 4011 of *LNAI*, pages 514–529, Heidelberg, June 2006. Springer-Verlag.

[Hotho *et al.*, 2003] A. Hotho, S. Staab, and G. Stumme. Ontologies improve text document clustering. In *Proceedings of the* $3^{rd}$ *IEEE International Conference on Data Mining, ICDM 2003*, pages 541–544, Melbourne, Florida, USA, December 2003.

[Hu *et al.*, 2009] J. Hu, G. Wang, F. H. Lochovsky, J-T. Sun, and Z. Chen. Understanding user's query intent with wikipedia. In *Proceedings of the* $18^{th}$ *International Conference on World Wide Web, WWW 2009*, pages 471–480, Madrid, Spain, April 2009. ACM Press.

[Huang and Lin, 2009] Zan Huang and Dennis K. J. Lin. The time-series link prediction problem with applications in communication surveillance. *INFORMS Journal on Computing*, 21(2) :286–303, 2009.

[Hull and Grefenstette, 1996] D. A. Hull and G. Grefenstette. Querying across languages : A dictionary-based approach to multilingual information retrieval. In *Proceedings of the* $19^{th}$ *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96*, pages 49–57, Zurich, Switzerland, August 1996. ACM.

[Hutchins, 2005] J. Hutchins. Example-based machine translation : a review and commentary. *Machine Translation*, 19(3-4) :197–211, 2005.

[Janssen, 2003] M. Janssen. Lexical translation and conceptual hierarchies. In *Proceedings of f the* $5^{th}$ *International Symposium on Language, Logic and Computation*, Tbilisi, Georgia, October 2003.

[Joho *et al.*, 2004] H. Joho, M. Sanderson, and M. Beaulieu. A study of user interaction with a concept-based interactive query expansion support tool. In *Proceedings of the 26th European Conference on Information Retrieval Research, ECIR 2004*, volume 2997 of *LNCS*, pages 42–56, Sunderland, UK, April 2004. Springer-Verlag.

[Jones *et al.*, 2000] K. S. Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval : development and comparative experiments. *Information Processing and Management*, 36(6) :779–840, 2000.

[Jorio *et al.*, 2007] L. Di Jorio, L. Abrouk, C. Fiot, D. Hérin, and M. Teisseire. Enrichissement d'ontologie basé sur les motifs séquentiels. In *Actes de la Plateforme AFIA 2007, Atelier Ontologies et gestion de l'hétérogénéité sémantique*, 2007.

[Jurafsky and Martin, 2008] D. Jurafsky and J. H. Martin. *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition.* Prentice Hall, 2008.

[Khan and Luo, 2002] L. Khan and F. Luo. Ontology construction for information selection. In *Proceedings of* $14^{th}$ *IEEE International Conference on Tools with Artificial Intelligence*, pages 122–127, Washington DC (USA), 4-6 2002.

[Kiliçaslan and Güner, 2011] Y. Kiliçaslan and E. S. Güner. Filtering machine translation results with automatically constructed concept lattices. In *Proceedings of the* $8^{th}$ *International Conference on Concept Lattices and Their Applications, CLA 2011*, pages 59–73, Nancy, France, October 2011.

[Kipper *et al.*, 2008] K. Kipper, A. Korhonen, N. Ryant, and M. Palmer. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42(1) :21–40, 2008.

[Knoth *et al.*, 2010] P. Knoth, T. Collins, E. Sklavounou, and Z. Zdrahal. Facilitating cross-language retrieval and machine translation by multilingual domain ontologies. In *Proceedings of the Workshop on Supporting eLearning with Language Resources and Semantic Data, at LREC 2010*, Malta, May 2010.

[Koehn and Hoang, 2007] P. Koehn and H. Hoang. Factored translation models. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 868–876, Prague, Czech Republic, June 2007.

[Koehn *et al.*, 2003] P. Koehn, F. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference*, pages 48–54, Edmonton, May-June 2003.

[Koehn *et al.*, 2007] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, and R. Zens. Moses : Open source toolkit for statistical machine translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics, Demonstation session*, june 2007.

[Koehn, 2004] P. Koehn. PHARAOH : A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In *Proceedings of* $6^{th}$ *Conference of The Association for Machine Translation In The Americas, AMTA'04*, volume 3265 of *LNCS*, pages 115–124, Washington, DC, USA, 2004. Springer-Verlag.

[Koester, 2006] B. Koester. Conceptual Knowledge Retrieval with FooCA : Improving Web Search Engine Results with Contexts and Concept Hierarchies. In *Proceedings of the* $6^{th}$ *IEEE International Conference on Data Mining, ICDM 2006*, volume 4065 of *LNCS*, pages 176–190, Hong Kong, China, December 2006. Springer-Verlag.

[Koolen and Kamps, 2011] M. Koolen and J. Kamps. Are semantically related links more effective for retrieval ? In *Proceedings of the* $33^{rd}$ *European Conference on IR Research, ECIR 2011*, volume 6611 of *LNCS*, pages 92–103, Dublin, Ireland, April 2011. Springer-Verlag.

[Kryszkiewicz, 2002] M. Kryszkiewicz. Concise representation of frequent patterns and association rules. Habilitation dissertation, Institute of Computer Science, Warsaw University of Technology, Warsaw, Poland, August 2002.

[Kum *et al.*, 2005] H. C. Kum, S. Paulsen, and W. Wang. Comparative study of sequential pattern mining models. In *Foundations of Data Mining and Knowledge Discovery, Studies in Computational Intelligence*, volume 6, pages 43–70. Springer-Verlag, 2005.

[Kumaran and Allan, 2008] G. Kumaran and J. Allan. Adapting information retrieval systems to user queries. *Information Processing and Management*, 44(6) :1838–1862, 2008.

[Kuramochi and Karypis, 2001] M. Kuramochi and G. Karypis. Frequent subgraph discovery. In *International Conference on Data Mining*, pages 313–320. IEEE Computer Society, 2001.

[Lafouge and Boukacem, 2004] T. Lafouge and B. Boukacem. Applications des lois informatiques en sciences de l'information : dualité, champ informatique d'usage et de production. *ISDM*, 17(165) :1–25, 2004.

[Latiri *et al.*, 2002] C. Latiri, S. Ben Yahia, and S. Elloumi. Conexion de galois floue : Extension et application au textmining. In *Proceedings of the* $9^{th}$ *International Conference IPMU'2002*, volume 3, pages 1407–1413, Annecy, France, July 2002.

[Latiri *et al.*, 2003a] C. Latiri, S. Ben Yahia, J. P. Chevallet, and A. Jaoua. Query expansion using fuzzy association rules between terms. In *Proceedings of the* $4^{th}$ *International Conference on Knowledge Discovery and Discrete Mathematics*, pages 231–241, Metz, France, September 2003.

[Latiri *et al.*, 2003b] C. Latiri, S. Ben Yahia, and G. Mineau. Conceptual Non-Redundant Association Rules Discovery : Application to Query Expansion. In *Proceedings of the First International Conference on Formal Concept Analysis, ICFCA03*, Darmstadt, Germany, February-March 2003.

[Latiri *et al.*, 2003c] C. Latiri, S. Ben Yahia, G. Mineau, and A. Jaoua. Découverte des règles associatives non redondantes : Application aux corpus textuels. In *Actes des troisièmes journées francophones en Extraction et Gestion des Connaissances EGC'2003, Revue d'Intelligence Artificielle*, volume 17, pages 131–144, Lyon, France, Janvier 2003.

[Latiri *et al.*, 2004] C. Latiri, J. P. Chevallet, S. Elloumi, and A. Jaoua. Une Extension de la Connexion de Galois Floue pour la Recherche d'Information. *I3 : Information - Intéraction - Intelligence*, 3(2) :73–116, Janvier 2004.

[Latiri *et al.*, 2005a] C. Latiri, W. Bellagha, and S. Ben Yahia. VIE-MGB : A Visual Interactive Exploration of Minimal Generic Basis of Association Rules. In *Proceedings of the* $3^{rd}$ *International Conference on Concept Lattices and their Applications, CLA'05*, pages 179–196, Olomouc, Czech Republic, September 2005.

[Latiri *et al.*, 2005b] C. Latiri, M. Mtir, and S. Ben Yahia. Méthode de construction d'ontologie de termes à partir du treillis d'Iceberg de Galois. In *Actes des cinquièmes journées francophones en Extraction et gestion des connaissances, EGC'2005*, volume RNTI-E-3 of *Revue des Nouvelles Technologies de l'Information*, pages 365–376, Paris, France, 2005. Cépaduès-Éditions.

[Latiri *et al.*, 2006] C. Latiri, L. Ben Ghezail, and M. Ben Ahmed. Fast-MGB : Nouvelle base générique minimale pour l'extraction de règles associatives. In *Actes des sixièmes journées francophones en Extraction et Gestion des Connaissances, EGC'2006*, volume RNTI-E-6 of *Revue des Nouvelles Technologies de l'Information*, pages 217–222, Lille, France, Janvier 2006. Cépaduès-Éditions.

[Latiri *et al.*, 2010a] C. Latiri, Y. Slimani, C. Nasri, and K. Smaïli. Extraction des séquences fermées fréquentes à partir de corpus parallèles : application à la traduction automatique. In *Actes des dixièmes journées francophones en Extraction et gestion des connaissances, EGC'2010*, volume RNTI-E-19 of *Revue des Nouvelles Technologies de l'Information*, pages 55–60, Hammamet, Tunisie, 2010. Cépaduès-Éditions.

[Latiri *et al.*, 2010b] C. Latiri, K. Smaïli, C. Lavecchia, and D. Langlois. Mining monolingual and bilingual corpora. *Intelligent Data Analysis*, 14(6) :663–682, 2010.

[Latiri *et al.*, 2011] C. Latiri, K. Smaïli, C. Lavecchia, D. Langlois, and C. Nasri. Phrase-based machine translation based on text mining and statistical language modeling techniques. *Proceedings of* $12^{th}$ *International Conference on Intelligent Text Processing and Computational Linguistics, CICLING'2011, in the International Journal of Computational Linguistics and Applications*, 2(1-2) :193–208, JAN-DEC 2011.

[Latiri *et al.*, 2012a] C. Latiri, L. Ben Ghezaiel, and M. Ben Ahmed. Proxemic conceptual network based on ontology enrichment for representing documents in ir. In *Proccedings of the* $18^{th}$ *International Conference on Knowledge Engineering and Knowledge Management, EKAW 2012, Volume 7603 of LNAI*, pages 72–86, Galway City, Ireland, October 2012. Springer-Verlag.

[Latiri *et al.*, 2012b] C. Latiri, H. Haddad, and T. Hamrouni. Towards an effective automatic query expansion process using an association rule mining approach. *Journal of Intelligent Information Systems*, 39(1) :209–247, 2012.

[Latiri, 2004] C. Latiri. Dérivation des règles d'association non redondantes à partir du texte et leur application pour l'expansion de requêtes en RI. Thèse de doctorat en informatique, École Nationale des Sciences de l'Informatique, Université de la Manouba (Tunisie), Avril 2004.

[Lavecchia *et al.*, 2007] C. Lavecchia, K. Smaïli, D. Langlois, and J. P. Haton. Using interlingual triggers for machine translation. In *Proceedings of the* $8^{th}$ *Annual Conference of the International Speech Communication Association, INTERSPEECH'07*, pages 2829–2832, Antwerp, Belgium, August 2007. ISCA.

[Lavecchia *et al.*, 2008] C. Lavecchia, K. Smaïli, and D. Langlois. Discovering phrases in machine translation by simulated annealing. In *Proceedings of the* $9^{th}$ *Annual Conference of the International Speech Communication Association, INTERSPEECH'08*, pages 2354–2357, Brisbane, Australia, September 2008. ISCA.

[Lavecchia, 2010] C. Lavecchia. *Les Triggers Inter-langues pour la Traduction Automatique Statistique*. PhD thesis, Université Nancy II, Juin 2010.

[Lesmo *et al.*, 2011] L. Lesmo, A. Mazzei, and Daniele D. Radicioni. Ontology based interlingua translation. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 6609 of *LNCS*, pages 1–12. Springer-Verlag, 2011.

[Levow *et al.*, 2005] G. Levow, D. W. Oard, and P. Resnik. Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management*, 41(3) :523 – 547, 2005.

[Li and Gaussier, 2010] B. Li and E. Gaussier. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the* $23^{rd}$ *International Conference on Computational Linguistics, COLING 2010*, pages 644–652, Beijing, China, August 2010. Tsinghua University Press.

[Li *et al.*, 2007] J. Li, G. Liu, H. Li, and L. Wong. Maximal biclique subgraphs and closed pattern pairs of the adjacency matrix : A one-to-one correspondence and mining algorithms. *IEEE Trans. on Knowl. and Data Eng.*, 19 :1625–1637, December 2007.

[Lin *et al.*, 2008] H. C. Lin, L. H. Wang, and S. M. Chen. Query expansion for document retrieval by mining additional query terms. *Information and Management Sciences*, 19(1) :17–30, 2008.

[Liquiere, 2007] M. Liquiere. Arc consistency projection : A new generalization relation for graphs. In *Proceedings of the* $15^{th}$ *International Conference on Conceptual Structures, ICCS 2007*, volume 4604 of *LNCS*, pages 333–346, Sheffield, UK, July 2007. Springer-Verlag.

[Liu and Wong, 2008] G. Liu and L. Wong. Effective pruning techniques for mining quasi-cliques. In *Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases - Part II, ECML PKDD '08*, pages 33–49, Antwerp, Belgium, September 2008. Springer.

[Liu *et al.*, 2009] H. Liu, J. Sun, and H. Zhang. *PostMining of Association Rules : Techniques for Effective Knowledge Extraction*, chapter V : Postprocessing for Rule Reduction using Closed Set. IGI Global, 2009.

[Lucchese *et al.*, 2006] C. Lucchese, S. Orlando, and R. Perego. Fast and memory efficient mining of frequent closed itemsets. *IEEE Trans. Knowl. Data Eng.*, 18(1) :21–36, 2006.

[Luxenburger, 1991] M. Luxenburger. Implications partielles dans un contexte. *Mathématiques, Informatique et Sciences Humaines*, 29(113) :35–55, 1991.

[Maedche and Staab, 2000] A. Maedche and S. Staab. Mining ontologies from text. In *Proceedings of the* $12^{th}$ *European Workshop on Knowledge Acquisition, Modeling and Management*, volume 1937. Springer-Verlag, 2000.

[Maedche *et al.*, 2002] A. Maedche, V. Pekar, and S. Staab. *Ontology Learning Part One - On Discovering Taxonomic Relations from the Web*, pages 301–322. Springer-Verlag, 2002.

[Makino and Uno, 2004] K. Makino and T. Uno. New algorithms for enumerating all maximal cliques. In Torben Hagerup and Jyrki Katajainen, editors, *Algorithm Theory - SWAT 2004*, volume 3111 of *LNCS*, chapter 23, pages 260–272. Springer, Berlin, Heidelberg, 2004.

[Marcu and Wong, 2002] D. Marcu and W. Wong. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2002*, pages 133–139, Philadelphia, PA, USA, July 2002. ACL.

[Markó *et al.*, 2005] K. Markó, S. Schulz, O. Medelyan, and U. Hahn. Bootstrapping dictionaries for cross-language information retrieval. In *Proceedings of the* $28^{th}$ *annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '05*, pages 528–535, Salvador, Brazil, August 2005. ACM.

[Masseglia *et al.*, 1998] F. Masseglia, F. Cathala, and P. Poncelet. The PSP Approach for Mining Sequential Patterns. In *Proceedings of the* $2^{nd}$ *European Symposium on Principles of Data*

*Mining and Knowledge Discovery*, pages 176–184, Nantes, France, September 1998. Springer-Verlag.

[Masseglia *et al.*, 2004] F. Masseglia, M. Teisseire, and P. Poncelet. Extraction de motifs séquentiels. problèmes et méthodes. *Ingénierie des Systèmes d'Information*, 9(3-4) :183–210, 2004.

[Masseglia *et al.*, 2009] F. Masseglia, P. Poncelet, and M. Teisseire. Efficient mining of sequential patterns with time constraints : Reducing the combinations. *Expert Systems with Applications*, 36(2, Part 2) :2677–2690, 2009.

[McCarley and Roukos, 1998] J. McCarley and S. Roukos. Fast document translation for cross-language information retrieval. In *Machine Translation and the Information Soup*, volume 1529 of *LNCS*, pages 150–157. Springer-Verlag, 1998.

[Mephu Nguifo, 1994] E. Mephu Nguifo. Une nouvelle approche basée sur le treillis de galois pour l'apprentissage de concepts. *Mathématiques, Informatique, Sciences Humaines*, 134 :19–38, 1994.

[Mihalcea and Moldovan, 2000] R. Mihalcea and D. Moldovan. Semantic indexing using wordnet senses. In *Proceedings of the ACL-2000 workshop on Recent Advances in Natural Language Processing and Information Retrieval*, pages 35–45, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[Müller *et al.*, 2004] H. M. Müller, E. E. Kenny, and P. W. Sternberg. Textpresso : An ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2(11) :1984–1998, 09 2004.

[Moha *et al.*, 2008] N. Moha, A. Rouane, P. Valtchev, and Y.l Guéhéneuc. Refactorings of Design Defects Using Relational Concept Analysis. In *Proceedings of the* $6^{th}$ *International Conference on Formal Concept Analysis, ICFCA 2008*, volume 4933 of *LNCS*, pages 289–304, Montreal, Canada, February 2008. Springer-Verlag.

[Mondary *et al.*, 2008] T. Mondary, S. Desprès, A. Nazarenko, and S. Szulman. Construction d'ontologies à partir de textes : la phase de conceptualisation. In *Proceedings of the* $19^{th}$ *French Knowledge Engineering Conference, IC 2008*, pages 87–98, Nancy, France, June 2008.

[Montiel-Ponsoda *et al.*, 2011] E. Montiel-Ponsoda, G. Aguado de Cea, A. Gómez-Pérez, and W.Peters. Enriching ontologies with multilingual information. *Natural Language Engineering*, 17(3) :283–309, 2011.

[Nagypál, 2005] G. Nagypál. Improving information retrieval effectiveness by using domain knowledge stored in ontologies. In *On the Move to Meaningful Internet Systems 2005 : OTM 2005 Workshops*, volume 3762 of *LNCS*, pages 780–789. Springer-Verlag, 2005.

[Nasri *et al.*, 2011] C. Nasri, K. Smaïli, and C. Latiri. Training statistical machine translation with multivariate mutual information. In *Proceedings of* $5^{th}$ *Language and Technology Conference*, pages 440–444, Poznan, Poland, November 2011.

[Nauer and Toussaint, 2010] E ; Nauer and Y. Toussaint. Crechaindo, un système itératif et interactif de classification par treillis de concepts, pour la recherche d'information sur le web. *Document numérique*, 13(1) :41–62, 2010.

[Navigli and Velardi, 2006] R. Navigli and P. Velardi. Ontology enrichment through automatic semantic annotation of on-line glossaries. In *Proceedings of* $15^{th}$ *International Conference, EKAW 2006, Podebrady, Czech Republic*, volume 4248 of *LNCS*, pages 126–140. Springer-Verlag, October 2006.

[Navigli, 2009] R. Navigli. Word sense disambiguation : A survey. *ACM Comput. Surv.*, 41 :1–69, February 2009.

[Neshatian and Hejazi, 2004] K. Neshatian and M. R. Hejazi. Text categorization and classification in terms of multiattribute concepts for enriching existing ontologies. In *Proceedings of the* $2^{nd}$ *Workshop on Information Technology and its Disciplines, WITID'04*, pages 43–48, Kish Island, Iran, February 2004.

[Newman and Girvan, 2004] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69, Feb 2004.

[Nie *et al.*, 1999] J. Y. Nie, M. Simard, P. Isabelle, and R. Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proceedings of the* $22^{nd}$ *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–81, Berkeley, CA, USA, August 1999. ACM.

[Nie *et al.*, 2012] J. Y. Nie, J. Gao, and G. Cao. Translingual mining from text data. In *Mining Text Data*, pages 323–359. Springer-Verlag, 2012.

[Nie, 2010] J. Y. Nie. *Cross-Language Information Retrieval.* Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2010.

[Nijssen and Kok, 2005] S. Nijssen and J. N. Kok. The Gaston Tool for Frequent Subgraph Mining. *Electron. Notes Theor. Comput. Sci.*, 127 :77–87, March 2005.

[Niles and Pease, 2003] I. Niles and A. Pease. Linking lexicons and ontologies : Mapping wordnet to the suggested upper merged ontology. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering, IKE 03*, pages 412–416, Las Vegas, USA, June 2003. CSREA Press.

[Oard, 1998] D. W. Oard. A comparative study of query and document translation for cross-language information retrieval. In *Proceedings of the* $3^{rd}$ *Conference of the Association for Machine Translation in the Americas, AMTA '98*, volume 1529 of *LNCS*, pages 472–483, Langhorne, PA, USA, October 1998. Springer-Verlag.

[Och and Ney, 2000] F. J. Och and H. Ney. Improved statistical alignment models. In *Proceedings of the* $38^{th}$ *Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Hong Kong, China, October 2000.

[Och and Ney, 2004] F. J. Och and H. Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4) :417–449, 2004.

[O'Madadhain *et al.*, 2005] J. O'Madadhain, J. Hutchins, and P. Smyth. Prediction and ranking algorithms for event-based network data. *SIGKDD Explorations*, 7(2) :23–30, 2005.

[Ortiz-Martinez, 2011] D. Ortiz-Martinez. *Advances in Fully-Automatic and Interactive Phrase-Based Statistical Machine Translation.* PhD thesis, Universidad Politécnica de Valencia, 2011.

[Ozaki and Ohkawa, 2009] T. Ozaki and T. Ohkawa. Discovery of correlated sequential subgraphs from a sequence of graphs. In *Proceedings of the* $5^{th}$ *International Conference on Advanced Data Mining and Applications, ADMA 2009*, volume 5678 of *LNCS*, pages 265–276, Beijing, China, August 2009. Springer-Verlag.

[Papineni *et al.*, 2001] K. Papineni, S. Roukos, T. Ward, and W-J. Zhu. Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the* $40^{th}$ *Annual Meeting on Association for Computational Linguistics, ACL'02*, pages 311–318, 2001.

[Parekh *et al.*, 2004] V. Parekh, J. Gwo, and T. W. Finin. Mining domain specific texts and glossaries to evaluate and enrich domain ontologies. In *Proceedings of the International Conference on Information and Knowledge Engineering, IKE'04*, pages 533–540, Las Vegas, Nevada, USA, June 2004. CSREA Press.

[Pasquier *et al.*, 1998] N. Pasquier, Y. Bastide, R. Touil, and L. Lakhal. Pruning closed itemset lattices for association rules. In *Proceedings of 14th International Conference Bases de Données Avancées, Hammamet, Tunisia*, pages 177–196, 26–30 October 1998.

[Pasquier *et al.*, 1999] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In *Proceedings of the* $7^{th}$ *International Conference on Database Theory, ICDT'99*, volume 1540 of *LNCS*, pages 398–416. Springer-Verlag, January 1999.

[Pasquier *et al.*, 2005] N. Pasquier, Y. Bastide, R. Taouil, G. Stumme, and L. Lakhal. Generating a condensed representation for association rules. *Journal of Intelligent Information Systems*, 24(1) :25–60, 2005.

[Pei *et al.*, 2000] Jian Pei, Jiawei Han, and Runying Mao. CLOSET : An Efficient Algorithm for Mining Frequent Closed Itemsets. In *Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 21–30, 2000.

[Pei *et al.*, 2001] J. Pei, J. Han, B. Mortazavi-asl, H. Pintoand Q. Chen, U. Dayal, and M. Hsu. PrefixSpan : Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In *Proceedings of the* $17^{th}$ *International Conference on Data Engineering*, pages 215–224, Washington, DC, USA, 2001. IEEE Computer Society.

[Pei *et al.*, 2005] Jian Pei, Daxin Jiang, and Aidong Zhang. On mining cross-graph quasi-cliques. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, KDD'05*, pages 228–238, Chicago, IL, USA, August 2005. ACM.

[Peters *et al.*, 2007] W. Peters, E. Montiel-Ponsoda, G. A. de Cea, and A. G. Perez. Localizing ontologies in owl. In *Proceedings of the OntoLex07 Workshop at the* $6^{th}$ *International Semantic Web Conference*, pages 13–22, Busan, Corea, November 2007.

[Peters *et al.*, 2012] C. Peters, M. Braschler, P. Clough, C. Peters, M. Braschler, and P. Clough. Cross-language information retrieval. In *Multilingual Information Retrieval*, pages 57–84. Springer-Verlag, 2012.

[Pfaltz and Taylor, 2002] J. L. Pfaltz and C. M. Taylor. Scientific knowledge discovery through iterative transformation of concept lattices. In *Proceedings of the Workshop on Discrete Applied Mathematics in conjunction with the* $2^{nd}$ *SIAM International Conference on Data Mining, SDM 2002*, pages 65–74, Arlington, Virginia, USA, April 2002.

[Polaillon *et al.*, 2007] G. Polaillon, M. Aufaure, B. LeGrand, and M. Soto. FCA for contextual semantic navigation and information retrieval in heterogeneous information systems. In *Proceedings of the* $18^{th}$ *International Workshop on Database and Expert Systems Applications, DEXA 2007*, pages 534–539, Regensburg, Germany, September 2007.

[Popov *et al.*, 2004] B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov, and A. Kirilov. Kim : a semantic platform for information extraction and retrieval. *Nat. Lang. Eng.*, 10 :375–392, September 2004.

[Potthast *et al.*, 2008] M. Potthast, B. Stein, and M. Anderka. A wikipedia-based multilingual retrieval model. In *Proceedings of the* $30^{th}$ *European Conference on IR Research, ECIR 2008*, volume 4956 of *LNCS*, pages 522–530, Glasgow, UK, 2008. Springer-Verlag.

[Priss and Old, 2007] U. Priss and L. Old. Bilingual word association networks. In *Conceptual Structures : Knowledge Architectures for Smart Applications*, volume 4604 of *LNCS*, pages 310–320. Springer-Verlag, 2007.

[Priss and Old, 2009] U. Priss and L. J. Old. Revisiting the potentialities of a mechanical thesaurus. In *Proceedings of the* $7^{th}$ *International Conference on Formal Concept Analysis, ICFCA 2009*, volume 5548 of *LNCS*, pages 284–298, Darmstadt, Germany, May 2009. Springer-Verlag.

[Qui and Frei, 1993] Y. Qui and H. P. Frei. Concept based query expansion. In *Proceedings of the* $16^{th}$ *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1993*, pages 160–169, Pittsburgh, PA, USA, June/July 1993. ACM Press.

[Rama and Borin, 2011] T. Rama and L. Borin. Estimating Language Relationships from a Parallel Corpus. A Study of the Europarl Corpus. In *Proceedings of the* $18^{th}$ *Nordic Conference of Computational Linguistics, NODALIDA 2011*, pages 161–167, Riga, Latvia, May 2011. Northern European Association for Language Technology (NEALT).

[Renders *et al.*, 2002] J. M. Renders, H. Déjean, and E. Gaussier. Assessing automatically extracted bilingual lexicons for clir in vertical domains : Xrce participation in the girt track of clef 2002. In *Proceedings of the Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002*, volume 2785 of *LNCS*, pages 363–371, Rome, Italy, September 2002. Springer-Verlag.

[Resnik, 1999] P. Resnik. Semantic similarity in a taxonomy : An information-based measure and its application to problems of ambiguity in natural language. In *Journal of Artificial Intelligence Research*, number 11, pages 95–130, 1999.

[Revuri *et al.*, 2006] S. Revuri, S. R. Upadhyaya, and P. S. Kumar. Using domain ontologies for efficient information retrieval. In *Proceedings of the* $13^{th}$ *International Conference on Management of Data*, pages 170–173, Delhi, India, December 2006. Tata McGraw-Hill Publishing.

[Riesa *et al.*, 2011] J. Riesa, A. Irvine, and D. Marcu. Feature-rich language-independent syntax-based alignment for statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011*, pages 497–507, Edinburgh, UK, July 2011. ACL.

[Rijsbergen, 1979] C.J. Van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.

[Robertson and Hull, 2000] S. E. Robertson and D. A. Hull. The trec-9 filtering track final report. In *Proceedings of the* $9^{th}$ *Text REtrieval Conference, TREC 9*, Gaithersburg, Maryland, November 2000.

[Robertson and Walker, 1994] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the* $17^{th}$ *annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR'94*, pages 232–241, Dublin, Ireland, 1994. Springer-Verlag.

[Rosenfeld, 1995] R. Rosenfeld. The cmu statistical language modeling toolkit and its use in the 1994 arpa csr evaluation. In *Proceeding of the Spoken Language Systems Technology Workshop*, pages 47–50, Austin, 1995.

[Rouquet and Nguyen, 2009] D. Rouquet and H. Nguyen. Multilinguïsation d'une ontologie par des corespondances avec un lexique pivot. In *Actes de la Conférence en Terminologie & Ontologie : Théories et Applications, TOTh'09*, Annecy, France, Juin 2009.

[Roussey *et al.*, 2010] C. Roussey, F. Harrathi, L. Maisonnasse, and S. Calabretto. Vers une approche statistique pour l'indexation sémantique des documents multilingues. In *Proceedings du XXVIIIème Congrès INFORSID*, pages 127–141, mai 2010.

[Rungsawang *et al.*, 1999] A. Rungsawang, A. Tangpong, P. Laohawee, and T. Khampachua. Novel query expansion technique using apriori algorithm. In *Proceedings of the* $8^{th}$ *Text REtrieval Conference, TREC 8*, pages 453–456, Gaithersburg, Maryland, November 1999.

[Ruthven, 2003] I. Ruthven. Re-examining the potential effectiveness of interactive query expansion. In *Proceedings of the* $26^{th}$ *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2003*, pages 213–220, Toronto, Canada, July/August 2003. ACM Press.

[Sadat, 2010] F. Sadat. Exploiting comparable corpora for cross-language information retrieval. In *PRICAI 2010 : Trends in Artificial Intelligence*, volume 6230 of *LNCS*, pages 662–667. Springer-Verlag, 2010.

[Salton and Buckely, 1988] G. Salton and C. Buckely. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5) :513–523, 1988.

[Salton and McGill, 1983] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[Sanderson, 1994] M. Sanderson. Word sense disambiguation and information retrieval. In *Proceedings of the* $17^{th}$ *Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 142–151. Springer-Verlag, 1994.

[Savoy, 2002] J. Savoy. Recherche d'information dans des corpus plurilingues. *Ingénierie des Systèmes d'Information*, 7(1-2) :63–93, 2002.

[Scott, 2011] J. Scott. Social network analysis : developments, advances, and prospects. *Social Network Analysis and Mining*, 1(1) :21–26, January 2011.

[Silva *et al.*, 2012] A. Silva, W. Meira, and M. J. Zaki. Mining attribute-structure correlated patterns in large attributed graphs. *PVLDB*, 5(5) :466–477, 2012.

[Singh and Husain, 2005] A. K. Singh and S. Husain. Comparison, selection and use of sentence alignment algorithms for new language pairs. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, ParaText'05, pages 99–106, Stroudsburg, PA, USA, 2005. ACL.

[Smalter *et al.*, 2008] A. Smalter, J. Huan, and G. Lushington. Structure-based pattern mining for chemical compound classification. In *Proceedings of the* $6^{th}$ *Asia Pacific Bioinformatics Conference*, volume 6 of *Series on Advances in Bioinformatics and Computational Biology*, pages 39–48, Kyoto, Japan, January 2008.

[Smucker *et al.*, 2007] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the* $16^{th}$ *International Conference on Information and Knowledge Management, CIKM 2007*, pages 623–632, Lisboa, Portugal, November 2007. ACM Press,.

[Song and Croft, 1999] F. Song and W. B. Croft. A general language model for information retrieval. In *Proceedings of the* $8^{th}$ *international conference on Information and knowledge management, CIKM'99*, pages 316–321, New York, NY, USA, 1999. ACM.

[Song *et al.*, 2007] M. Song, I. Song, X. Hu, and R. B. Allen. Integration of association rules and ontologies for semantic query expansion. *Data and Knowledge Engineering*, 63(1) :63 – 75, 2007.

[Srikant and Agrawal, 1996] R. Srikant and R. Agrawal. Mining sequential patterns : Generalizations and performance improvements. In *Proceedings of the* $5^{th}$ *International Conference on Extending Database Technology, EDBT'96*, volume 1057 of *LNCS*, pages 3–17, Avignon, France, March 1996. Springer-Verlag.

[Steinberger *et al.*, 2006] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, and D. Tufis. The jrc-acquis : A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the* $5^{th}$ *International Conference on Language Resources and Evaluation, LREC'06*, pages 2142–2147, Genoa, Italy, May 2006.

[Steinberger *et al.*, 2011] R. Steinberger, B. Pouliquen, M. A. Kabadjov, J. Belyaeva, and E. Van der Goot. Jrc-names : A freely available, highly multilingual named entity resource. In *Proceedings of the conference on Recent Advances in Natural Language Processing, RANLP 2011*, pages 104–110, Hissar, Bulgaria, September 2011.

[Stolcke, 2002] A. Stolcke. SRILM - an extensible language modeling toolkit. In *Procceding of the* $7^{th}$ *International Conference on Spoken Language Processing, ICSLP2002*, Denver, Colorado, USA, September 2002. ISCA.

[Stumme and Maedche, 2001] G. Stumme and A. Maedche. Fca-merge : Bottom-up merging of ontologies. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI 2001, Seattle, Washington, USA.*, pages 225–234, August 2001.

[Stumme *et al.*, 2002] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal. Computing Iceberg Concept Lattices with Titanic. *Journal on Knowledge and Data Engineering*, 2(42) :189–222, 2002.

[Stumme, 2005] G. Stumme. Ontology merging with formal concept analysis. In *Proceedings of the Dagstuhl Seminar on Semantic Interoperability and Integration.* IBFI, Schloss Dagstuhl, Germany, 2005.

[Sun *et al.*, 2006] R. Sun, C. Ong, and T. Chua. Mining dependency relations for query expansion in passage retrieval. In *Proceedings of the* $29^{th}$ *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2006*, pages 382–389, Seattle, Washington, USA, August 2006. ACM Press.

[Szathmary *et al.*, 2007] L. Szathmary, A. Napoli, and S. O. Kuznetsov. Zart : A multifunctional itemset mining algorithm. In *Proceedings of the Fifth International Conference on Concept Lattices and Their Applications, CLA 2007*, volume 331, Montpellier, France, October 2007.

[Talvensaari *et al.*, 2007] T. Talvensaari, J. Laurikkala, K. Järvelin, M. Juhola, and H. Keskustalo. Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Trans. Inf. Syst.*, 25(1), February 2007.

[Tang and Liu, 2010] L. Tang and H. Liu. Graph mining applications to social network analysis. In *Managing and Mining Graph Data*, volume 40 of *Advances in Database Systems*, pages 487–513. Springer-Verlag, 2010.

[Tangpong and Rungsawang, 2000] A. Tangpong and A. Rungsawang. Applying association rules discovery in query expansion process. In *Proceedings of the* $4^{th}$ *World Multi-Conference on Systemics, Cybernetics and Informatics, SCI 2000*, Orlando, Florida, USA, July 2000.

[Toumouh *et al.*, 2011] A. Toumouh, D. Widdows, and A. Lehireche. Parallel corpora and word-space models : using a third language as an interlingua to enrich multilingual resources. *International Journal of Information and Communication Technology*, 3(4) :299–313, 2011.

[Uno *et al.*, 2004] T. Uno, M. Kiyomi, and H. Arimura. LCM ver. 2 : Efficient Mining Algorithms for Frequent/Closed/Maximal Itemsets. In *Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations, FIMI'04*, volume 126, Brighton, UK, November 2004. CEUR-WS.org.

[Valarakos *et al.*, 2004] A. Valarakos, G. Paliouras, V. Karkaletsis, and G. Vouros. A name-matching algorithm for supporting ontology enrichment. In G. Vouros and T. Panayiotopoulos,

editors, *Methods and Applications of Artificial Intelligence*, volume 3025 of *LNCS*, pages 381–389. Springer-Verlag, 2004.

[Vallet *et al.*, 2005] D. Vallet, M. Fernández, and P. Castells. An ontology-based information retrieval model. In *The Semantic Web : Research and Applications*, volume 3532 of *LNCS*, pages 103–110. Springer-Verlag, 2005.

[Velardi *et al.*, 2001] P. Velardi, P. Fabriani, and M. Missikoff. Using text processing techniques to automatically enrich a domain ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems, FOIS '01*, pages 270–284, New York, NY, USA, October 2001. ACM.

[Ventresque *et al.*, 2008] A. Ventresque, S. Cazalens, P. Lamarre, and P. Valduriez. Improving interoperability using query interpretation in semantic vector spaces. In *The Semantic Web : Research and Applications*, volume 5021 of *LNCS*, pages 539–553. Springer-Verlag, 2008.

[Voorhees, 1993] E. M. Voorhees. Using wordnet to disambiguate word senses for text retrieval. In *Proceedings of the* $16^{th}$ *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1993*, pages 171–180, Pittsburgh, PA, USA, June/July 1993. ACM Press.

[Voorhees, 1994] Ellen M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the* $17^{th}$ *Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 61–69, Dublin, Ireland, July 1994. ACM Press.

[Vossen *et al.*, 2008] P. Vossen, E. Agirre, N. Calzolari, C. Fellbaum, S. Hsieh, C. Huang, H. Isahara, K. Kanzaki, A. Marchetti, M. Monachini, F. Neri, R. Raffaelli, G. Rigau, M. Tesconi, and J. VanGent. KYOTO : a system for mining, structuring and distributing knowledge across languages and cultures. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco, 2008. European Language Resources Association.

[Wang and Han, 2004] J. Wang and J. Han. BIDE : Efficient mining of frequent closed sequences. In *Proceedings of the* $20^{th}$ *International Conference on Data Engineering*, pages 79–90, Boston, USA, March/April 2004. IEEE Computer Society.

[Wang and Oard, 2005] J. Wang and D. W. Oard. Document and query expansion using side collections and thesauri. In *Proceedings of* $6^{th}$ *Workshop of the Cross-Language Evalution Forum, CLEF 2005*, volume 4022 of *LNCS*, pages 800–809, Vienna, Austria, September 2005. Springer-Verlag.

[Wang *et al.*, 2003] J. Wang, J. Han, and J. Pei. CLOSET+ : searching for the best strategies for mining frequent closed itemsets. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 236–245, Washington, DC, USA, August 2003.

[Wille, 1982] R. Wille. Restructuring lattice theory : an approach based on hierarchies of concepts. In I. Rival, editor, *Ordered sets*, pages 445–470, Dordrecht–Boston, 1982. Reidel.

[Wille, 1989] R. Wille. Knowledge acquisition by methods of formal concept analysis. In E. Diday, editor, *Data analysis, learning symbolic and numeric knowledge*, pages 365–380, New York–Budapest, 1989. Nova Science Publishers.

[Wörlein *et al.*, 2005] M. Wörlein, T. Meinl, I. Fischer, and M. Philippsen. A Quantitative Comparison of the Subgraph Miners MoFa, gSpan, FFSM, and Gaston. In *Proceedings of the* $9^{th}$ *European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD 2005*, volume 3721 of *LNCS*, pages 392–403, Porto, Portugal, October 2005. Springer-Verlag.

[Wu and He, 2010] D. Wu and D. He. A study of query translation using google machine translation system. In *Proceedings of the International Conference on Computational Intelligence and Software Engineering, CiSE*, pages 1–4, Wuhan, China, December 2010. IEEE.

[Wu and Palmer, 1994] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *Proceedings of the* $32^{nd}$ *annual meeting of the Association for Computational Linguistics*, pages 133–138, New Mexico, USA, June 1994.

[Yan and Gregory, 2009] B. Yan and S. Gregory. Detecting communities in networks by merging cliques. In *IEEE International Conference on Intelligent Computing and Intelligent Systems, ICIS 2009*, pages 832–836, November 2009.

[Yan *et al.*, 2003] X. Yan, J. Han, and R. Afshar. Clospan : Mining closed sequential patterns in large databases. In *Proceedings of the* $3^{rd}$ *International Conference on Data Mining*, pages 166–177, San Francisco, CA, USA, May 2003. SIAM.

[Yang and Yao, 2005] X. Yang and Y. Yao. Conceptual query expansion. In *Proceedings of the Atlantic Web Intelligence Conference, AWIC 2005*, volume 3528 of *LNCS*, pages 190–196, Lodz, Poland, June 2005. Springer-Verlag.

[Zaki and Hsiao, 2002] M.J. Zaki and C. Hsiao. CHARM : An efficient algorithm for closed association rule mining. In *Proceedings of the* $2^{nd}$ *SIAM International Conference on Data Mining, SDM 2002*, pages 457–473, Arlington, VA, USA, April 2002.

[Zaki, 2001] M. J. Zaki. Spade : An efficient algorithm for mining frequent sequences. *Machine Learning*, 42 :31–60, 2001.

[Zaki, 2004] M. J. Zaki. Mining non-redundant association rules. *Data Mining and Knowledge Discovery*, 9(3) :223–248, 2004.

[Zens *et al.*, 2002] R. Zens, F. J. Och, and H. Ney. Phrase-based statistical machine translation. In *Proceedings of the* $25^{th}$ *Annual German Conference on AI : Advances in Artificial Intelligence*, KI '02, pages 18–32, London, UK, 2002. Springer-Verlag.

[Zhang *et al.*, 2006] G. Zhang, A. D. Troy, and K. Bourgoin. Bootstrapping ontology learning for information retrieval using formal concept analysis and information anchors. In *Proceedings of the* $14^{th}$ *International Conference on Conceptual Structures, Aalborg University, Denmark*, July 2006.

[Zhang *et al.*, 2009] H. Zhang, M. Zhang, H. Li, and C. L. Tan. Fast Translation Rule Matching for Syntax-based Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1037–1045, Singapore, August 2009. ACL and AFNLP.

[Zhong *et al.*, 2012] N. Zhong, Y. Li, and S. Wu. Effective pattern discovery for text mining. *IEEE Trans. Knowl. Data Eng.*, pages 30–44, 2012.