Accès à l'information : vers une hybridation fouille de données et traitement automatique des langues

Thierry Charnois

▶ **To cite this version:**

Thierry Charnois. Accès à l'information : vers une hybridation fouille de données et traitement automatique des langues. Traitement du texte et du document. Université de Caen, 2011. tel-00657919

HAL Id: tel-00657919
https://theses.hal.science/tel-00657919

Submitted on 13 Feb 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Accès à l'information : vers une hybridation fouille de données et traitement automatique des langues

Habilitation à diriger des recherches

Thierry CHARNOIS

GREYC - CNRS UMR 6072
Université de Caen
Campus Côte de Nacre
Bd Maréchal Juin
F-14032 Caen Cédex France

Thierry.Charnois@unicaen.fr

soutenue le 1er décembre 2011

Composition du jury :

Bruno Crémilleux, Professeur, Université de Caen, examinateur
Béatrice Daille, Professeur, Université de Nantes, examinatrice
Patrice Enjalbert, Professeur, Université de Caen, examinateur
Brigitte Grau, Professeur, ENSIIE Evry, rapporteure
François Jacquenet, Professeur, Université Saint-Etienne, examinateur
Pascal Poncelet, Professeur, Université Montpellier 2, rapporteur
Pierre Zweigenbaum, Directeur de Recherches CNRS, LIMSI, rapporteur

Table des matières

1 Introduction **1**
- 1.1 Parcours de recherche . . . 1
- 1.2 Positionnement . . . 4
- 1.3 Plan du mémoire . . . 4

2 Compréhension automatique et graphes conceptuels **7**
- 2.1 Graphes conceptuels comme formalisme de représentation de la langue . . . 8
- 2.2 Expériences . . . 9
  - 2.2.1 Positionnement . . . 10
  - 2.2.2 Fondements linguistiques . . . 11
  - 2.2.3 Information temporelle et graphes conceptuels . . . 14
  - 2.2.4 Système développé . . . 15
- 2.3 Bilan . . . 17

3 Analyse linguistique pour l'accès à l'information **19**
- 3.1 Extraction d'information pour la recherche intra-documentaire . . . 19
  - 3.1.1 Présentation du projet GéoSem . . . 20
  - 3.1.2 Analyse spatiale et temporelle . . . 21
  - 3.1.3 Analyse discursive . . . 23
  - 3.1.4 Bilan . . . 25
- 3.2 Extraction d'information en biomédical . . . 26
  - 3.2.1 Le projet Bingo . . . 26
  - 3.2.2 Reconnaissance d'entités nommées biologiques . . . 27
  - 3.2.3 Notre système . . . 27
- 3.3 Analyse d'opinion . . . 29
  - 3.3.1 Projet Ontopitex en cours . . . 29
  - 3.3.2 Extraction de constituants détachés axiologiques . . . 31
  - 3.3.3 Annotation manuelle pour l'évaluation et poursuites . . . 33
  - 3.3.4 Autres poursuites . . . 34
- 3.4 Bilan . . . 34

4 Fouille de données pour le TAL **35**
- 4.1 Indices linguistiques pour la fouille : les défis DEFT . . . 37
- 4.2 Contraintes et fouille récursive pour l'extraction de relation entre entités nommées . . . 39
  - 4.2.1 Méthode proposée pour l'apprentissage des patrons linguistiques . . . 41

4.2.2 Discussion 45
4.3 Exploration de motifs pour l'apprentissage de patrons 46
4.3.1 Motifs séquentiels d'itemsets 46
4.3.2 Apprentissage et exploration de patrons 47
4.4 Les motifs LSR pour la reconnaissance d'entités nommées 50
4.4.1 Aperçu de la méthode 50
4.4.2 Motifs LSR : un nouveau type de motif 51
4.4.3 Utilisation des motifs LSR pour la détection d'entités nommées . . . 53

5 Bilan et perspectives **59**

Annexe **63**

Bibliographie **69**

Chapitre 1

Introduction

1.1 Parcours de recherche

Mes activités de recherche concernent le TAL – traitement automatique des langues – et l'accès à l'information par des méthodes d'analyse linguistique, ainsi que l'utilisation de méthodes de fouille de données en TAL. Elles s'étendent sur une douzaine d'années depuis la soutenance de ma thèse préparée au LIPN et soutenue en 1999, et se sont poursuivies au sein du GREYC. Beaucoup de mes travaux ont été réalisés en collaboration avec des collègues de différents laboratoires dans le cadre de projets de recherche labellisés, ou informels, et souvent interdisciplinaires. Mon travail de recherche bénéficie fortement de ces multiples rencontres et collaborations.

Après avoir suivi le DEA informatique et linguistique de l'université Paris-Sorbonne où j'ai été initié à la recherche par Jean-Pierre Desclés et Thierry Chanier, mon intérêt pour l'informatique et la linguistique m'a naturellement conduit vers un sujet de thèse en TAL. J'ai eu le plaisir d'être accueilli au LIPN dans le cadre d'un projet de recherche [1] pour une thèse dirigée par Brigitte Biébow, Daniel Kayser et Sylvie Szulman. Le sujet portait sur la « détection d'anomalies sémantiques dans les textes de spécification », un sujet directement lié à la représentation des connaissances pour la compréhension automatique des textes. Sous forme d'une collaboration informelle avec Delphine Battistelli (université Paris-Sorbonne) et Tassadit Amghar (université d'Angers), ce travail sur la sémantique de la langue s'est poursuivi après ma thèse pour prendre en compte la dimension temporelle dans la représentation des textes narratifs et a montré l'intérêt d'utiliser le formalisme des graphes conceptuels et sa mise en œuvre sur cette problématique.

Lors de mon arrivée au GREYC comme enseignant chercheur en 2001, j'ai naturellement intégré l'équipe DoDoLa – Données Document Langue – autour des travaux menés dans l'axe intitulé « sémantique et TAL ». De façon plus générale, l'équipe DoDoLa s'est construite sur une problématique commune de recherche autour des questions de production, d'extraction, d'accès et de diffusion de l'information à partir de données, que celles-ci soient sous une forme structurée ou non. J'y ai poursuivi des activités en TAL dans la li-

1. Projet CNRS / Cnet-France Télécom numéro 931B143 intitulé « Aide à la rédaction de spécifications formelles, passage de l'informel au formel » auquel j'ai participé à partir de novembre 1994 et jusqu'à la fin du projet en novembre 1996.

gnée des travaux sur la compréhension de textes mais dans un contexte nouveau. En effet, la prolifération des données textuelles concomitante au développement du web a fait émerger d'importants besoins en extraction et en recherche d'information dans les textes. Ce contexte m'a amené à m'intéresser à la problématique générale de l'accès à l'information pertinente dans des corpus volumineux. La dynamique de l'équipe sur ce sujet et divers projets m'ont permis d'avoir des collaborations avec des linguistes pour mettre au point ou expérimenter des modèles linguistiques (par exemple l'encadrement du discours), des géographes et des biologistes pour valider des résultats d'analyse (comme la reconnaissance d'entités nommées biologiques dans les textes). Patrice Enjalbert m'a ainsi proposé de participer au projet GéoSem [2], un projet visant à développer des méthodes d'analyse sémantique pour la recherche documentaire appliquée à des corpus de géographie humaine. D'autre part, j'ai été sollicité par Bruno Crémilleux, coordonnateur du projet Bingo [3] pour m'investir dans ce projet qui portait sur la thématique des bases de données inductives et le développement de méthodes d'extraction de connaissances à partir de base de données hétérogènes, avec des applications dans le domaine de la biologie moléculaire. L'axe TAL du projet auquel j'ai participé en collaboration avec Nicolas Durand et Jiri Kléma concernait l'extraction d'informations textuelles pour les intégrer en tant que connaissances dans les processus de fouille de données. Les résultats obtenus dans ce projet ont permis de définir et de concevoir le projet Bingo*2*, accepté par l'ANR en 2007. Enfin, toujours sur le volet TAL, je suis aussi membre du projet Ontopitex [4], actuellement en cours, sur la thématique de l'analyse d'opinion. Avec Agata Jackiewicz (Paris 4) et Stéphane Ferrari (GREYC) je m'occupe plus spécifiquement de la mise au point et de l'expérimentation de patrons permettant d'extraire des constituants détachés porteurs de jugement ; un volet apprentissage de ce travail vise à découvrir automatiquement les patrons (*cf. infra*).

L'expérience acquise sur les projets cités précédemment a montré que l'acquisition manuelle des ressources et modèles est un processus lourd et coûteux, et de plus ces ressources sont peu voire pas génériques. Notre participation aux défis fouille de textes (éditions DEFT 2006–2008), et des réflexions menées avec Bruno Crémilleux et Jiri Kléma suite au projet Bingo, nous ont ainsi conduit à étudier l'apport de méthodes d'apprentissage automatique en TAL. Cette réflexion s'est formalisée avec une collaboration internationale avec Jiri Klema (Czech Technical University, Prague) sur le projet "Heterogeneous Data Fusion for Genomic and Proteomic Knowledge Discovery" [5] qui a porté sur le croisement de diffé-

2. Le projet GéoSem (traitements sémantiques pour l'information géographique), soutenu par le CNRS dans le cadre du programme interdisciplinaire « Société de l'information », a réuni les laboratoires ESO (géographie, Caen), ERSS (linguistique, Toulouse 2), LIUPPA et IKER (informatique et linguistique, Pau), GREYC (Caen) et s'est déroulé de 2002 à 2006.

3. Le projet Bingo (Bases de données INductives et GénOmique), a été réalisé dans le cadre de l'ACI masse de données en collaboration avec les laboratoires CGMC (CNRS UMR 5534, Lyon), EURISE (Laboratoire Hubert Curien - UMR CNRS 5516, Saint-Etienne) et LIRIS (CNRS UMR 5205, Lyon), de 2004 à 2007. `http://www.info.unicaen.fr/~bruno/bingo/`

4. Le projet ANR Ontopitex (Modèles linguistiques et ontologies. Extraction informatique et caractérisation d'opinions et de jugements d'évaluation dans les textes) a pour partenaires les laboratoires LaLIC (Université Paris IV), CRISCO (Caen), GREYC (Caen) et les sociétés Noopsis (Caen) et TecKnowMetrix (Grenoble) et se déroule de fin 2009 à fin 2012. `https://ontopitex.greyc.fr/`

5. Projet PHC Barrande : bilateral Programme Hubert Curien project "Heterogeneous Data Fusion for Genomic and Proteomic Knowledge Discovery", projet impliquant les laboratoires CGMC (Lyon), GREYC (Caen) et le Department of Cybernetics, Czech Technical University (Prague, République tchèque). De

rentes sources d'information autour du transcriptome pour la découverte de connaissances en biologie, comme par exemple les groupes de synexpression. Ce travail a été approfondi avec le projet BINGO*2*[6].

Ces expériences m'ont motivé pour déposer un projet de recherche autour de l'utilisation de méthodes de fouille de données pour le TAL dans le cadre d'un congé de recherche en 2009. Grâce à son obtention (6 mois), j'ai pu intensifier mes travaux dans cette voie et ainsi acquérir des compétences en fouille. Simultanément, j'ai été responsable de la tâche « Learning information extraction rules » du projet BINGO*2*. J'ai aussi eu l'opportunité de co-« encadrer » deux chercheurs post-doctorants (Marc Plantevit[7], puis Peggy Cellier[8]) sur cette problématique. Ce contexte a été particulièrement enrichissant et stimulant pour développer des activités de recherche sur l'intégration de méthodes de fouille en TAL et une dynamique sur cette nouvelle activité s'est mise en place dans notre équipe. Ainsi, pour compléter nos travaux en TAL cités précédemment, nous avons conçu des méthodes de fouille de données séquentielles pour la découverte de patrons linguistiques pour les tâches de reconnaissance d'entités nommées, d'extraction de relation entre entités nommées, ou encore, d'acquisition de constituants porteurs de jugement ou de qualification. Le succès de cette activité est aussi visible à travers le soutien de la région Basse-Normandie qui permet de renforcer cette activité pour l'année 2011-2012 avec l'arrivée de deux nouveaux chercheurs post-doctorants à partir d'octobre de cette année académique : l'un sur la conception d'algorithmes de fouille intégrant des connaissances linguistiques sous forme de contraintes, le second sur l'utilisation d'outils de fouille pour l'analyse et l'exploration de corpus en sciences humaines. Par ailleurs, depuis décembre dernier, je co-encadre avec Maroua Bouzid (GREYC) une thèse portant sur l'extraction d'information pour la capitalisation de connaissances, thèse financée par la société EADS dans le cadre d'une CIFRE.

Comme l'indique le titre de ce mémoire, l'objectif à terme est de tendre vers une *hybridation* et de fournir de nouvelles méthodes pour guider la découverte de connaissances à partir de textes en combinant les méthodes de fouille de données et de TAL. L'idée centrale est de concevoir un processus convergent qui mette en interaction les deux types de méthodes. Les activités du projet ANR Hybride[9], dont je suis le coordinateur pour le GREYC, et qui démarre ce 1er décembre, s'inscrivent directement dans cette problématique.

janvier 2008 à décembre 2009.

6. Projet ANR BINGO*2* intitulé "Knowledge discovery for and by inductive queries in post-genomic applications" impliquant les laboratoires CGMC et LIRIS (Lyon), LHC (Saint-Etienne) et GREYC (Caen) pour une durée de 4 ans (janvier 2008 à décembre 2010). `https://bingo2.greyc.fr/`

7. De septembre 2008 à août 2009, docteur de l'université de Montpellier (LIRMM), ensuite MC à l'université de Lyon 1.

8. De septembre 2009 à août 2010, docteur de l'université de Rennes (IRISA), ensuite MC à l'INSA de Rennes

9. Ce projet (programme blanc de l'ANR) intitulé « Hybridation de la fouille de données et du traitement automatique de la fouille de données » associe, outre le GREYC, les laboratoires INRIA-NGE Nancy, MODYCO Paris-Ouest et Orphanet INSERM Paris pour une durée de 4 ans (décembre 2011 – novembre 2015).

1.2 Positionnement

Notre démarche s'appuie sur la prise en compte de l'aspect sémantique de la langue sous forme de modélisation linguistique et de ressources linguistiques pour le développement d'applications. La méthodologie de l'accès à l'information est donc ici vue comme un accès au sens à partir d'une modélisation linguistique relative à l'application visée plutôt qu'un accès aux formes de surface comme l'utilisent par exemple actuellement les méthodes prédominantes des moteurs de recherche. En retour, la formalisation du modèle linguistique et son expérimentation sur corpus visent à améliorer notre connaissance des phénomènes traités et permettre un retour qualitatif, c'est-à-dire explicatif, sur le modèle.

L'intérêt pour les méthodes de fouille fondées sur les motifs et la volonté de les intégrer au TAL – notamment pour acquérir automatiquement des ressources linguistiques et alimenter à moindre coût les systèmes de TAL – procède de la même démarche. Une telle combinaison vise d'une part à repenser la manière dont les processus de TAL peuvent tirer bénéfice de méthodes de fouille spécifiquement adaptées au texte, et d'autre part à renforcer les processus de fouille, en fonction des spécificités de la donnée textuelle, par une prise en compte d'informations de nature linguistique visant notamment à sélectionner l'information la plus pertinente.

1.3 Plan du mémoire

Ce mémoire se décline en deux grandes parties. La première présente mes travaux menés plus spécifiquement avec des méthodes de TAL (chapitre 2 et 3) et la seconde mes travaux plus récents visant à intégrer les méthodes de fouille en TAL (chapitre 4).

Le chapitre 2 synthétise mes premiers travaux qui ont trait à la compréhension automatique. Ceux-ci visent à obtenir une représentation formelle de l'énoncé permettant ensuite de réaliser des inférences. Dans ce contexte, nous avons particulièrement étudié et implémenté le formalisme de représentation des graphes conceptuels dans le cadre de deux travaux dont l'originalité principale est de produire une représentation sémantique profonde du texte dans ce formalisme pour réaliser des inférences [Charnois 2000, Amghar *et al.* 2001].

Le chapitre 3 porte sur les méthodes d'analyse orientées vers l'accès à l'information. Nous décrivons trois travaux réalisés dans le cadre de contrats de recherche. Le premier concerne la recherche documentaire sur des documents géographiques [Bilhaut *et al.* 2003a] : le développement d'analyses locales et la mise en oeuvre du modèle linguistique de l'encadrement du discours ont permis à notre équipe de développer un moteur de recherche exploitant les annotations fournies par les analyses. L'extraction d'information dans les textes biomédicaux est l'objet de l'activité suivante décrite dans ce chapitre. Nous présentons une méthode d'extraction d'entités nommées simple et légère avec peu de ressources et sans analyse syntaxique [Charnois *et al.* 2006]. Enfin, nous présentons nos premières expérimentations dans le domaine de l'analyse de l'opinion, expérimentations visant à fournir des observables pour une étude des constituants détachés porteurs de jugements [Jackiewicz *et al.* 2009a]. Ces travaux montrent l'intérêt de cette approche, mais aussi leurs limites

et notamment l'importance des ressources qui impactent directement les performances des outils. C'est pourquoi, je me suis intéressé à la problématique de l'apprentissage des ressources linguistiques qui est au centre du chapitre suivant.

La deuxième partie du rapport (chapitre 4) décrit comment l'utilisation de méthodes de fouille de données fondée sur les motifs a permis d'apporter une contribution au problème de l'acquisition des ressources en TAL et, en retour, à réfléchir sur les méthodes de fouille en elles-mêmes. Nous avons ainsi été amenés à en concevoir de nouvelles plus adaptées à ce type de matériau. D'autre part, nous pensons que ces développements apportent aussi des contributions à des problématiques ne relevant pas nécessairement de la fouille de données textuelles, comme par exemple la possibilité de fouiller des bases de données séquentielles où les séquences sont elles-mêmes composées d'ensembles d'attributs. Pour apprendre des patrons linguistiques en vue de l'extraction de relations entre entités nommées (comme les interactions entre gènes) ainsi que leur caractérisation [Cellier *et al.* 2010a], nous avons conçu une méthode de fouille de données fondée sur les motifs séquentiels et sur une fouille récursive des motifs eux-mêmes [Cellier *et al.* 2010b]. Une originalité de notre approche est de ne pas utiliser d'analyse syntaxique tout en permettant de produire des résultats symboliques utilisables comme ressources linguistiques. Pour la reconnaissance d'entités nommées, nous proposons une méthode fondée sur un nouveau type de motifs intégrant une séquence et son contexte [Plantevit *et al.* 2009]. Enfin, nous proposons un cadre s'appuyant sur l'analyse de concepts logique pour explorer les motifs et identifier les patrons linguistiques [Cellier *et al.* 2011].

En fin de document, nous présentons nos perspectives de recherche. L'objectif est de fournir de nouvelles méthodes de fouille de données et de TAL génériques à l'extraction de connaissances dans les textes, l'idée étant de tirer profit d'un enrichissement mutuel entre les deux approches et leur utilisation combinée.

Chapitre 2

Compréhension automatique et graphes conceptuels

La compréhension automatique de textes (CA) est un des problèmes les plus ardus et des plus passionnants dans le domaine du TAL. Mes premiers travaux de recherche se situent dans la lignée des systèmes qui se sont intéressés à cette question. Je décris les grands traits les principes généraux de ces systèmes [Charnois et Enjalbert 2005] (ou [Sabah 2000] pour une synthèse historique) et présente ensuite mes travaux de cette époque autour des graphes conceptuels. Le bilan que j'en ai tiré a orienté pour une grande part la suite de mes recherches.

Comprendre automatiquement la langue a été l'un des objectifs ambitieux de *l'intelligence artificielle* (IA) autour des années 1970-90. Pour bien appréhender cette problématique, il est intéressant de la resituer dans le cadre du programme de l'IA qui vise à faire reproduire par des ordinateurs des activités humaines (raisonnement, perception, organisation et planification de l'action, et bien sûr, communication langagière). À cette période, comprendre automatiquement un texte, c'est être capable de produire une représentation informatique du *sens du texte*. Il est important de souligner que le sens s'entend globalement comme toute l'information véhiculée par le texte. À l'issue de ces analyses, et dans l'absolu, la représentation obtenue n'est pas dédiée à un type d'application particulier mais est conçue pour être utilisable par tout type d'application de TALN (traduction automatique, résumé automatique, dialogue homme-machine, *etc.*). On constate donc que le processus de représentation est primordial dans cette approche – l'application est en second plan. Ce processus nécessite :

- un formalisme exploitable par des procédures informatiques pour permettre des inférences et suffisamment expressif pour représenter le sens d'un énoncé, la logique et les graphes conceptuels étant les deux modèles majeurs de l'IA ;
- des mécanismes permettant de calculer la représentation formelle à partir de l'énoncé et de diverses connaissances linguistiques (lexicales et syntaxiques par exemple) associées au texte et extra-linguistiques (connaissances relatives au domaine du discours).

Pour ce faire, nombre de systèmes se sont appuyés sur les niveaux de la langue (du mot au texte) pour effectuer des analyses symboliques (morphologique, syntaxique, sémantique,

pragmatique), et génériques, correspondant à ces différents paliers et sur des théories linguistiques (modèles syntaxiques et sémantiques). C'est un processus de traduction aboutissant à la représentation formelle visée – du texte à l'analyse syntaxique, puis de la syntaxe à une représentation logique ou conceptuelle.

2.1 Graphes conceptuels comme formalisme de représentation de la langue

Le modèle des graphes conceptuels (en abrégé GC) est issu des réseaux sémantiques critiqués pour leur manque de rigueur et d'interprétation formelles (cf. [Woods 1975, Brachman 1979]). Il a été introduit par [Sowa 1984] et utilisé dans de nombreux domaines de l'IA. Il a notamment été considéré comme un formalisme bien adapté pour la représentation sémantique d'énoncés et la modélisation de phénomènes linguistiques (par exemple, dans les projets de compréhension automatique de textes tels que Kalipsos [Bérard-Dugourd *et al.* 1988] et Menelas [Consortium Menelas 1995], dans les travaux sur la modélisation de la causalité [Nazarenko 1993], ou encore dans l'étude de la métonymie [Amghar *et al.* 1995].

Ce formalisme présente en effet de multiples intérêts. Il fournit un bon compromis entre capacités inférentielles et pouvoir expressif tout en préservant des propriétés formelles de complétude et de cohérence des opérations. Un GC est un graphe, biparti [1] et étiqueté, muni d'un support [2], qui définit le langage et permet de représenter des connaissances, et sur lequel sont définies des opérations (notamment la spécialisation de sommets, la jointure et la projection entre GC) qui permettent de raisonner (les principales notions du domaine sont données en annexe). De plus, la (relative) proximité de ce modèle avec la langue est intéressante : typiquement, à un mot on peut associer un sommet concept, et une relation comme un cas sémantique [Fillmore 1968] va se représenter par un sommet relation.

Le problème de la complexité L'un des problèmes posés par l'implémentation du modèle des GC est celui de la complexité des opérations. La recherche d'une projection d'un GC sur un GC est un problème NP-Complet [Chein et Mugnier 1992]. Deux types de propositions sont possibles : i) envisager un algorithme général avec retours arrières pour le calcul des solutions. Voir par exemple [Leclère 1995] qui propose un algorithme fondé sur le développement d'un arbre de recherche des solutions avec un coût algorithmique de l'ordre de $O(m^n)$ m et n étant le nombre de sommets. Même si les graphes utilisés en langue sont « petits », il n'est pas rare d'avoir à manipuler des graphes de plusieurs dizaines de sommets ; ii) restreindre l'opération de projection à des classes de graphes pour lesquelles l'opération devient polynomiale : graphes sans cycles [Mugnier 1995], ou autres restrictions données par [Willems 1995] ou [Liquière et Brissac 1994]. De plus, la jointure maximale qui a été souvent utilisée pour simuler la compositionnalité en langue a été peu étudiée d'un point de vue algorithmique. La recherche d'une jointure maximale

1. On distingue les sommets concepts et les sommets relations

2. Un support comprend essentiellement trois ensembles partiellement ordonnés – types de concepts, types de relations entre concepts, marqueurs de concepts (qui peuvent être génériques pour représenter une classe de concept ou individuels pour une instance) –, des signatures de relations et une relation de conformité entre les marqueurs et les concepts.

au sens de la cardinalité (voir [Chein et Mugnier 1992]) est aussi un problème NP-Complet.

Nous avons proposé un algorithme de recherche exhaustive pour la projection et la jointure maximale et une heuristique gloutonne pour obtenir en temps polynomial les solutions [Charnois 2000]. L'idée est de tirer parti des caractéristiques des graphes de notre domaine qui représentent des énoncés linguistiques. Les relations conceptuelles (*e.g.* cas sémantique) sont binaires : on peut représenter les graphes sous forme de triplets ($concept - relation - concept$) pour optimiser la recherche d'appariements. Surtout, il y a peu de mots identiques (peu de sommets identiques), donc peu de solutions optimales : à chaque étape de la recherche on choisit une solution optimale locale [3]. La recherche est ainsi déterministe, et conduit généralement, et pour le type de graphes mentionné, à un résultat global optimal ; ce que nous avons pu vérifier dans la pratique.

2.2 Expériences

J'ai exploité et implémenté ce formalisme dans le cadre de deux travaux dont l'originalité principale est de produire une représentation sémantique profonde du texte sous forme de GC pour réaliser des inférences. L'idée forte est d'employer l'opération de jointure maximale plus puissante que la jointure dirigée plus souvent employée (par exemple par [Bérard-Dugourd *et al.* 1988]) pour produire, par combinaison des différents éléments de l'énoncé, la représentation sémantique (application du principe de compositionnalité). Dans le premier cas [4], il s'agissait de représenter des textes spécifiant des services de télécommunication. Nous avons montré comment passer d'un énoncé textuel donné sous forme de relations casuelles [5] à une représentation sous forme de graphes conceptuels enrichie par l'insertion des connaissances du domaine permettant d'obtenir une représentation sémantique profonde de l'énoncé [Charnois 1999].

Je décris plus précisément la deuxième expérience [6] fondée sur les mêmes principes [Amghar *et al.* 2002]. L'objectif de ce travail consiste à produire une représentation de la signification de textes (courts) relatant des mouvements spatiaux et temporels. La représentation vise à pouvoir spécifier dans le temps des positions relatives d'objets entre eux et par rapport à des lieux. Cette procédure nécessite la mise en place de raisonnements réalisés à partir de l'information aspectuo-temporelle, et d'un niveau de description lexicale et grammaticale suffisamment fin. Pour ces raisons, nous nous appuyons sur les travaux de J.P. Desclés sur la modélisation des significations des verbes de mouvement – sous forme de *Schèmes Semantico-Cognitifs* – [Desclés 1990] d'une part, et sur la modélisation du temps linguistique et de ses valeurs aspectuelles [Desclés 1989] d'autre part. Une première approche permettant de construire la Représentation Sémantico-Cognitive d'un texte (RSC) en tenant compte des significations lexicales et de la temporalité narrative a été proposée

3. Pour un triplet donné du premier graphe, on choisit le triplet à apparier dans l'autre graphe qui conserve le plus grand nombre d'appariements possibles – on élimine les triplets qui ne satisfont pas les deux triplets appariés –, puis on itère sur les autres triplets du premier graphe restant à apparier.

4. Travail réalisé en thèse.

5. Structure issue de l'analyse syntaxique et dans laquelle les prédicats verbaux ou nominaux sont reliés à leurs arguments par des cas sémantiques.

6. Travail post-doctoral.

dans [Battistelli et Vazov 1997]. Nous avons étendu ce travail pour développer une chaîne complète menant d'un texte à sa représentation sous forme de graphe conceptuel, formalisme choisi pour ses capacités expressives et inférentielles. Le système a été expérimenté sur un corpus de constats d'accidents dans lequel le traitement des informations aspectuo-temporelles revêt un aspect crucial. Un extrait de ce corpus est donné par la figure 2.1 (nommé *texte A* par la suite).

> *(1) Etant arrêté momentanément sur la file droite du Bd des Italiens, (2) j'avais mis mon clignotant, (3) j'étais à l'arrêt. (4) Le véhicule B arrivant sur ma gauche (5) m'a serré de trop près et (6) m'a abîmé tout le côté avant gauche.*

FIGURE 2.1 – Extrait du corpus des constats d'accident

Dans le modèle de Desclés, un texte se représente par deux structures : la première modélise la représentation sémantique des propositions du texte, et la seconde le diagramme temporel exprimant les contraintes temporelles entre les différentes situations décrites dans le texte. La prise en compte de ces deux structures et des liens qu'elles entretiennent nous a amenés à modifier le modèle des graphes conceptuels simples, et à envisager les modes d'interaction entre temps, aspect (grammatical) et signification des lexèmes verbaux.

2.2.1 Positionnement

La théorie aspectuo-temporelle proposée par J.-P. Desclés se distingue à la fois des travaux de [Kamp 1979] et de [Reichenbach 1947]. Dans le premier cas, du fait qu'elle distingue trois grands types de procès (*état, processus* et *événement*) au lieu de deux (*état* et *événement* chez Kamp) mais aussi du fait qu'elle propose une stratégie explicite de calcul de ces valeurs primitives prenant en compte différents types de marqueurs temporels [Desclés *et al.* 1991]. Dans le deuxième cas, du fait qu'elle propose d'exprimer les entités temporelles primitives à l'aide d'intervalles (topologiques) et non de points.

Les modèles linguistiques sur la sémantique aspectuelle ayant donné lieu des implémentations effectives sont peu nombreux (*cf.* [Gosselin 2005, p. 174]). Outre le modèle de Desclés, celui de [Gosselin 1996] prend aussi en compte l'aspect qu'il représente au moyen de structures d'intervalles. La principale différence entre les deux modèles est que celui de Desclés ne propose pas de manière explicite une modélisation de l'aspect lexical et de l'aspect grammatical, ni surtout de l'interaction entre ces deux modes d'expression de l'aspect. Le modèle de Gosselin intègre lui explicitement cette différenciation et cherche à en modéliser l'interaction.

Notre travail cherche à rendre compte, dans le cadre d'un modèle formel (les GC), de cette interaction aspect lexical/aspect grammatical en utilisant les concepts linguistiques aspectuels du modèle de Desclés. Sur le plan de la modélisation linguistique, cette interaction est décrite en détail dans la thèse de [Battistelli 2000].

Les premiers travaux en TAL qui traitent du problème de la représentation de l'information temporelle au sein des GC sont ceux de [Sowa 1984] qui donne les bases d'un traitement du temps en TALN, utilisant les graphes conceptuels emboîtés. Cette proposition est équivalente à celle que l'on pourrait faire en utilisant des logiques modales. Ici, les

opérateurs modaux sont des étiquettes temporelles portant sur des intervalles temporels organisés en fonction des relations définies par [Allen 1983]. Sans considérer les problèmes issus de l'usage des logiques modales pour la représentation du temps linguistique [Bestougeff et Ligozat 1989], le point faible le plus important de cette approche concerne deux aspects liés à l'expression des connaissances temporelles dans la langue. Tout d'abord, les entités temporelles sont des points, ce qui ne convient pas aux données temporelles linguistiques. Et deuxièmement, le nombre de temps grammaticaux considérés semble insuffisant (seulement trois temps : le passé, le présent et le futur). De plus, le problème des valeurs aspectuelles n'est pas mentionné. Les travaux de [Moulin et Coté 1992] proposent une première extension du modèle des GC pour traiter du temps linguistique. L'expression des valeurs sémantiques des temps des verbes y est fondée sur les travaux de [Reichenbach 1947] et les relations entre intervalles qui sont représentées sont celles d'Allen. [Moulin 1993] note que sa première proposition est adaptée à la représentation d'une proposition, mais elle reste insuffisante pour exprimer la structure temporelle d'un texte. Il étend sa proposition en distinguant trois niveaux de représentation. Son nouveau modèle est partiellement basé sur les travaux de [Desclés 1989] en particulier en ce qui concerne la taxonomie des procès (*i.e. états, événements et processus*). Ceci lui permet de spécifier la structure temporelle d'un discours constituée de *situations temporelles*, d'*intervalles temporels*, et de *relations temporelles*. Ces approches sont complémentaires de notre travail mais s'en différencient par l'utilisation de graphes conceptuels emboîtés et se situent dans le cadre de la génération de textes.

2.2.2 Fondements linguistiques

Représentation Sémantico-Cognitive (RSC) d'une proposition

La sémantique des verbes est exprimée à l'aide de structures abstraites, composées de primitives sémantico-cognitives ; ces structures sont appelées *Schèmes Sémantico-Cognitifs* (SSC). La signification du verbe « arriver » par exemple est représentée par une λ-expression à l'aide du SSC suivant [7] :

[SSC du verbe "arriver"] $\lambda x.\lambda loc.\ CONTR\ (MOUVTor\ (\epsilon_0(ex\ loc)\ x)\ \ (\epsilon_0(fr\ loc)\ x))\ x$

En contexte, les variables d'un SSC sont instanciées par les arguments d'une relation prédicative donnée (composée d'un prédicat verbal et de ses arguments) ; la structure obtenue est appelée la *Représentation Sémantico-Cognitive* (RSC). Dans un texte, à chaque proposition identifiée correspond alors une RSC. Soient les deux propositions :

(1) *la voiture arrivait dans le virage*
(2) *j'avais mis mon clignotant*

7. Cette structure exprime qu'une entité x contrôle (CONTR) un mouvement orienté (MOUVTor) entre deux situations statiques (Sit1 et Sit2). Dans la première (Sit1), x est repérée (ϵ_0) par rapport à l'extérieur (ex) d'un lieu loc ; dans la seconde (Sit2), x est repérée (ϵ_0) par rapport à la frontière (fr) du même lieu loc.

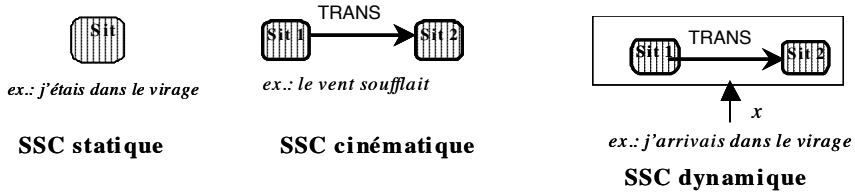


FIGURE 2.2 – Typologie des SSC

La RSC de (1) est construite à partir de la relation prédicative (RP) qui lui est sous-jacente à savoir *<arriver_dans, le_virage, la_voiture>* et du SSC de « arriver » en instanciant les variables de celui-ci par les arguments de la RP, à savoir */le_virage/* et */la_voiture/*. Le résultat est donné par (1'). Le même processus permet de construire la RSC (2') [8] associée à (2).

$(1')CONTR(MOUVTor\ (\epsilon_0(ex\ le\ _virage)la_voiture)\ (\epsilon_0(in\ le_virage)la_voiture))\ la_voiture$
$(2')CONTR\ (CHANGTor\ (\epsilon_{act}\ (ex\ en_fonctionnement)\ le_clignotant)$
$(\epsilon_{act}\ (in\ en_fonctionnement)\ le_clignotant))\ je$

Dans (1') et (2'), les entités (repérées ou servant de repère dans une relation de repérage) sont typées : *la_voiture*, *le_clignotant* et *je* ont le type sémantico-cognitif INDIVIDU et les entités *le_virage* et *en_fonctionnement* ont le type sémantico-cognitif LIEU (lieu spatial pour le premier, lieu d'activité pour le second). D'une manière générale (figure 2.2), un SSC peut être de type statique (dans le cas d'une seule situation de repérage (Sit)), cinématique (dans le cas où il y a transition (TRANS) entre deux situations statiques(Sit1 et Sit2)) ou dynamique (dans le cas où il s'agit d'une situation cinématique elle-même sous le contrôle d'une entité, comme dans les exemples (1) et (2)). La figure 2.3.a donne les 6 RSC [9] associées aux 6 propositions du texte(A).

Afin d'exhiber les relations temporelles entre les différentes RSC associées aux propositions d'un texte, nous construisons le diagramme temporel du texte. Par ailleurs nous faisons appel aux valeurs aspectuelles des propositions pour prendre en compte l'interaction entre aspect lexical et aspect grammatical.

Valeurs aspectuelles et diagramme temporel d'un texte

L'organisation conceptuelle du domaine aspectuo-temporel repose sur la trichotomie état, processus et événement ; cette trichotomie trouve une interprétation dans le modèle

8. (2') exprime que l'entité *je* contrôle (CONTR) un changement orienté (CHANGTor) entre deux situations statiques (Sit1 et Sit2) concernant l'entité *le_clignotant*, repérée successivement (ϵ_{act}) par rapport au lieu d'activité *en_fonctionnement* visualisé dans son extérieur (*ex*) en Sit1 et dans son intérieur (*in*) en Sit2.
Une expression comme *mettre son clignotant* n'exprime pas un déplacement spatial de l'entité *clignotant*, mais l'entrée en activité du système que décrit un clignotant. Par analogie avec les lieux spatiaux on peut alors décrire les états d'activité d'un système, conçus comme des lieux, en utilisant les primitives de spécification topologique.

9. dans lesquelles e_0 et e_{act} représentent respectivement ϵ_0 et ϵ_{act}.

RSC_1	&2 (e_{act} A_L'ARRET A) (e_0 FR(la_file_dr_du_bd_des_It) A)
RSC_2	CONTR (CHANGTor (e_{act} EX(EN_FONCTIONNEMENT)clig(A)) (e_{act} IN(EN_FONCTIONNEMENT)clig(A)) A
RSC_3	(e_{act} A_L'ARRET A)
RSC_4	CONTR (MOUVTor (e_0 EX(GAUCHE(A) B) (e_0 FR(GAUCHE(A) B)) B
RSC_5	CONTR (MOUVTor (e_o (LOC A) B) (e_o FR(LOC A) B)) B
RSC_6	&2 (E1> E2) (CONTR (CHANGTor (e_0 IN(E1) tout_le_côté_avt_gauche (A)) (e_0 IN(E2) tout_le_côté_avt_gauche (A)) B)

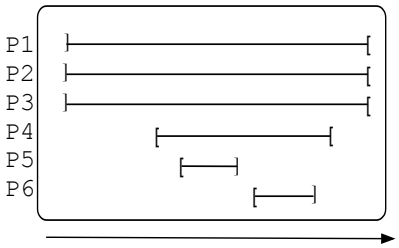


FIGURE 2.3 – a) RSC du texte (A) — b) Diagramme temporel de (A)

	PROCESSUS	EVENEMENT	ETAT (résultatif)	ETAT (d'activité)
CINEM /DYNA	Sit 1 → Sit 2 ex.: je roulais vers Marseille	Sit 1 → Sit 2 ex.: j'ai roulé jusqu'à Marseille	Sit 1 → Sit 2 ex.: j'avais roulé jusqu'à Marseille	Sit 1 → Sit 2 ex.: j'étais en route pour Marseille
STAT	Ø	EVENEMENT Sit ex.: j'ai habité cette maison	ETAT (descriptif) Sit ex.: j'habitais cette maison	ETAT (conséquent) Sit ex.: j'avais habité cette maison

FIGURE 2.4 – RSC et visées aspectuelles

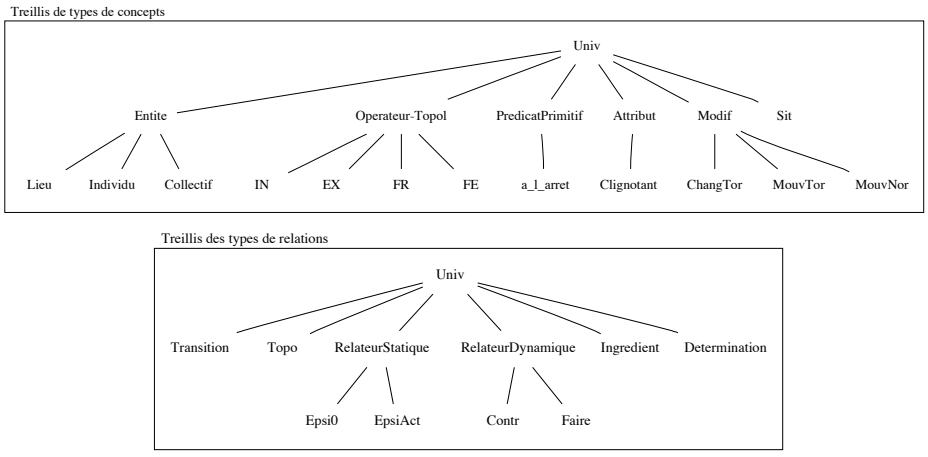


FIGURE 2.5 – Extrait des treillis de concepts et de relations (le plus petit élément - le type absurde - est omis pour alléger les figures)

topologique [Desclés 1989]. La valeur aspectuelle d'une proposition est calculée à partir d'indices (grammaticaux et lexicaux) présents dans le contexte, selon une stratégie d'*exploration contextuelle* présentée dans [Desclés *et al.* 1991]. Chaque valeur aspectuelle correspond alors à une zone de validation représentée par un intervalle topologique d'un certain type, situé sur un axe représentant le référentiel temporel de l'énonciateur. Une valeur d'état est représentée par un intervalle ouvert ; une valeur d'événement par un intervalle fermé ; une valeur de processus par un intervalle semi-ouvert à droite [10]. Les relations temporelles entre intervalles sont calculées à partir de la prise en compte de la sémantique des connecteurs temporels et d'un ensemble de règles appliquées par défaut [11]. La figure 2.4 présente l'interaction entre les différentes catégories de RSC en fonction des différents types de visée aspectuelle (de type état, événement ou processus) qui peuvent être envisagés [12]. Dans les rectangles grisés sont indiquées les situations encodées lexicalement ; les rectangles blancs désignent les situations qui sont focalisées selon le type de la valeur aspectuelle. Un rectangle complètement vide indique que la situation n'est pas encodée lexicalement i.e. qu'aucune description de cette situation n'est fournie par la signification du prédicat verbal et de ses arguments. La figure 2.3.b fournit le diagramme temporel du texte (A).

2.2.3 Information temporelle et graphes conceptuels

Nous représentons un texte par deux types de Graphes Conceptuels Simples (GCS). Le premier modélise la représentation sémantico-cognitive du texte et le second le diagramme temporel exprimant les contraintes temporelles entre les différentes situations décrites dans le texte. Afin de représenter et traiter l'information temporelle nous avons légèrement étendu le modèle des GCS.

Les marqueurs temporels

Nous représentons l'information temporelle par un marqueur spécifique. Un sommet concept est ainsi étiqueté par un triplet *<type,marqueur1,marqueur2>* où *type* et *marqueur1* correspondent respectivement au type et au marqueur du concept (*cf.* note 2 page 8), et *marqueur2* est un marqueur aspectuo-temporel. Il faut noter que ce deuxième marqueur est un marqueur individuel de l'ensemble des marqueurs de DT (le diagramme temporel). Dans notre application, ce marqueur joue le rôle de lien co-référent entre l'intervalle temporel associé à la proposition du texte dans DT et les valeurs aspectuo-temporelles de la RSC correspondante. Si pour un sommet concept donné cette information n'est pas instanciée, le marqueur *absurde* est utilisé par convention afin d'empêcher la fusion de deux sommets ayant même type et même $marqueur1$, mais ayant des $marqueur2$ différents : nous faisons l'hypothèse que le même objet considéré à des moments différents doit correspondre à différentes représentations. Nous avons redéfini l'opération de jointure maximale pour tenir compte de ce marqueur.

10. Par exemple, les propositions (1) et (2) auront respectivement pour valeur aspectuelle *processus* et *état résultant*.

11. règles appliquées en l'absence de marqueurs temporels explicites tels que les connecteurs temporels ; [Battistelli 2000] présente en détail l'ensemble de ces règles ainsi que le processus de construction des RSC.

12. C'est à ce niveau que nous abordons le problème de l'interaction entre aspect lexical et aspect grammatical.

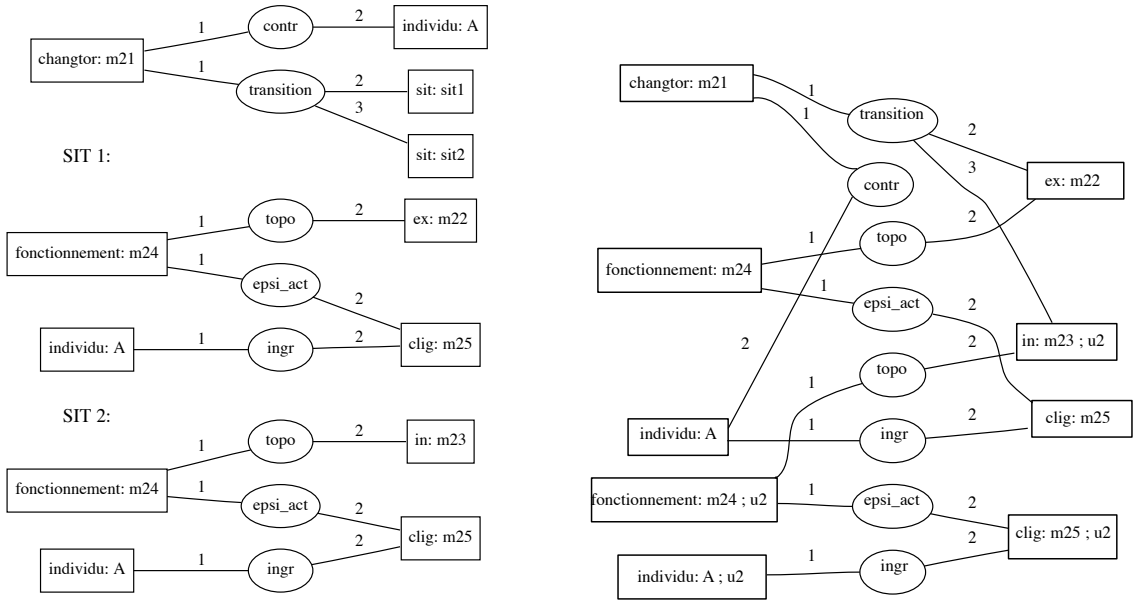


FIGURE 2.6 – Représentation de la proposition textuelle *J'avais mis mon clignotant* : à gauche la RSC associée, à droite le GC obtenu après propagation des marqueurs temporels et jointures

La représentation des intervalles

Nous utilisons des intervalles typés par des bornes ouvertes ou fermées afin de représenter l'information aspectuelle (état, événement, processus). Les relations entre intervalles que nous avons implémentées sont celles proposées par [Allen 1983].

```
ETAT : ]----[
type ETAT(x) is [point]->(])->[intervalle: *x] ->([)->[point]

EVENEMENT : [----]
type EVENEMENT(x) is [point]->([)->[intervalle: *x] ->(])->[point]

PROCESSUS : [----[
type PROCESSUS(x) is   [point]->([)->[intervalle: *x]->([)->[point]
```

Nous exprimons les relations d'Allen sous forme de définitions de relations, par exemple [13] :

```
relation BEFORE(x,y) is   [point:*p1]->(|)->[procès: *x]->(|)->
   [point:*p2]->(<)->[point:*p3]->(|)->[procès: *y]->(|)->[point:*p4]

relation MEETS(x,y) is   [point:*p1]->(|)->[procès: *x]->(|)->
   [point*p2]->(|)->[procès: *y]->(|)->[point:*p3]
```

Ces primitives permettent de représenter le diagramme temporel du texte.

2.2.4 Système développé

La représentation finale du texte est réalisée par trois modules qui, à partir de la RSC de chaque proposition textuelle et du DT, génèrent deux GCS. Ils représentent respec-

13. le type de concept *procès* est un super type des types *état,événement et processus* et le type de relation '|' un super type de '[' et ']'

tivement les informations sémantiques et temporelles. L'ontologie utilisée (figure 2.5) est celle proposée par J-P. Desclés. Un analyseur implémenté en Prolog sous forme de grammaire DCG traduit une RSC en un GCS. Les GCS ainsi obtenus ne contiennent que des informations d'ordre lexical. L'information aspectuo-temporelle est propagée par le second module.

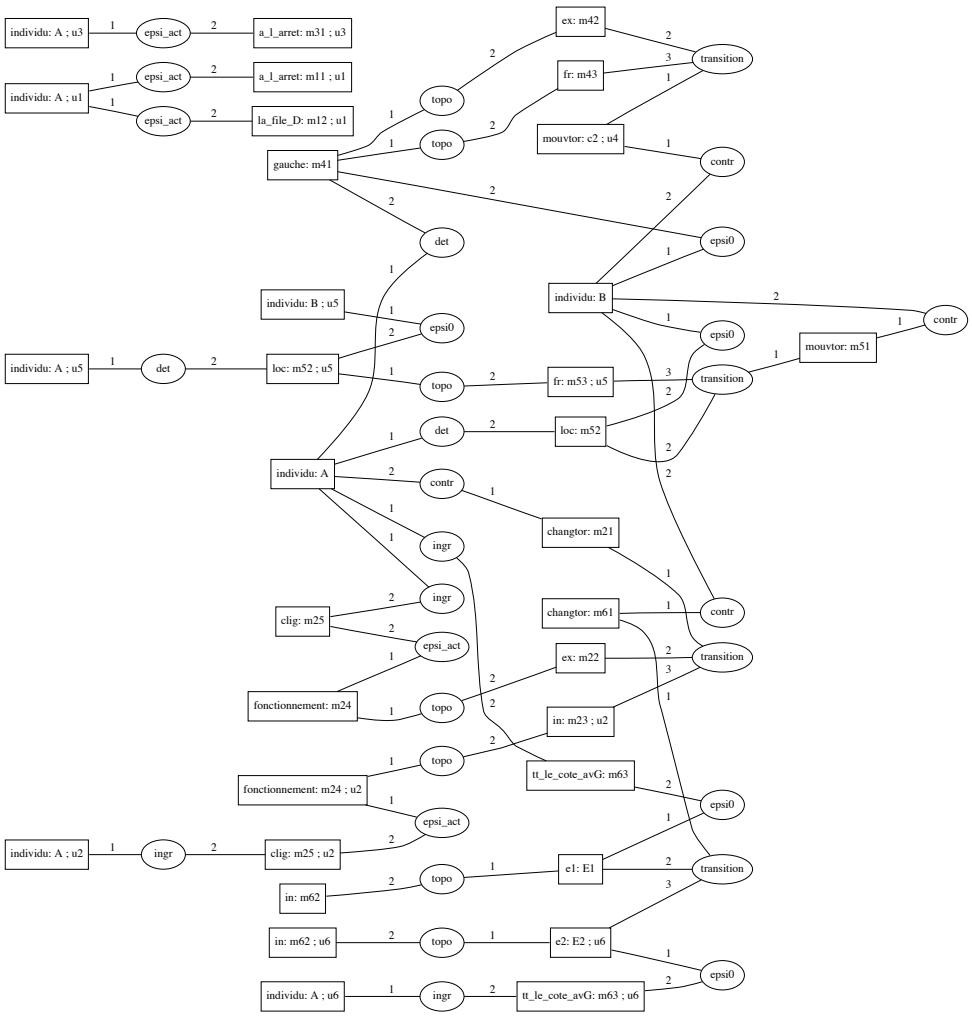


FIGURE 2.7 – Représentation sémantique du texte (A)

La propagation de marqueurs temporels

Ce module propage le marqueur de l'intervalle temporel correspondant à la proposition considérée dans la partie adéquate du GCS en fonction de la valeur aspectuelle de l'action décrite. Par exemple, si la valeur aspectuelle d'une proposition est EVENEMENT, l'ensemble des concepts du composant nommé *sit*2 est marqué temporellement par le marqueur du c-sommet correspondant au procès du DT. Cette implémentation est conforme à l'ensemble des règles données par la figure 2.4.

Une méthode incrémentale

En suivant un principe de compositionnalité, la représentation sémantique intégrant les informations aspectuo-temporelles est obtenue en combinant les RSC par l'opération de jointure maximale [14]. Dans le cas de RSC de type cinématique ou dynamique, l'étape succédant à la propagation des marqueurs temporels consiste à remplacer, par jointure maximale, les sommets $sit1$ et $sit2$ par les GCS correspondants. L'opération est répétée pour tous les GCS associés aux RSC. Enfin, la combinaison – par jointures maximales successives – de tous ces GCS produit la représentation du texte. Les figures 2.6 et 2.7 donnent respectivement le graphe associé à la proposition (2) du texte (A) et le graphe associé à l'intégralité du texte (A). Dans cette représentation, le troisième champ des sommets concepts (lorsqu'il existe [15]) est le marqueur temporel. C'est aussi le marqueur individuel d'un concept de type intervalle du graphe DT (*cf. supra*) ; il permet ainsi de retrouver l'intervalle temporel associé et ses relations aux autres intervalles dans DT.

2.3 Bilan

Les limites et les acquis des travaux en compréhension automatique sont bien connus ([Kayser 2000], [Charnois et Enjalbert 2005], [Nazarenko 2006], [Sabah 2006]). Je souligne simplement quelques aspects qui m'ont paru importants pour la suite de mes travaux.

En résumé, ce type d'approche visait une compréhension globale du texte reposant sur des théories linguistiques et sur des analyses complètes, et une compréhension générique (valable dans tout domaine). Les résultats de ces recherches portent sur des aspects théoriques et méthodologiques importants (modélisation informatique de théories linguistiques, développement de formalismes de représentation [16]). Cependant les usages de la représentation en termes de tâches applicatives n'ont pas répondu aux attentes. D'une part, les systèmes se sont heurtés à des phénomènes inhérents à la langue (entre autres : ambiguïté syntaxique, sémantique, co-référence, glissement de sens...) difficiles à traiter, et non encore maîtrisés, mais dont la prise en compte est nécessaire dans le cadre d'analyse complète et profonde. D'autre part, l'objectif de généricité est aussi dans l'absolu trop ambitieux en raison de l'impossibilité de disposer des connaissances de tous les domaines possibles.

Mes travaux sur les graphes conceptuels dans le cadre de la compréhension ont montré l'intérêt d'une modélisation linguistique pour permettre des analyses fines et produire une représentation sémantique des énoncés, mais sur des corpus de spécialité afin de limiter les connaissances nécessaires. En effet, les méthodes et outils développés à cette période permettent de concevoir une représentation de domaines particuliers et un traitement de certaines régularités des formes (structures syntaxico-sémantiques, organisations discursives) et d'envisager des systèmes opérationnels. Cette démarche s'est concrétisée dans les tâches d'extraction d'information pour lesquelles il s'agit d'identifier des éléments de sens, des informations ciblées [Hobbs et Riloff 2010]. Parallèlement, à la fin des années 90,

14. L'algorithme de jointure implémenté est décrit dans [Charnois 2000].

15. sinon le concept n'est pas marqué temporellement.

16. On peut noter que le modèle RDF (*Resource Description Framework*) utilisé pour annoter des données du Web sémantique, à large échelle donc, est directement issu des travaux de l'IA puisqu'un graphe RDF est simplement un type de réseau sémantique.

le développement d'internet a permis la disponibilité de corpus nombreux et volumineux et a fait naître de nouveaux besoins en termes de recherche documentaire et d'accès à l'information. J'ai donc souhaité travaillé sur ces problématiques, tout en conservant une démarche linguistique, pour orienter l'analyse en prenant en compte l'objectif et le domaine applicatifs.

Chapitre 3

Analyse linguistique pour l'accès à l'information

Dans le prolongement de mes travaux sur la compréhension, je me suis intéressé à la problématique de l'extraction d'information, et plus généralement à celle de l'accès à l'information pour des usages applicatifs liés aux besoins croissants en recherche documentaire. Par rapport aux travaux sur la compréhension automatique, ce n'est plus une représentation sémantique profonde du texte qui est visée, mais le repérage d'informations, ou de segments de textes, ciblés dans un domaine applicatif déterminé et restreint. De plus, le corpus de textes à traiter est aussi spécialisé. Comme souligné à la fin du chapitre précédent, ces restrictions permettent de viser des systèmes opérationnels pour la tâche applicative.

Nous décrivons dans ce chapitre les travaux que nous avons menés dans le cadre de projets interdisciplinaires. Nos travaux se caractérisent par une méthodologie linguistique et une perspective applicative. Sur le plan méthodologique, il s'agit de mettre au point des analyses symboliques développées à partir d'une observation ou d'une description linguistique, souvent en collaboration avec des linguistes, et destinées à annoter les informations extraites des textes. Soulignons que les collaborations interdisciplinaires (géographes, biologistes, et surtout linguistes) sont fructueuses pour la mise au point du modèle sur l'objet d'étude tant en amont de l'expérimentation qu'en cours d'expérimentation pour valider ou affiner le modèle. D'autre part, les informations extraites sont exploitées pour des usages applicatifs : recherche documentaire (projet GeoSem), extraction de connaissances (projet Bingo), exploration de corpus et analyse d'opinion (projet Ontopitex).

3.1 Extraction d'information et analyse discursive pour la recherche intra-documentaire

Cette section présente notre travail en sein du projet GéoSem, un projet de recherche d'information documentaire s'appuyant sur des analyses linguistiques pour indexer les corpus et présenter à l'utilisateur des passages correspondant à sa requête.

Les techniques qui sont mises en œuvre dans le domaine de la recherche d'information, notamment les moteurs de recherche sur le web, sont largement dominées par les méthodes statistiques traitant la forme du mot pour indexer les documents avec parfois

des pré-traitements morphologiques (lemmatisation, stemmatisation), sans prise en compte de la dimension sémantique de la langue. Cette dimension semble pourtant nécessaire pour interpréter la requête de l'utilisateur rédigée dans sa langue, améliorer l'indexation, et au final obtenir une meilleure mise en correspondance entre la requête et l'indexation. D'autre part, les systèmes de recherche d'information retournent une liste de documents que l'utilisateur doit explorer pour trouver l'information qu'il recherche. Pour l'aider à repérer des zones porteurs de cette information, sont apparus des travaux qui proposent un découpage du texte en segments thématiquement homogènes [Enjalbert et Bilhaut 2005]. Dans la lignée des travaux de [Hearst 1994], une première série d'approches se fondent sur la notion de cohésion lexicale [Halliday et Hasan 1976] en exploitant la distribution des mots ou leur récurrence lexicale comme indicateur d'homogénéité thématique. Plus récemment, des travaux ont exploité des marqueurs linguistiques de la structure discursive, comme par exemple les expressions *au sujet de X, à propos de X, en ce qui concerne X...*, qui « instancient des cadres thématiques », c'est à dire introduisent des zones textuelles relatives au thème *X*. Ce type d'indices est aussi exploité en complément d'une segmentation s'appuyant sur la récurrence lexicale [Ferret *et al.* 2001, Couto *et al.* 2004]. Ces approches visent donc à délimiter des segments thématiques au sein des textes mais ne fournissent pas d'indexation du document. Enfin, certains travaux se sont attachés à identifier dans un texte les expressions temporelles et les annoter [Muller et Tannier 2004]. Dans le domaine de l'information géographique, plusieurs projets comme GIPSY [Woodruff et Plaunt 1994] et SPIRIT [Vaid *et al.* 2005] visent à repérer les expressions spatiales géoréférencées des textes et à les indexer.

3.1.1 Présentation du projet GéoSem

Le projet GéoSem a pour originalité de proposer des procédures avancées de recherche d'information fondées sur des analyses sémantiques à un niveau local (intra-phrastique) et textuel afin d'indexer les documents par les représentations produites par ces analyses. L'objectif applicatif consiste à permettre à un utilisateur de trouver rapidement une information dans un document ou un ensemble de documents à partir de requêtes pouvant porter sur les trois dimensions *espace – temps – phénomène*. Ce projet s'est donné comme objet d'expérimentation des corpus du domaine de la géographie humaine [1] et l'idée est de tirer parti de l'une des caractéristiques de l'information géographique qui est de relier le phénomène observé, et décrit, à une localisation spatiale et, souvent, à une période temporelle. L'autre caractéristique de ce type de corpus est de comporter des textes longs. Comme évoqué en introduction, il faut alors pouvoir délimiter un ou plusieurs passages pertinents au regard de la requête de l'utilisateur pour lui apporter une aide dans sa recherche. Voici un exemple de requêtes visées par le projet :

- Retard scolaire dans l'Ouest de la France dans les années 1950
- Politique de sécurité maritime dans la Manche
- Évolution du vote FN dans le Grand Ouest entre 1981 et 2002

En réponse à une requête de ce type, le système doit retourner à l'utilisateur un passage d'un document, ou un ensemble de passages, satisfaisant le ou les critères spécifiés (ce sont essentiellement les dimensions temporelles et spatiales qui ont été investies dans le

1. Ce domaine inclut la géographie urbaine, sociale, économique, politique, *etc.*

projet). Le problème se situe donc dans le champ de la recherche d'information mais à un niveau intra-documentaire. Dans GéoSem le moteur de recherche s'appuie donc sur une indexation linguistique des documents à deux niveaux :

1. une indexation sémantique. Cette première phase utilise des ressources linguistiques pour identifier et procéder à une analyse sémantique des expressions référant à la temporalité ou à l'espace qui sont des dimensions majeures dans le document géographique. La sémantique produite par ces analyses est associée, sous forme de représentation symbolique, aux expressions extraites ;
2. une indexation des passages. Cette étape met à contribution une analyse discursive s'appuyant sur la notion de cadre de discours introduite par Charolles [Charolles 1997] afin d'établir, pour des segments textuels à déterminer, une indexation incluant potentiellement les trois dimensions (espace, temps, phénomène).

En phase de recherche documentaire, le moteur de recherche doit établir une forme de correspondance sémantique entre les descripteurs et la requête.

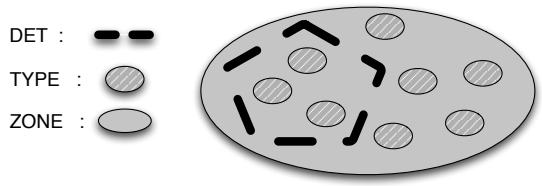
Je présente dans la suite mes principales contributions qui portent sur l'analyse spatiale en collaboration avec Patrice Enjalbert, Yann Mathet puis Séverine Beaudet [Bilhaut *et al.* 2003a;b] et sur la structuration en cadres de discours [Bilhaut *et al.* 2003c] avec l'équipe toulousaine et Frédérik Bilhaut, Patrice Enjalbert et Yann Mathet.

3.1.2 Analyse spatiale et temporelle

Suivant notre démarche de modélisation linguistique *a priori*, nous avons réalisé une étude sur corpus et une description fine des expressions spatiales qui a permis de dégager une structure canonique de la sémantique de ces expressions en lien avec son expression syntaxique. Schématiquement, à partir de toponymes dénotant des zones géoréférencées, la langue procède par application d'un certain nombre d'opérateurs : opérateurs spatiaux (ou géométriques) tels que *le nord/sud... de, le triangle X Y Z, de X à Y, etc.* ; et opérateurs de sélection d'entités au sein d'une zone donnée, selon divers types de critères : sociologiques, administratifs, physiques... (tableau 3.1). L'analyseur que nous avons développé [2] repère cette organisation et la traduit en structures symboliques (structures de traits récursives) exploitables pour une indexation de segments de textes (figure 3.1). L'analyseur temporel est conçu selon les mêmes principes. Les points d'arrêt sont ici les dates ; des opérateurs définissent des intervalles (*de X à Y, entre X et Y, les années X)* soumis à leur tour à une nouvelle classe d'opérateurs (*le début de X, aux alentours de X...*). Les deux types d'analyse font appel à des mécanismes compositionnels forts pour produire la représentation de l'expression. Remarquons que le principe de compositionnalité est ici tout à fait adapté

2. L'implémentation est réalisée en Prolog sous forme de grammaires DCG en utilisant les techniques GULP de Covington pour les structures de traits, sur la plateforme *LinguaStream* [Bilhaut et Widlöcher 2006]. L'analyseur prend en entrée le corpus préalablement étiqueté par un analyseur morphologique (*TreeTagger* [Schmid 1994]). Par exemple pour l'espace, la grammaire comporte environ 160 règles syntagmatiques et un petit lexique créé manuellement contient environ 200 entrées, incluant des termes grammaticaux (déterminants, prépositions...) ainsi que des termes administratifs (départements, villes, régions, *etc.*) et socio-économiques. Une base lexicale contenant plus de 100 000 lieux nommés français (gazetteer) donnant le type administratif et les coordonnées géographiques du lieu est également utilisée comme ressource [Bilhaut *et al.* 2003a;b].

aux formes particulières des expressions spatiales et temporelles.



DET	TYPE		ZONE	
	administratif	*qualification*	*position*	*entité géographique*
				Paris
La plupart	départements		du nord de	la France
quelques	villes	maritimes		de la Normandie
les	départements	ruraux	situés au sud de	la Loire

TABLE 3.1 – Structure des expressions spatiales

Tous les départements du nord de la France / Quelques villes maritimes de la Normandie

(a)
$$\begin{bmatrix} det : [\, type : exhaustif \,] \\ type : [\, ty_zone : departement \,] \\ zone : \begin{bmatrix} egn : \begin{bmatrix} ty_zone : pays \\ nom : France \end{bmatrix} \\ loc : interne \\ position : nord \end{bmatrix} \end{bmatrix}$$

(b)
$$\begin{bmatrix} det : [\, type : relatif \,] \\ type : \begin{bmatrix} ty_zone : ville \\ geo : maritime \end{bmatrix} \\ zone : \begin{bmatrix} egn : \begin{bmatrix} ty_zone : region \\ nom : Normandie \end{bmatrix} \\ loc : interne \end{bmatrix} \end{bmatrix}$$

FIGURE 3.1 – Expressions spatiales et leur représentation sémantique. L'exemple (a) montre une quantification exhaustive stipulant que toutes les entités du type administratif *département* localisées dans la zone sont sélectionnées. La zone correspond à la partie *nord* de la *France*. En (b), la détermination *quelques* est relative (seule une partie des éléments donnés par le type est à considérer). Le type stipule que dans la zone, on ne doit considérer que les villes qui sont maritimes.

Les représentations sémantiques ainsi obtenues sont exploitées pour l'indexation des textes. Elles restent relativement « abstraites » et ne peuvent être directement utilisées par un moteur de recherche qui doit comparer des représentations entre elles (requête / index). Une seconde phase va transformer les structures symboliques en des structures « référentielles » [3]. Les représentations issues de l'analyse spatiale sont traduites en représentation géométrique sous la forme d'une entité située dans un espace géolocalisé (boîte rectangulaire représentée par la latitude et longitude de ses coins nord-ouest et sud-est) et typé [Bilhaut *et al.* 2007]. Pour ce faire une base d'entités nommées géographiques spécifiant leur géolocalisation, ainsi que des règles géométriques sont utilisées. Par exemple, on obtient pour l'expression *Les départements maritimes du nord de la France* :

3. Ce travail a été réalisé par Frédiérik Bilhaut, Franck Dumoncel et Nicolas Hernandez qui ont aussi développé le moteur de recherche évoqué ci-dessous.

```
spatial: (entité: (type: département, identifiant: libre, qualificatif: maritime)
          boîte : (latitude NO, longitude NO, latitude SE, longitude SE))
```

Le calcul de la boîte pose des problèmes de sémantique spatiale (voir pour plus de détails les réflexions que nous donnons dans [Bilhaut *et al.* 2003b]) : en particulier l'ambiguïté interne/externe dans l'expression *au nord de Paris*[4] qui suppose une analyse du contexte de l'expression assez lourde. Une stratégie serait de mettre à jour une sorte de « cadre spatial », c'est-à-dire le contexte granulaire du passage sélectionné[5]. En outre, une étude approfondie des prépositions spatiales serait nécessaire pour affiner les interprétations sémantiques comme dans *autour de / aux alentours de / près de... Paris*. Pour ce qui concerne les expressions temporelles, l'interprétation référentielle produit une approximation de la période évoquée représentée par un intervalle entre deux dates.

Évaluation L'analyseur spatial a été mis au point sur une partie du texte de [Hérin 1994] qui décrit le système scolaire et son évolution au cours des cinquante dernières années et testé sur une autre partie du corpus, puis sur deux autres corpus l'«ATLAS transmanche[6] » et l'«Atlas Politique[7] ». 900 expressions ont été reconnues sur le premier corpus et le nombre d'erreurs observées est extêmement faible. De plus, les tests menés sur les deux autres corpus ne font pas apparaître de nouvelles formes d'expression. En revanche, les silences observés ne sont pas marginaux. Ils proviennent de : i) des co-références non-traitées (par exemple, *ces régions* ou encore *[...] partage clairement la France en deux : la moitié méridionale où [...]*). ii) l'incomplétude du lexique : par exemple, le vocabulaire de géographie maritime (port, rade, estuaire...) manquait.

Les principes de l'analyseur ont été repris dans différents contextes. En premier lieu, une transposition sur l'allemand a été réalisée par Zvenja Tantzen [Tantzen 2004]. Une variante adaptée au corpus et à la tâche du projet PIV[8] a été développée par [Sallaberry *et al.* 2007]. On peut considérer que ces diverses expérimentations permettent de valider la méthode d'analyse.

3.1.3 Analyse discursive

L'analyse discursive concerne la structuration des documents et son repérage. Rappelons qu'il s'agit dans le cadre du projet de présenter à l'utilisateur des extraits d'une base documentaire en réponse à une requête associant trois critères (par exemple, *évolution des effectifs scolaires dans l'Ouest dans les années 1980*). Il est important de noter que repérer une corrélation entre ces trois dimensions dans un texte implique de dépasser le cadre de la phrase comme le montre l'exemple suivant caractéristique du genre textuel considéré :

4. S'agit-il de la partie à l'intérieur de Paris ou la banlieue nord ?

5. Par exemple, si ce contexte est la France ou la région parisienne, alors on penchera sans doute vers l'interprétation externe pour l'expression *au nord de Paris*, sinon interne si le contexte est celui de Paris intra-muros.

6. «Atlas électronique Transmanche» : réalisation commune de géographes des universités de Caen et de Porsmouth.

7. P. Buléon, Quarante années d'évolution politique de l'Ouest de la France, Rapport interne ESO, 2001.

8. Projet Pyrénées Itinéraires Virtuels soutenu par la Communauté d'Agglomération Pau Pyrénées (CDAPP) et la Médiathèque Intercommunale à Dimension Régionale (MIDR).

(1) **De 1965 à 1985**, le nombre de collégiens et de lycéens a augmenté de 70%, mais selon des rythmes et avec des intensités différents selon les académies et les départements. Faible *dans le Sud-Ouest et le Massif central*, modérée *en Bretagne et à Paris*, l'augmentation a été considérable *dans le Centre-Ouest, en Alsace, dans la région Rhône-Alpes et dans les départements de la grande banlieue parisienne* où les effectifs ont souvent plus que doublé. [Hérin 1994]

Dans l'exemple (1) l'expression *de 1965 à 1985* a une portée qui s'étend sur tout le paragraphe. Pour effectuer une analyse permettant de repérer cette structure, l'approche développée est fondée sur la théorie de l'« encadrement du discours » définie par [Charolles 1997]. Ce modèle décrit des segments dits « cadres de discours » homogènes par rapport à un critère sémantique (par ex. une localisation spatiale ou temporelle) spécifié par une expression détachée en initiale de phrase (*De 1965 à 1985* dans l'extrait (1)) et nommée « introducteur de cadre » (dorénavant IC). Les IC sont présentés comme des marqueurs d'indexation délimitant un univers de discours (spatial, temporel...) sous forme de « contenus propositionnels dans des blocs homogènes relativement à un critère spécifié par le contenu de l'introducteur » (op. cité p. 24). Du fait de la présence massive de critères spatiaux et temporels, la notion de cadre de discours a semblé particulièrement bien adaptée à la segmentation discursive du corpus.

Nous nous sommes donc intéressés à la détection automatique de cadres spatio-temporels. Le processus s'appuie sur le marquage des expressions spatiales et temporelles. Une expression de ce type détectée en position détachée et en début de phrase est traitée comme étant un IC. Le problème de la délimitation du cadre, c'est-à-dire de la portée de l'IC est plus délicat en l'absence de marques explicites de clôture. Nous nous sommes appuyés sur une étude linguistique menée par l'équipe toulousaine du projet [Draoulec et Pery-Woodley 2001] décrivant plusieurs indices linguistiques qui ont été exploités dans le processus et qui se sont avérés relativement fiables. Il est à noter qu'une marge d'erreurs d'une ou deux phrases sur la terminaison du cadre a peu d'incidences pour l'application en termes de recherche documentaire. Les indices utilisés sont de trois types :

- cohésion sémantique : les représentation sémantiques des expressions spatiales et temporelles permettent d'évaluer la compatibilité référentielle d'une expression temporelle (respectivement spatiale) avec l'IC de type temporel (resp. spatiale). L'apparition d'une incompatibilité sera considérée comme un indice de rupture. Par exemple, l'apparition de l'expression *l'année 2000* interromprait le cadre introduit par *de 1965 à 1985* de l'exemple (1) ;
- temps des verbes : à l'intérieur d'un cadre, le passage présent à un non-présent marque la fermeture du cadre. Certaines exceptions sont prises en compte comme le changement de temps à l'intérieur de parenthèses ;
- fin de paragraphes : ces marques sont considérés comme des indices forts de fin de cadre.

Ces critères sont assez simples et opérationnels pour une implémentation avec des résultats satisfaisants (voir *infra*). Ils ne prennent cependant pas en compte l'ensemble des

configurations mises au jour par l'étude linguistique [9].

Afin d'évaluer la segmentation des cadres temporels, Stéphane Ferrari, Frédérik Bilhaut, Antoine Widlöcher et Marion Laignelet ont réalisé une phase d'évaluation reposant sur l'annotation manuelle des cadres [Ferrari *et al.* 2005] par un expert sur le corpus [Hérin 1994]. Mes collègues ont utilisé les mesures quantitatives classiques avec un taux de précision de 92% (74 cadres ont été alignés avec les cadres de l'expert sur les 80 produits par le système, soit un bruit correspondant à 6 cadres) et un rappel de 47% (156 cadres annotés par l'expert). Parmi les 6 cadres erronés, 3 correspondent à un oubli de l'expert, 2 à un problème de protocole et 1 erreur d'analyse en amont de la détection des cadres. L'analyse des cadres « oubliés » par le système montre deux types de phénomènes : i) une part importante de ces oublis est due à la non reconnaissance des introducteurs comme *À l'occasion du mouvement national...* où la référence à un événement historique est difficilement calculable. Ce type de silence est relatif à un module en amont (l'analyse temporelle) de la reconnaissance des cadres ; ii) d'autres cadres oubliés ne correspondent pas au modèle linguistique (par exemple, les introducteurs de cadre non détachés comme *Pour la fin des années 1980 à défaut d'autres statistiques...*).

Remarquons que les critères linguistiques stricts et objectifs ne sont pas disponibles pour délimiter avec exactitude les cadres de discours, ou autres structures discursives de ce type. En particulier, la clôture d'un cadre, selon le critère sémantique de l'introducteur, peut être sujette à différentes interprétations selon les annotateurs humains ou conduire à repérer des zones en fin de cadre qualifiées de « floues » par [Ferrari *et al.* 2005]. L'évaluation en termes de précision et de rappel, par nature binaire, n'est par conséquent pas satisfaisante sur ce type d'objets mais nécessiterait la prise en compte de marges d'erreurs pour leur délimitation (voir pour notre équipe les réflexions et travaux de [Ferrari 2010] et [Mathet et Widlöcher 2011]).

Moteur de recherche Une fois établie la segmentation du document à indexer, l'analyse de la composante phénomène est effectuée : un calcul de type $tf.idf$ est appliqué afin d'isoler pour chaque segment les termes statistiquement saillants et discriminants pour indexer les documents.

Au final, le système produit une indexation sur les trois dimensions (phénomène/espace/temps) qui est exploitée par un moteur de recherche capable de répondre à des requêtes portant simultanément sur ces trois dimensions. Les résultats sont présentés à l'utilisateur en mettant en valeur des passages déterminés dynamiquement, délimités à l'aide des bornes des cadres sélectionnés [Bilhaut *et al.* 2007].

3.1.4 Bilan

J'ai participé à la mise en œuvre de l'analyse spatio-temporelle (plus spécifiquement sur la dimension spatiale) et discursive. La première produit une indexation des documents

9. Par exemple, l'enchâssement possible de cadres, ou encore le fait qu'une expression référentiellement incompatible puisse ne pas clore un cadre si elle apparaît dans une proposition relative, donc en arrière plan par rapport à la ligne générale du discours.

et la seconde permet de segmenter le texte en cadres de discours thématiquement homogènes selon le critère sémantique donné par l'introducteur. En ce qui concerne l'analyse spatio-temporelle, les principales idées sont directement issues des systèmes de compréhension automatique tels que mentionnés au chapitre précédent. La différence majeure réside dans la spécialisation du domaine traité, et dans le type d'analyse qui est ici local et ciblé. Ces restrictions rendent possibles des analyses fines, et relativement fiables, qui produisent pour chaque expression une représentation sémantique. Au niveau de la recherche d'information, ces représentations permettent d'effectuer des requêtes sur les dimensions spatiales et temporelles hors de portée des méthodes traditionnelles s'appuyant sur des descripteurs lexicaux.

Les améliorations à envisager portent d'une part sur les problèmes de sémantique référentielle évoqués précédemment, et particulièrement sur l'interprétation géorenférencée des expressions spatiales. Les approximations opérées ont cependant permis de développer un système opérationnel. D'autre part, même si les formes des expressions (spatiales et temporelles) semblent relativement stables d'un corpus à l'autre, l'adaptation des analyseurs à d'autres types de corpus suppose de nouvelles ressources (lexicales essentiellement). La problématique de leur acquisition automatique reviendra régulièrement dans ce chapitre et fera l'objet du suivant.

3.2 Extraction d'information en biomédical

Le domaine de la biologie génétique et médicale est riche en corpus de textes. La base MedLine, disponible publiquement via PubMed, compte ainsi plus de 20 millions de publications recensées, et chaque année un million de références sont ajoutées. Les besoins d'outils automatiques pour accéder à des informations ciblées et pertinentes sont donc énormes, par exemple pour la mise à jour de bases de connaissances, pour la veille technologique. La masse de données textuelles disponible fournit par ailleurs un champ d'expérimentation particulièrement attractif pour le TAL, et depuis une bonne quinzaine d'années de nombreux travaux en traitement automatique des langues appliqué à la biologie (BioNLP) ont vu le jour [Zweigenbaum *et al.* 2007].

3.2.1 Le projet Bingo

À la fin du projet GéoSem, j'ai choisi d'intégrer le projet de recherche Bingo et j'ai commencé à m'intéresser au domaine du traitement des textes biomédicaux pour expérimenter, à large échelle, les méthodes d'extraction d'information que nous avions précédemment développées. Le projet Bingo est un projet centré sur le développement de nouvelles méthodes de fouille de données avec des applications dans le champ de la biologie moléculaire. L'un des axes du projet porte sur l'utilisation de ressources textuelles notamment dans les phases de post-traitement de motifs pour la prise en compte de la connaissance du domaine et des contextes d'analyse. En effet, dans de nombreux contextes, comme l'analyse du transcriptome, un grand nombre de motifs sont extraits car *a priori* intéressants. L'idée est d'exploiter les informations extraites comme connaissance du domaine biologique afin de proposer des motifs satisfaisant des propriétés à la fois dans les données d'expression de gènes et dans les textes associés à ces gènes [Klema *et al.* 2008].

Je relate dans cette section une expérimentation conduite avec Nicolas Durand et Jiri Kléma (post-doctorants en 2005-2006) dans le cadre du projet Bingo. L'approche choisie est proche de celle suivie pour l'analyse des expressions spatiales mais ici on s'intéresse à quatre types d'information : noms de protéines, de familles protéiniques, localisation de la protéine (composant cellulaire), fonctions biologiques [Charnois *et al.* 2006].

3.2.2 Reconnaissance d'entités nommées biologiques

La reconnaissance des entités nommées biologiques revient à localiser une sous-chaîne dans la phrase et à lui attribuer une des catégories pré-définies (protéine, fonction biologique...). Plusieurs particularités rendent le problème difficile [Leser et Hakenberg 2005]. En effet, un gène ou une protéine peut être nommé par un sigle, un terme composé de plusieurs mots, ou encore par une partie ou une abréviation du terme composé ou du sigle. De plus, l'absence de nomenclatures figées du fait de néologismes, la polysémie (les mots *'pigs'*, *'set'*, *'she'* ou encore *'clock'* désignent des gènes), l'existence de plusieurs dénominations pour un même gène, la présence de signes de ponctuation divers au sein même des termes complexifient encore le processus de reconnaissance. Les familles d'approches les plus répandues dans la littérature sont de deux types. D'une part les approches à base d'apprentissage automatique qui tirent parti de la disponibilité de corpus annotés dans le domaine pour apprendre des règles d'extraction ou de classification (voir [Chang *et al.* 2006, Nédellec *et al.* 2006] pour un état de l'art et [Smith *et al.* 2008] pour un aperçu des systèmes utilisés lors du challenge BioCreative II). Nous reviendrons dans le chapitre suivant sur la problématique de l'apprentissage en TAL. L'autre type d'approche s'inscrit dans le courant du TAL symbolique [Cohen et Hunter 2004, Cohen et Hersh 2005] avec des méthodes à base de dictionnaires, ou de règles d'extraction. Les techniques à base de dictionnaires tentent un appariement exact ou partiel avec des mesures de distances comme [Tsuruoka et Tsujii 2003]. Elles sont simples à mettre en œuvre mais souffrent d'un taux de couverture bas. Un autre type de méthode effectue la reconnaissance à partir de règles qui peuvent reposer sur des expressions régulières telles que dans [Fukuda *et al.* 1998], un des premiers systèmes à base de règles. Les systèmes les plus complexes utilisent au préalable une analyse syntaxique de la phrase par exemple, [Gaizauskas *et al.* 2003].

3.2.3 Notre système

Principes Rappelons que l'on vise à extraire des entités nommées biologiques (noms de protéines, familles de protéines, localisation, fonctions biologiques) et à les typer. Afin d'obtenir rapidement un système opérationnel, nous nous sommes tournés vers le développement de règles légères, faciles à implémenter et sans analyse syntaxique. L'observation du corpus a révélé des formes récurrentes qui apparaissent dans la position qui précède immédiatement l'entité (voir figure 3.2 p. 28). L'idée, similaire à la méthode d'exploration contextuelle proposée par [Wonsever et Minel 2001], a donc consisté à localiser un contexte gauche favorable à la reconnaissance d'une entité nommée. Le processus de reconnaissance quant à lui ne fait pas d'hypothèse sur la forme de l'entité et consiste en une analyse en négatif : les éléments ne pouvant faire partie d'une entité (comme certaines ponctuations et catégories de mots : verbes conjugués, pronoms relatifs, *etc.*) permettent de stopper

Tag 1
Tag 2
GeneID 2023
ENO1 NNE PPH MPB1 MBP-1
This gene encodes one of three enolase isoenzymes ***found in*** *mammals; it encodes alpha-enolase, a homodimeric soluble enzyme, and also encodes a shorter monomeric structural lens protein, tau-crystallin. The two proteins are made from the same message. The full length protein, the isoenzyme,* ***is found in*** *the cytoplasm. The shorter protein is produced from an alternative translation start,* ***is localized to*** *the nucleus, and has been found to bind to an element in the c-myc promoter. A pseudogene has been identified that is* ***located on*** *the other arm of the same chromosome.*
Tag 3
GeneID 51192
CKLF C32 CKLF1 CKLF2 CKLF3 CKLF4 UCK-1
The product of this gene is a cytokine. Cytokines are small proteins that have an essential ***role in*** *the immune and inflammatory responses. This gene is one of several chemokine-like factor genes* ***located in*** *a cluster on chromosome 16. The protein encoded by this gene is a potent chemoattractant for neutrophils, monocytes and lymphocytes. It also can stimulate the proliferation of skeletal muscle cells. This protein may play important* ***roles in*** *inflammation and in the regeneration of skeletal muscle. Alternatively spliced transcript variants encoding different isoforms have been identified.*
Tag 6
GeneID 64409
WBSCR17 GALNT16 GALNTL3 DKFZp434I2216
This gene encodes an N-acetylgalactosaminyltransferase, which has 97% sequence identity to the mouse protein. This gene is deleted in Williams syndrome, a multisystem developmental disorder caused by the deletion of contiguous genes at 7q11.23.

FIGURE 3.2 – Extrait du corpus – Les contextes gauches identifiant une protéine sont soulignés, une localisation ou une fonction biologique sont en gras.

la reconnaissance (identifiant la frontière droite de l'entité). Ce principe avait déjà été exploité pour l'extraction de terminologie par [Bourigault 1994]. La figure 3.3 montre le schéma général des règles (et un exemple de règle) qui se résume en l'identification du contexte gauche (qui permet aussi de typer l'entité), puis l'extraction de l'entité jusqu'à la reconnaissance de la frontière droite.

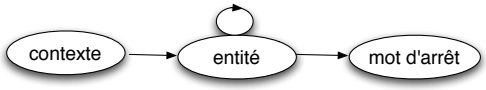
Hormis les règles, le seul traitement linguistique est une analyse morphologique et un étiquetage grammatical opérés par l'outil *TreeTagger*. L'ensemble des traitements a été effectué sous *LinguaStream*, ce qui a facilité l'utilisation de certains éléments précédemment annotés dans la chaîne de traitements : le processus opère donc en plusieurs passes (par exemple, pour reconnaître la famille de protéine X dans *X are class of FAMILY*, où FAMILY est une famille de protéine préalablement annotée).

Démarche, résultats et discussion L'étude linguistique pour l'acquisition des contextes et le développement des règles a porté sur une petite partie du corpus [10] (200 résumés soit

10. Le corpus est constitué de *summaries* qui ont été extraits du site *entrez gene* `http://www.ncbi.nlm.nih.gov/gene` à partir de la liste de gènes présents dans la matrice d'expression des gènes utilisée par ailleurs pour extraire des concepts dans les expérimentations de fouille de données (11000 gènes). Au final

5% de l'ensemble). Au final, on obtient une représentation de chaque résumé sous la forme de base d'annotations XML correspondant à des connaissances sur chaque gène, connaissances exploitables pour la fouille [Klema *et al.* 2008].

Le système a été évalué sur un échantillon à l'aide des mesures classiques en termes de rappel et précision. Le rappel estimé se situe aux alentours de 70% et la précision entre 80 et 90% selon les types d'entités, contre 80-95% en rappel et précision pour les meilleurs systèmes de la littérature. Par rapport aux autres systèmes à base de règles nos résultats sont comparables avec une bonne précision mais un rappel insatisfaisant. Notons que l'observation linguistique et préliminaire du corpus a permis de développer rapidement une grammaire légère, utilisant peu de ressources, pour extraire différents types d'entités nommées biologiques utilisables comme connaissances du domaine fiables (en qualité). En revanche, une étude approfondie des cas de silence est nécessaire pour améliorer le rappel et concevoir de nouvelles règles. Le coût de leur conception manuelle, mais aussi de leur adaptation à d'autres corpus, nous a orientés vers des méthodes pour les acquérir automatiquement. La section 4.4.2 traitera précisément de cette question avec l'idée (tirée de ce travail) d'intégrer le contexte dans les règles apprises.



prot(type :pro..name :Name) –> $'the', $protein, $'encoded', lemma :'be', np(Name).
np(N) –> ls_token(N,_), stop.
np(N) –> ls_token(N1,_), np(N2), {concat(N1,N2,N)}.
stop –> verb ; punctuation ; relative_pronoun ; triggerWord.

FIGURE 3.3 – Structure des règles de reconnaissance des entités nommées biologiques, suivi d'un extrait de la grammaire (écrite en Prolog). Cette grammaire localise le contexte *The protein encoded by this gene is*, puis concatène tous les éléments qui suivent (formant l'entité ; ici de type *pro*, pour protéine) jusqu'à rencontrer un élément d'arrêt (règle *stop*).

3.3 Analyse d'opinion

3.3.1 Projet ONTOPITEX en cours

J'aborde dans cette section un travail en cours sur un troisième domaine d'application relatif à l'extraction de l'opinion exprimée dans les textes. Ce travail est mené dans le cadre du projet ONTOPITEX qui vise la mise en place d'outils automatiques d'analyse sémantique fine des opinions et des jugements exprimés au sein de documents textuels numériques. L'approche suit un point de vue multidisciplinaire, consistant à croiser les regards de linguistes et d'informaticiens pour proposer des modèles opératoires qui s'appuient en amont sur une expertise de la langue.

le corpus comprend environ 4 000 résumés (2,33 Mo, 65 000 paragraphes) correspondant à autant de gènes avec un résumé associé (7 000 gènes n'ont pas de résumé associé, à la date de la constitution du corpus). On dispose donc potentiellement d'un ensemble de connaissances à extraire pour ces gènes.

Le projet offre donc un cadre intéressant pour articuler étude linguistique et analyse informatique et alimenter mutuellement les deux processus. Je décris plus précisément mes contributions [Jackiewicz *et al.* 2009a, Ferrari *et al.* 2009b] sur la mise en œuvre d'une expérimentation informatique dont le but est de fournir des observables permettant une étude des constituants détachés à valeur axiologique (porteurs de jugement) ainsi qu'une étude sur l'évaluation de cette expérimentation. Un deuxième aspect de mon travail portant sur l'apprentissage de ressources dans ce contexte sera présenté au chapitre suivant (section 4.3).

Positionnement L'analyse d'opinion, ou fouille d'opinion, est un champ de recherche qui est apparu en TAL assez récemment autour de la problématique de la recherche informatisée de jugements, opinions, sentiments ou émotions exprimés dans les documents textuels comme les articles de la presse journalistique ou les blogs. Cette problématique renvoie à des enjeux économiques ou sociaux d'importance en matière de veille économique, veille d'image, analyse de tendance, *etc.*

Les travaux relatifs menés en informatique peuvent être classés selon trois axes principaux : établir des ressources lexicales, catégoriser des textes, dans leur globalité, et enfin analyser des expressions plus locales, comme des phrases [11].

En ce qui concerne les ressources lexicales, l'objectif est de les constituer de manière semi-automatique. Les travaux existants se distinguent par le type de propriétés lexicales étudiées : déterminer le caractère subjectif ou objectif des termes, ou la polarité des subjectifs (positive, négative), ou encore préciser plus finement leur sens comme dans [Whitelaw *et al.* 2005] ou dans les travaux relatifs à SentiWordNet [Esuli et Sebastiani 2006], associant une combinaison des notions de subjectivité et de polarité à chaque SynSet de WordNet. L'atelier FODOP08 (FOuille des Données d'OPinions) a été l'occasion de vérifier l'existence de travaux comparables sur le français [Harb *et al.* 2008].

De très nombreux travaux s'attachent à catégoriser des textes dans leur intégralité, en termes de polarité globale (texte positif, négatif ou neutre). La campagne d'évaluation DEFT07 avait cet objectif, pour la langue française. Les domaines d'applications sont nombreux et divers ; citons l'analyse de critiques de films [Turney 2002], [Pang et Lee 2004] et une partie de DEFT07, de textes politiques (un corpus de réactions à des propositions de lois dans DEFT07), ou les critiques de produits. Les techniques utilisées sont variées, majoritairement issues des domaines de la fouille de données et de l'apprentissage automatique.

Une série de travaux s'attachent enfin à mener des analyses plus locales, à déterminer le caractère objectif ou subjectif ou encore la polarité de mots, d'expressions complexes ou de phrases dans leur intégralité, en contexte. Les approches relatives s'attachent à la classification de phrases et de propositions pour distinguer les opinions des faits dans un système de Question/Réponse [Yu et Hatzivassiloglou 2003], pour proposer un résumé des points sur lesquels portent les critiques émises par les consommateurs dans les travaux de [Hu et Liu 2004], ou encore à l'annotation en contexte d'expressions d'opinion [Legallois et

11. Je reprends ici un état de l'art que nous avons publié dans [Ferrari *et al.* 2009b]

Ferrari 2006, Vernier et Ferrari 2007]. La pertinence des annotations proposées est liée aux applications visées. On retrouve la simple détermination du caractère subjectif ou la polarité, mais une tendance se dégage pour proposer des annotations d'autres caractéristiques comme les cibles (objets des évaluations), les sources (émetteurs), l'intensité, *etc.* [Wiebe *et al.* 2005, Read *et al.* 2007].

La plupart des approches sont donc fondées sur des méthodes statistiques qui donnent les meilleurs résultats dans des tâches de classification telles que celles de DEFT'07 où l'analyse de l'opinion porte sur un texte complet en termes de polarité globale. Elles sont cependant pauvres d'un point de vue linguistique et peu adaptées à une analyse fine des différents jugements et opinions exprimés au sein des textes. Notre approche s'inscrit dans le cadre d'une analyse locale d'expressions d'opinion. Pour cela, une première expérimentation a porté sur l'extraction des constituants détachés à valeur axiologique (porteurs de jugement) à partir d'une étude linguistique menée par Agata Jackiewicz.

3.3.2 Extraction de constituants détachés axiologiques

Nous nous intéressons au phénomène discursif du détachement, et plus particulièrement à l'expression d'opinion au sein de constituants périphériques extra-prédicatifs ciblant des individus ou des organisations. Pour plus de détails sur les aspects linguistiques, nous renvoyons le lecteur à [Jackiewicz *et al.* 2009a]. Les éléments en gras dans les exemples (2) à (5) illustrent le type d'expressions qui nous intéressent :

(2) *Mais,* **en politicien expérimenté**, *élu pour la première fois à la Knesset il y a trente-cinq ans, il a su résister aux roquettes de ses adversaires politiques.*

(3) **Ni trop sentimental**, **ni trop énergique**, *il maîtrise,* **avec une finesse quasi mozartienne**, *un lyrisme généreux.*

(4) **Militant mais opportuniste**, **franc-tireur mais habile**, **sociable mais anticonformiste**, *le directeur de l'Opéra de Paris sait manier les paradoxes pour parvenir à ses fins.*

(5) **Figure légendaire de l'opposition au régime communiste**, **éminent professeur d'histoire médiévale**, **ministre des affaires étrangères de la Pologne de 1997 à 2000**, *Bronislaw Geremek avait été élu au Parlement européen en 2004.*

Ces expressions, constituants périphériques à la prédication principale, dénotent souvent des appréciations, sous formes réduites et clairement délimitées. Les constituants périphériques peuvent se caractériser par un ensemble de critères tels que : délimitation par des signes de ponctuation, possibilité d'effacement, (relative) liberté de position, statut de prédication seconde, existence d'un référent sous-jacent. L'étude linguistique menée manuellement par Agata Jackiewicz [12] a ainsi mis en évidence une vingtaine de patrons linguistiques, ainsi qu'un lexique (d'environ 550 termes) constitutif de ces expressions, correspondant à trois principales catégories :

- Constructions détachées :
  - groupes adjectivaux (*imprévisible et fantasque, X*) ;
  - constructions absolues (*l'oeil vigilant, X*) ;
  - participes (*réputé pour son caractère bourru, X*).

12. Étude réalisée sur un corpus constitué manuellement d'articles du journal « Les Échos » issus de la catégorie « Portraits ».

- Compléments circonstanciels :
  - Adverbes (*courageusement, X*) ;
  - groupes prépositionnels (avec patience, X).
- Groupes nominaux (GN) : avec ou sans déterminants (*Femme de tête,X* ; *X, le maestro de la désinflation*).

Expérimentation La poursuite de cette première étude a consisté en une extension des patrons proposés et une évaluation sur de nouvelles données. Pour mener à bien ce travail, nous avons constitué un nouveau corpus, d'un genre proche de celui ayant permis l'analyse linguistique, regroupant 884 articles de type « Portrait » ou « Biographie » issus du journal Le Monde de la période juillet à décembre 2002.

Résultats La mise en œuvre de 10 patrons parmi les 20 observés a été intégrée sous forme de règles Prolog dans un composant d'une chaîne d'analyses de la plate-forme *Linguastream*. Leur application sur ce deuxième corpus mène aux résultats suivants. 1966 expressions ont été repérées dans le corpus, dans 580 des 884 articles (65%). Leur répartition montre que 395 articles contiennent plus d'une expression, 107 plus de 5, 21 seulement plus de 10 ; le maximum d'expressions par article est 23. La dispersion sur le corpus est donc assez hétérogène. Nous renvoyons à [Jackiewicz *et al.* 2009a] pour une analyse plus détaillée de ces résultats.

L'observation qualitative des résultats permet de dégager plusieurs pistes intéressantes pour notre étude. Nous constatons une forte concentration d'expressions au sein de quelques articles, quand 35% n'en contiennent aucune, que nous expliquons par des préférences concernant les modes d'expression qui seraient propres aux auteurs. Les patrons concernant les groupes adjectivaux sont, sans surprise, les plus fréquemment rencontrés. Ceci va dans le sens des nombreux travaux de la communauté TAL qui concentrent leurs efforts sur les adjectifs pour des analyses en termes de subjectivité et de polarité. On remarque aussi quelques phénomènes intéressants à la lecture des sorties, notamment de nombreuses énumérations, parfois au sein de phrases averbales (7), et l'utilisation de connecteurs permettant des jeux d'opposition ou de renforcement, des modalités temporelles (8) (*et, tantôt, sans, à la fois, mais, toujours, jamais*, etc.). Autre fait notable : la position détachée contribue à l'interprétation axiologique de certains adjectifs (*ouvert, européen*) (9).

(6) *L'acteur d'Anouilh et de Chabrol, de Diderot et de Pinter, reste un voyageur traqué de l'âme. Inquiétant, ambigu, fascinant. (...) Laconique et inquiétant, discret et courtois, il cultive l'anonymat jusqu'à l'étrangeté.*

(7) *Froidement, avec méthode, en chemise blanche et cravate noire.*

(8) *Réputée rigide, parfois cassante, elle est toujours difficile à manoeuvrer. (...) Décriée, souvent méprisée par les caciques du parti, elle jouit malgré tout d'une véritable popularité parmi les militants.*

(9) *Elle participe à ce Portugal qui va vite, ouvert, européen, décomplexé.*

Sans avoir mené d'analyse quantitative systématique, nous pouvons toutefois qualifier les principaux cas de bruit et de silence. Certains sont bien sûr dûs à des erreurs de

prétraitement (segmentation, étiquetage grammatical). Concernant le bruit, nous notons les cas suivants : (i) la cible est un objet, pas une personne (*baroques et tendres, ils...* en parlant des livres) ; (ii) le cas des propriétés sans charge axiologique (*moustachu, aux yeux noirs, à la voix douce*) ; (iii) l'ambiguïté de la structure (*en bon parrain* versus *en pleine chiraquie*). Quant au silence, il est essentiellement dû aux facteurs suivants : (i) constructions contraintes ne laissant pas de place à des modifieurs ; (ii) restrictions lexicales fortes pour certains patrons.

3.3.3 Annotation manuelle pour l'évaluation et poursuites

Pour obtenir une évaluation quantitative de ces phénomènes à partir d'un corpus de référence, une campagne d'annotation manuelle a été mise en place. Elle s'est déroulée en deux temps [Ferrari *et al.* 2009a].

Une première campagne « ouverte » demandait aux annotateurs de délimiter les segments détachés, les qualifier (axiologique versus qualificatif), d'ajouter des traits correspondant à des propriétés sur le jugement (polarité, prototypicalité, intensité), et, si possible, de repérer la cible du jugement. Les taux d'accord [13] entre les 4 annotateurs se sont révélés mauvais, voire très mauvais : 0,13 sur la simple délimitation des segments (les 6 accords deux à deux vont de -0,1 à 0,36) et -0,06 sur les seuls segments axiologiques (les 6 accords deux à deux variant entre 0,15 à 0,53). Devant la multiplicité des tâches proposées, les experts se sont probablement concentrés sur des objectifs différents et leur lecture des textes en a sans doute été influencée.

Une deuxième campagne, qui sollicite moins l'attention, a été effectuée. À partir des 181 segments repérés lors de la première phase d'annotation, il s'agissait simplement de préciser si le segment était de nature axiologique ou non. Les taux d'accord entre les 6 annotateurs sont nettement meilleurs : 0,72 (les taux deux à deux variant de 0,56 à 0,87).

Ces expériences ont mis en lumière la complexité et le coût de cette tâche. Nous avons mené des réflexions pour diminuer ce coût et envisagé deux approches :

- une annotation par les experts des segments qu'ils jugent constituants détachés axiologiques sur textes bruts ;
- une annotation automatique préalable de segments pour que les experts aient simplement à vérifier ou éventuellement à corriger les résultats de l'analyse automatique.

La première approche est la plus simple, et convient pour tester les patrons mentionnés *supra*, mais il convient de noter le coût qui reste non négligeable et la portée restreinte de l'annotation : une annotation complexe incluant plusieurs facettes du jugement (polarité, intensité, cible...) serait plus intéressante pour une analyse automatique de ces facettes. Mais comme l'ont montré nos deux campagnes, il convient de ne pas complexifier les facettes à annoter et il faudrait envisager, si l'on poursuit à l'extrême, une campagne pour chaque propriété étudiée... La deuxième approche est quant à elle séduisante, mais comporte le risque que les silences ne soient pas vus par les experts.

13. mesure *Kappa*

Nous avons exploré ce type d'approche automatique mais en utilisant des méthodes d'apprentissage pour acquérir des segments potentiellement porteurs de jugement et un outil d'exploration pour permettre leur validation aisément. Ce travail est décrit en section 4.3.

3.3.4 Autres poursuites

Les expressions extraites ont fait l'objet d'une étude approfondie de leurs caractéristiques sémantiques par [Jackiewicz 2010]. Par ailleurs, avec Stéphane Ferrari nous avons expérimenté quelques patrons sur un corpus volumineux (l'ensemble des articles du journal « Le Monde » sur la période 1987-2002 représentant 2,2 Go de données) pour acquérir des lexiques. Par exemple, sur le patron $<$ *{réputé, salué} pour ART_POSS* [*ADJ*] *NOM* $>$, nous avons fixé la partie gauche pour collecter la partie droite c'est-à-dire un nom (éventuellement précédé d'un adjectif). Cette expérience a permis de collecter 590 attestations pour ce patron qui ont aussi été étudiées finement par A. Jackiewicz qui a pu dégager des catégories sémantiques [14]. L'ensemble des patrons et lexiques fournit ainsi des ressources linguistiques. Une réflexion est en cours sur leur utilisation au sein d'outils de TAL en vue de l'extraction et de l'analyse de certaines facettes de l'opinion (polarité, prototypicalité, intensité) dans les textes.

3.4 Bilan

Les travaux présentés dans ce chapitre s'appuient sur un modèle ou une étude linguistique sur un phénomène ciblé pour extraire de l'information des textes à des fins applicatives (recherche documentaire, extraction de connaissances) ou encore pour affiner un modèle linguistique en fournissant des observables (analyse d'opinion). Le développement des ressources linguistiques se fait à partir de l'observation des corpus. Plus précisément, le concepteur de règles d'extraction cherche des régularités linguistiques dans les textes. Ce processus est long donc coûteux en temps de développement. En terme de résultat, les règles ont généralement une bonne précision mais souffrent d'un rappel moins bon (une bonne couverture des règles est difficile à assurer lorsque la conception des règles est manuelle). Automatiser ce processus nous a ainsi paru une voie à explorer que nous proposons et développons au chapitre suivant.

14. À titre d'illustration, les expressions récurrentes mettent en évidence une caractéristique saillante et valorisée d'une cible : *sentiments nationalistes, jugements tranchés, sens caustique...* [Jackiewicz 2010] .

Chapitre 4

Fouille de données pour le TAL

L'analyse automatique de textes requiert des ressources linguistiques (lexiques, patrons linguistiques...). Elles sont de plus dédiées à un domaine spécifique. Nous nous sommes alors intéressés aux méthodes permettant de les acquérir automatiquement.

L'apprentissage automatique en TAL a déjà fait l'objet de nombreux travaux (voir [Hobbs et Riloff 2010] et [Tellier 2009]). Historiquement, deux familles d'approches peuvent être considérées. Les approches numériques ou statistiques telles que les modèles de Markov caché, machines à vecteur de support, champs aléatoires conditionnels, *etc.* sont extrêmement efficaces sur les tâches de classification. Les tâches d'extraction sont vues par ces approches comme un problème de classification où le principe consiste à parcourir le texte et à décider à quelle classe appartient un mot (par exemple pour la reconnaissance d'entités nommées). Ces approches sont entièrement automatiques et obtiennent les meilleurs résultats dans les challenges (voir par exemple [Smith *et al.* 2008] pour la reconnaissance d'entités nommées biologiques, [Krallinger *et al.* 2008] pour la détection d'interaction entre protéines), mais elles nécessitent des corpus annotés pour l'apprentissage des modèles, ce qui est un processus coûteux. De plus, les classifieurs ont un fonctionnement de type boîte noire car les modèles obtenus, à base de descripteurs numériques, sont difficilement interprétables ou exploitables comme ressources linguistiques du point de vue qualitatif ou explicatif. De leur côté, les approches symboliques visent à apprendre des ressources linguistiques. On peut citer les approches dans la lignée de [Hearst 1992] concernant l'acquisition des relations lexicales (exemple l'hyperonymie). Les méthodes relatives à l'apprentissage des patrons exprimant des relations entre entités nommées sont majoritairement de type supervisé. Les techniques utilisées peuvent être fondées sur la programmation logique inductive ou l'apprentissage relationnel (comme pour le système RAPIER [Califf et Mooney 1999]), ou encore sur la recherche de chemins pertinents sur l'arbre syntaxique de la phrase (par exemple [Schneider *et al.* 2009] pour l'extraction d'interaction entre gènes dans les textes biologiques). Certains travaux comme ceux de [Riloff 1996] effectuent l'apprentissage sur un corpus brut préalablement découpé en textes considérés comme pertinents ou non pour les relations cherchées. Le principe consiste à amorcer l'apprentissage avec quelques règles d'extraction élaborées manuellement (*seed patterns*) pour générer l'ensemble des patrons candidats et un processus sélectionne les candidats pertinents pour la tâche visée. Pour prendre en compte les diverses formes syntaxiques que peuvent prendre les relations entre entités nommées, relations qui peuvent être distantes dans la phrase, l'apprentissage

des règles, ainsi que leur application, sont réalisés à partir d'une analyse plus ou moins profonde de la phrase.

Les méthodes dites de fouille de données à base de motifs ont pour objectif la découverte d'informations nouvelles à partir de bases de données [Han 2005, Agrawal *et al.* 1993]. Le cœur des processus de fouille est la recherche de régularités dans les données ; on appelle motifs de telles régularités. Par exemple, à partir de situations biologiques décrites par un ensemble de gènes, un motif est un ensemble de gènes qui se retrouvent fréquemment dans de nombreuses situations biologiques. De plus, si on retrouve dans les textes biologiques ces mêmes gènes en interaction, ce motif présente alors un intérêt biologique potentiel : il est candidat pour être un groupe de « synexpression » (les gènes sont simultanément sur-exprimés dans des données différentes).

Outre l'extraction de motifs fréquents, la fouille de données présente aussi l'intérêt de pouvoir être appliquée dans un cadre non supervisé. Dans le domaine de la fouille de textes, de nombreux travaux ont pour but l'extraction de règles d'association [1]. Les règles extraites sont considérées comme de nouvelles connaissances découvertes [Janetzko *et al.* 2004, Roche *et al.* 2004] ou encore elles peuvent être utilisées pour caractériser des textes [Turmel *et al.* 2003]. Ces travaux se situent dans un paradigme ensembliste, c'est-à-dire sans intégrer la dimension séquentielle de la langue. Nous verrons que cette dimension est un point central dans nos contributions.

Contributions Dans ce contexte, nous avons proposé d'exploiter les méthodes de fouille de données fondées sur des motifs séquentiels pour : i) la prise en compte de la relation d'ordre temporelle entre les données textuelles ; ii) le caractère symbolique et interprétable des motifs qui permet une analyse qualitative ; iii) la possibilité de découvrir les motifs dans un cadre non supervisé. Dans ce chapitre, je présente notre démarche pour la découverte de patrons linguistiques à travers trois contributions :

(1) l'extraction de relations entre entités nommées (section 4.2) ;
(2) la découverte de patrons pour extraire des constituants détachés porteurs de jugement ou de qualification (section 4.3) ;
(3) la reconnaissance d'entités nommées (section 4.4).

Notre démarche est caractérisée par les idées suivantes :

Sur le plan du TAL, l'apprentissage ne nécessite pas d'analyse syntaxique ni de ressources linguistiques autres que les corpus d'apprentissage et l'utilisation d'un analyseur de type *treeTagger*. Pour la tâche (1), seules les entités nommées sont pré-étiquetées - et non les relations comme dans les méthodes supervisées, et pour la deuxième tâche, les corpus sont bruts. L'apprentissage est donc non supervisé sur ces tâches. La fouille est effectuée sur des traits linguistiques (qui sont ici limités à la forme du mot, à son lemme, et aux informations morpho-syntaxiques, mais d'autres attributs linguistiques sont possibles). Le nombre de motifs obtenu est fortement restreint, ce qui permet de limiter le coût de la

1. Motif exprimé par une relation de la forme $A \rightarrow B$ où A et B sont des ensembles disjoints signifiant que si on a A, alors on trouvera aussi **fréquemment** B.

validation manuelle, (sur la tâche (3), les motifs sont validés automatiquement car l'apprentissage est supervisé) et ceux-ci sont applicables en tant que patrons linguistiques. Sur la tâche (2), nous montrons que les patrons découverts peuvent être composés de différents types d'information associés à un mot (lemme, catégorie grammaticale...). Ils peuvent être très spécifiques – composés de la forme seule des mots –, être très généraux – catégorie grammaticale seule –, ou combiner les informations.

D'un point de vue fouille, nous montrons l'intérêt de concevoir de nouvelles méthodes de fouille combinant la découverte de motifs séquentiels composés d'items (simples éléments) ou d'itemsets (ensemble d'éléments) à partir de données textuelles. Ainsi, pour la reconnaissance d'entités nommées, nous proposons un nouveau type de motif qui combine une séquence et son voisinage sous forme ensembliste. Sur le plan linguistique, ce voisinage permet la prise en compte du contexte gauche et droit pour la reconnaissance de l'entité. Pour l'extraction des motifs, notre approche se caractérise par l'utilisation de contraintes variées en plus de la fréquence des motifs, et notamment la prise en compte d'informations linguistiques dans le processus de fouille sous forme de contraintes. En s'appuyant sur le cadre de la fouille récursive sur les ensembles d'items, nous avons adapté les techniques de fouille récursive aux séquences. Nous montrons qu'il est ainsi possible, sur des données caractérisées par une dimension séquentielle, de bénéficier du paradigme de la fouille récursive qui permet de découvrir un nombre borné de motifs récursifs ayant des capacités de généralisation. La tâche (1) illustre l'apport de ces motifs pour l'extraction et la caractérisation de relations entre entités nommées. Comme nous l'indiquons dans le dernier chapitre, nous pensons que ces méthodes ont aussi une portée dans d'autres problématiques relevant de la fouille de données.

4.1 Indices linguistiques pour la fouille : les défis DEFT

J'ai commencé à m'intéresser à la fouille de données lors de défis DEFT et du projet Bingo (*cf.* section 3.2 p. 26). Les défis DEFT éditions 2006, 2007 et 2008, auxquels j'ai pris part avec plusieurs membres de notre équipe [2] ont été l'occasion de combiner des compétences en TAL et en fouille de données. Nous avons mis en œuvre un classifieur supervisé fondé sur des règles d'association et s'appuyant sur des indices linguistiques fournis par des chaînes de traitement adaptées aux tâches spécifiques de chaque défi.

Contexte Les trois défis DEFT auxquels nous avons participé avaient les objectifs suivants [Widlöcher *et al.* 2006, Vernier *et al.* 2007, Charnois *et al.* 2008] :

- lors du défi DEFT 2006, il s'agissait de reconnaître des segments thématiques appartenant à différents domaines (juridique, scientifique, politique). Il s'agissait plus précisément de déterminer les premières phrases de chaque segment, ce qui peut revient à attribuer à une phrase donnée une catégorie (début de segment ou non) ;
- la tâche de DEFT'07 consistait à attribuer une classe d'opinion à chaque texte du corpus. Les corpus proposés étaient de quatre types : critiques de films et de livres,

2. Selon les années ont participé : Frédérik Bilhaut, Antoine Doucet, Patrice Enjalbert, Stéphane Ferrari, Nicolas Hernandez, Yann Mathet, François Rioult, Antoine Wildöcher

critiques de jeux vidéo, relectures d'articles scientifiques et débats parlementaires. Il s'agissait d'attribuer pour l'ensemble d'un texte une catégorie parmi trois valeurs (positif, neutre, négatif) ou deux pour les textes parlementaires (favorable ou défavorable à l'amendement).
– enfin, le thème de l'édition 2008 de DEFT concernait la classification en genres (Le Monde, Wikipedia) et catégories thématiques (sport, art, économie, *etc.*).

Les domaines d'application étaient donc divers, mais ont en commun le fait d'être traités comme des problèmes de classification. Les méthodes les plus employées par les participants de ces défis sont celles qui utilisent des classifieurs numériques (machines à vecteur de support, réseaux de neurones, méthodes probabilistes...). Sur l'ensemble des défis, ce type de méthodes a donné les meilleurs résultats, relativement au F-score [3] [Azé *et al.* 2006, Paroubek *et al.* 2007, Hurault-Plantet *et al.* 2008].

Réalisations L'originalité de l'approche proposée par notre équipe repose sur des traitements symboliques combinant traitements linguistiques et fouille de données. Cette approche se répartit ainsi :
– une phase de modélisation pour dégager des critères linguistiques génériques et pertinents pour la tâche et le domaine ;
– une chaîne de traitement pour repérer ces indices et les marquer dans le texte ;
– un processus de fouille produisant des règles de classification (règles d'association concluant sur la classe) à partir du marquage textuel des indices ;
– une application du modèle obtenu sur le corpus de test.

J'illustre notre méthode sur la tâche de classification en genres de l'édition 2008. La phase de modélisation s'est appuyée sur une étude des deux corpus à traiter qui a permis de dégager des hypothèses quant aux marques linguistiques caractéristiques des deux types de textes. Ceux de Wikipedia, par nature encyclopédique, nous ont semblé relever d'un style définitoire que souligne l'usage fréquent de marques comme « être un », « désigne », « définit », *etc.* À l'opposé, les articles du Monde plus narratifs incluent des formes langagières diverses : marques énonciatives typiques de l'interview (pronoms personnels des premières et deuxièmes personnes, marques de citation), marques interrogatives, exclamatives, anaphoriques, formes impersonnelles, temps verbaux du passé, futur, conditionnel (versus le présent atemporel de la définition). Le repérage de ces marques a été réalisé en tête de texte, c'est-à-dire en première phrase ; l'hypothèse étant que cette position joue un rôle privilégié dans l'organisation discursive, et en particulier, que certaines marques discriminantes des genres textuels sont situées dans cette position.

Ces différentes marques ont été repérées dans les textes (plus précisément sur la première phrase). En bout de chaîne, un composant produit une matrice dans laquelle chaque ligne correspond à un texte du corpus et chaque colonne à un attribut étiqueté par un indice. La valeur de cet indice est le nombre d'occurrences de l'objet ou 0 si l'indice est absent de la première phrase. Les règles de classification – c'est-à-dire les règles d'association – sont produites automatiquement à partir de la matrice. Quelques expériences rapides ont permis de constater que la seule règle $/\hat{e}tre + d\acute{e}terminant \rightarrow Wikipedia/$, de confiance 73% (présente 4703 fois dans les 6398 textes de classe Wikipedia et uniquement 446 fois

3. La mesure utilisée est la moyenne harmonique entre la précision et le rappel.

dans les 88825 textes du Monde), permet d'obtenir sur l'échantillon d'apprentissage un F-score moyen de 85.2%. Le F-score théorique d'un classifieur utilisant uniquement cette règle est de 86%, qui est le résultat obtenu sur les données de test. Nous avons tenté de perfectionner cette règle et d'utiliser des règles moins fréquentes et linguistiquement justifiées, mais les performances étaient moins bonnes et nous n'avons pu améliorer les performances du classifieur. Le bon résultat obtenu confirme cependant notre hypothèse de départ sur ce simple indice. Les autres indices étant nettement moins fréquents sur la première phrase – et probablement disséminés sur l'ensemble du texte –, les règles correspondantes sont peu utilisées par le classifieur.

Bilan Du point de vue de la modélisation, l'intérêt de cette approche est donc de pouvoir observer la pertinence de marqueurs linguistiques à partir des régularités exhibées par les règles d'association (comme nous avions commencé à le faire lors de notre troisième participation et que nous venons de décrire). Le calendrier des défis n'a cependant pas permis ce regard linguistique lors de nos deux premières participations. Sur les trois participations, les résultats obtenus par notre méthode se situent autour de la moyenne des participants. Nous pensons que l'originalité forte de notre participation a été de montrer l'intérêt de l'utilisation de méthodes de fouille sur les textes, notamment pour faire émerger des régularités linguistiques, ainsi que le besoin d'outils de fouille mieux adaptés au matériau linguistique. Cet axe de recherche est devenu une thématique de recherche au sein de l'équipe et une réelle dynamique s'est ainsi instaurée. Suite à nos participations à ces défis, nous avons commencé à investir les méthodes de fouille pour le TAL et la problématique de l'accès à l'information. Ce travail, toujours en cours, est décrit dans la suite de ce chapitre.

4.2 Contraintes et fouille récursive pour l'extraction de relation entre entités nommées

Cette section présente notre méthode pour la découverte des règles d'extraction entre entités nommées que nous avons appliquée pour la détection d'interactions entre gènes [Charnois *et al.* 2009] ainsi que pour la caractérisation des interactions [Cellier *et al.* 2010a]. Une originalité de notre méthode est de tirer parti de la capacité des méthodes de fouille séquentielle à faire émerger des régularités sous forme de motifs afin de produire des résultats symboliques – patrons linguistiques – intelligibles par un utilisateur humain, et sans utiliser d'analyse syntaxique[4]. L'idée clef est d'appliquer un processus de fouille récursif et d'intégrer des contraintes pour l'extraction des motifs, nous verrons que la combinaison de ces deux techniques permet de se focaliser sur un nombre borné de motifs faisant sens d'un point de vue linguistique.

Contexte La fouille de données présente en effet l'intérêt de découvrir de nouvelles connaissances dans des bases de données, et en particulier, la fouille séquentielle permet de prendre en compte une relation d'ordre entre les données (*e.g.*, le temps) [Agrawal et

4. Nous n'écartons pas *a priori* ce type d'analyse, mais nos premières expérimentations ont montré des résultats tout à fait satisfaisants sans mettre en place une procédure souvent qualifiée de « traitement lourd » [Nazarenko 2006, p. 60] . Des expérimentations seront à réaliser, par exemple, avec une analyse de surface en *chunks* afin d'en mesurer l'impact.

Srikant 1995]. Nous donnons ci-dessous de façon intuitive quelques notions utiles pour la suite :

- une séquence est une liste ordonnée de littéraux appelés *items* (à chaque item est associée une estampille temporelle représentée par un entier : $\langle(4, a), (5, b), (7, c), (8, d)\rangle$ [5]). Un motif séquentiel est simplement une séquence ;
- une séquence s' est une sous-séquence de s (s contient s') si tous les éléments de s' sont présents dans s et dans le même ordre (on note $s' \preceq s$). Par exemple, $\langle b\ d\rangle \preceq \langle a\ b\ c\ d\rangle$. Dans le tableau 4.1 page 41, le motif $m_1 = \langle AGENE\ inhibit@vvd\ AGENE\rangle$ est une sous-séquence des séquences $S2$ et $S3$;
- le support absolu d'une séquence s dans une base de séquences SDB est le nombre de séquences de SDB qui contiennent s. Le motif m_1 a donc un support de 2 dans la base du tableau 4.1 puisqu'il est contenu dans $S2$ et $S3$;
- étant donnés une base de séquences SDB et un seuil de support minimal *minsup*, le problème de l'extraction de motifs séquentiels fréquents est de retourner l'ensemble complet des séquences s contenues dans SDB qui ont un support supérieur ou égal à *minsup*. Si on reprend l'exemple du tableau 4.1, pour un support minimal fixé à 2, on obtient ainsi l'ensemble des motifs ci-dessous :

  $\langle AGENE\rangle$; $\langle AGENE\ AGENE\rangle$; $\langle AGENE\ AGENE\ .@sent\rangle$; $\langle AGENE\ mrna@np\rangle$; $\langle AGENE\ mrna@np\ .@sent\rangle$; $\langle AGENE\ .@sent\rangle$; $\langle AGENE\ in@in\rangle$; $\langle AGENE\ in@in\ .@sent\rangle$; $\langle AGENE\ inhibit@vvd\rangle$; $\langle AGENE\ inhibit@vvd\ AGENE\rangle$; $\langle AGENE\ inhibit@vvd\ AGENE\ .@sent\rangle$; $\langle AGENE\ inhibit@vvd\ .@sent\rangle$; $\langle of@in\rangle$; $\langle of@in\ AGENE\rangle$; $\langle of@in\ AGENE\ mrna@np\rangle$; $\langle of@in\ AGENE\ mrna@np\ .@sent\rangle$; $\langle of@in\ AGENE\ .@sent\rangle$; $\langle of@in\ mrna@np\rangle$; $\langle of@in\ mrna@np\ .@sent\rangle$; $\langle of@in\ .@sent\rangle$; $\langle mrna@np\rangle$; $\langle mrna@np\ .@sent\rangle$; $\langle .@sent\rangle$; $\langle in@in\rangle$; $\langle in@in\ .@sent\rangle$; $\langle inhibit@vvd\rangle$; $\langle inhibit@vvd\ AGENE\rangle$; $\langle inhibit@vvd\ AGENE\ .@sent\rangle$; $\langle inhibit@vvd\ .@sent\rangle$;

Il faut noter que la fouille séquentielle est un paradigme plus générique que celui des n-grammes. Un n-gramme peut être vu comme une instance spécifique d'un motif séquentiel. Le désavantage des n-grammes réside dans la taille qui est fixée *a priori* (par le n), alors que les motifs séquentiels découverts sont *a priori* de taille quelconque (on peut cependant contraindre la taille si besoin). D'autre part, les items (*i.e.* les mots du texte si un item représente un mot) des séquences ne sont pas nécessairement contigus ; ce qui offre le grand avantage de pouvoir découvrir des relations distantes entre items (ou mots).

L'extraction de motifs doit faire face à deux grands problèmes que sont la définition d'outils efficaces sur de grands volumes de données et la sélection de motifs potentiellement intéressants. En effet, le processus d'extraction de motifs peut retourner une collection trop importante de motifs pour être exploitable. L'extraction de motifs sous contraintes est généralement considérée comme une solution pour améliorer à la fois la qualité des motifs extraits et le processus de fouille ; ainsi, bon nombre d'approches génériques pour l'extraction sous contraintes de motifs ensemblistes et de motifs séquentiels sont-elles proposées dans la littérature [Pei *et al.* 2002, Garofalakis *et al.* 1999, Leleu *et al.* 2003]. Par exemple, la contrainte *gap* impose que pour que S_1 soit contenue dans S_2 il faut que chaque couple d'items adjacents de S_1 ne soit pas séparés dans S_2 par plus d'un certain nombre d'items.

5. Par simplification, et lorsqu'elles ne sont pas nécessaires à la lecture, les estampilles temporelles ne seront pas mentionnées, on notera alors : $\langle a, b, c, d\rangle$

ID	Sequence
...	...
S1	⟨ *Recent@jj study@nns have@vhp suggest@vvn that@in/that AGENE may@md be@vb vital@jj for@in regulation@nn of@in AGENE mrna@np expression@nn and@cc telomerase@nn activity@nn .@sent* ⟩
S2	⟨ *injection@nn of@in AGENE mrna@np in@in xenopus@np embryo@nns inhibit@vvd the@dt AGENE mediate@vvd dorsal@jj axis@nn duplication@nn .@sent* ⟩
S3	⟨ *interestingly@rb ,@, AGENE inhibit@vvd ligand@nn -@ : dependent@jj AGENE -@ : AGENE heterodimerization@nn ,@, when@wrb constitutively@rb express@vvn in@in yeast@nn .@sent* ⟩
...	...

TABLE 4.1 – Extrait d'une base de séquences contenant des interactions : la colonne de gauche indique l'identifiant de la séquence donnée dans la colonne de droite. Chaque item de la séquence représente un mot donné par son lemme suivi de l'information issue de *Treetagger*, par exemple *jj* adjectif, *nns* nom commun au pluriel, *vvp* pour un verbe au participe passé, *etc.*, voir :
www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html. Les phrases associées aux séquences sont :
"*Recent studies have suggested that c-myc may be vital for regulation of hTERT mRNA expression and telomerase activity.* " (S1)
"*Injection of frpHE mRNA in Xenopus embryos inhibited the Wnt-8 mediated dorsal axis duplication.*" (S2)
"*Interestingly, L7 inhibited ligand-dependent VDR-RXR heterodimerization, when constitutively expressed in yeast.* " (S3)

Ce nombre est appelé *maxgap*. Dans notre approche, nous utilisons à la fois des contraintes syntaxiques (au sens formel, comme la taille du motif, le *gap*, *etc.* [Pei *et al.* 2002]), mais aussi des contraintes sur l'information linguistique.

4.2.1 Méthode proposée pour l'apprentissage des patrons linguistiques

Aperçu de la méthode Le fonctionnement général de la méthode est donné figure 4.1 et se compose classiquement d'une phase d'apprentissage et d'une phase d'application des patrons. Une base de séquences textuelles est initialement créée à partir d'un ensemble de phrases contenant des relations entre entités nommées (interactions entre gènes) dont les mots sont étiquetés par l'outil *TreeTagger* (étape 1). Dans l'étape 2, les motifs séquentiels fréquents sont extraits à partir de la base de séquences textuelles. L'étape 3 vise à contraindre ces motifs afin de sélectionner un sous-ensemble de motifs séquentiels qui respecte un ensemble de contraintes. L'ensemble obtenu peut encore être trop important pour être manuellement validé, nous réduisons aussi cet ensemble en procédant à une fouille récursive des motifs qui permet de limiter et de maîtriser le nombre de motifs (étape 4). L'ensemble des motifs restants peut alors être analysé et validé par un expert (étape 5). Les motifs séquentiels validés forment alors l'ensemble des patrons linguistiques qui sont ensuite appliqués pour détecter des interactions entre gènes dans des textes biomédicaux (étape 6).

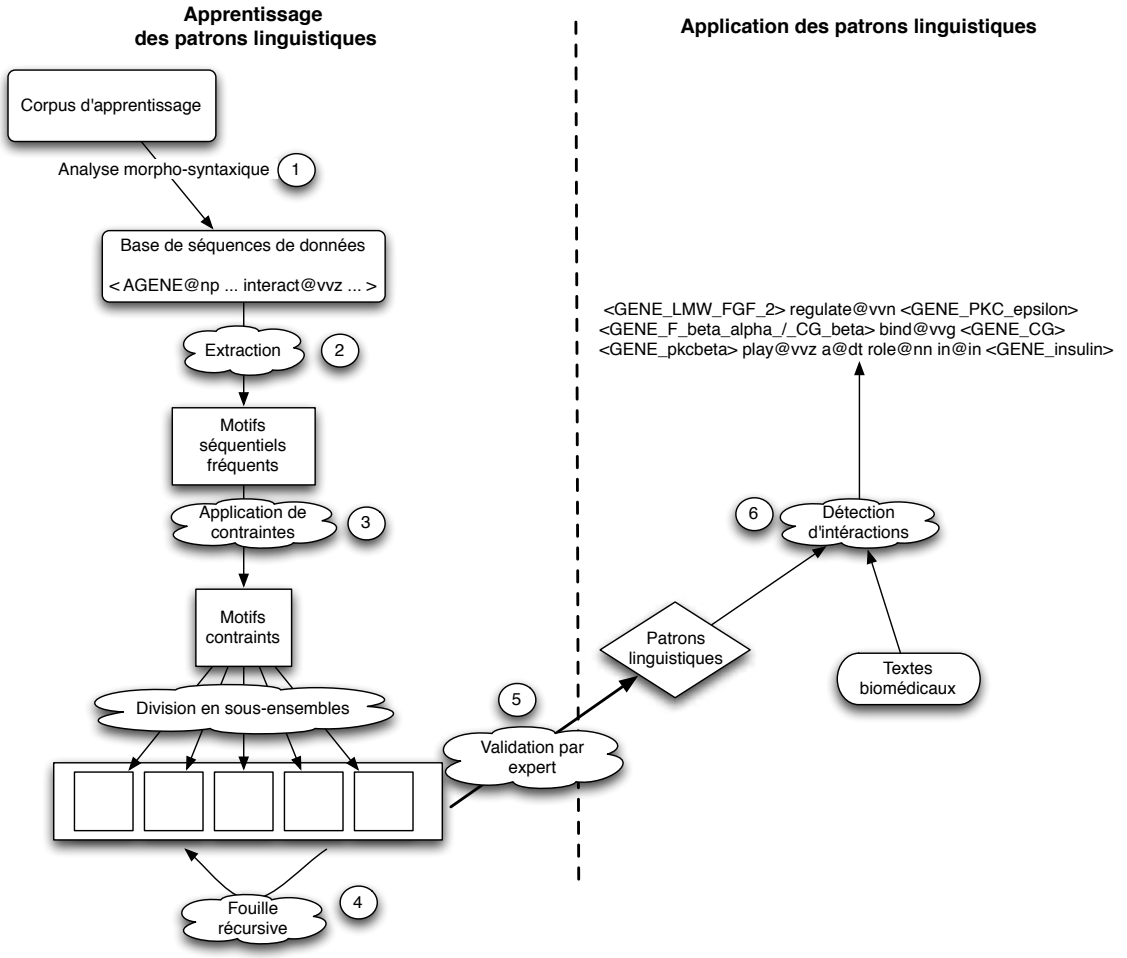


FIGURE 4.1 – Schéma général de l'approche

Données d'entrée pour l'extraction des motifs Appliquer la découverte de motifs séquentiels à des données textuelles suppose de pré-traiter les textes pour les adapter à la fouille et fixer le seuil de support :

Pré-traitement Une séquence de données se caractérise par sa durée et sa granularité : en termes de données textuelles, il faut associer à la séquence un type de segment textuel – phrase dans le cas des relations entre entités nommées, mais ce peut être la proposition, le paragraphe... – et à l'item, une unité du segment considéré – ici le mot auquel nous accolons les informations issues de *TreeTagger*, par exemple, $interact@vvz$ – et nous remplaçons les entités nommées par un item particulier jouant le rôle de jeton générique ($AGENE$) : un extrait de la base de séquences à partir de laquelle les motifs sont extraits est donné par le tableau 4.1.

Seuil de support Le choix du seuil de support minimal est un problème récurrent en fouille de données. Si le seuil de support est trop élevé, le risque est d'extraire uniquement des généralités qui n'apporteront rien à l'utilisateur. Si le seuil de support est trop faible, l'ensemble des motifs fréquents extraits peut être extrêmement volumineux rendant impossible toute utilisation. L'option choisie est de considérer

un seuil de support faible, afin de conserver le maximum d'informations, notamment les motifs peu fréquents dans les textes mais pouvant être pertinents, et de réduire l'ensemble des motifs séquentiels fréquents en introduisant des contraintes supplémentaires différentes de la contrainte de fréquence utilisée et aussi la démarche de fouille récursive.

Idées clefs Pour réduire l'ensemble des motifs extraits et cibler les motifs intéressants, nous utilisons trois contraintes, dont deux de type linguistiques, et un processus de fouille récursif :

Contraintes Les contraintes permettent d'intégrer des connaissances du domaine. Les relations entre entités nommées comme les interactions entre gènes s'expriment par un verbe ou un nom, et deux entités nommées. On ne gardera donc que les motifs satisfaisant ces deux contraintes. Afin de réduire la redondance des motifs séquentiels, nous considérons les séquences maximales fréquentes (par rapport à l'inclusion de séquences) : un motif séquentiel fréquent $S1$ est maximal s'il n'existe pas de motif séquentiel fréquent $S2$ tel que $S2$ contient $S1$. L'ensemble des motifs fréquents extraits ($SAT(\mathcal{C}_G)$) doit donc satisfaire la conjonction de ces trois contraintes.

Fouille récursive Même si l'application des contraintes réduit le nombre de motifs, cet ensemble peut rester trop important pour être analysé et validé par un utilisateur humain. La fouille récursive de motifs est une approche qui permet de limiter et de maîtriser la taille de la collection motifs extraits. Si une telle procédure a déjà été introduite dans le contexte de la fouille ensembliste [Soulet 2007], nous avons montré sa pertinence, ainsi que d'importantes propriétés formelles (correction et arrêt de l'algorithme), dans le cadre de la fouille séquentielle [Cellier *et al.* 2010b].
L'idée est de réduire le nombre de motifs extraits précédemment pour obtenir les motifs les plus représentatifs. Le principe de la fouille récursive est d'appliquer la fouille aux motifs extraits (les motifs extraits sur une base de séquences deviennent à leur tour la base de séquences qui va être fouillée), et de réitérer tant que le nombre de motifs extraits reste supérieur à un seuil fixé k. Dans la pratique, nous appliquons ce processus sur l'ensemble des motifs obtenus en utilisant les contraintes précédentes et divisés en sous-ensembles E_{X_i} où le sous-ensemble E_{X_i} regroupe tous les motifs séquentiels de $SAT(\mathcal{C}_G)$ contenant l'item X_i. Plus formellement, $E_{X_i} = \{s \in SAT(\mathcal{C}_G)\ t.q.\ \langle X_i \rangle \preceq s\}$. Le but étant de conserver des motifs pour chaque forme de relation sémantique, on considère le verbe ou le nom comme la tête sémantique de la relation, et chaque item X_i vaut alors un nom ou un verbe spécifique ($interact@vvz$ et $interact@vvn$ étant deux items X_i distincts). Chaque sous-ensemble E_{X_i} est ainsi fouillé récursivement afin d'extraire les motifs séquentiels fréquents vérifiant les trois contraintes globales introduites précédemment. Dans [Cellier *et al.* 2010b] nous avons montré que la récursivité s'arrête, et que le nombre de motifs séquentiels extraits vérifiant les trois contraintes est inférieur ou égal à k, le seuil fixé au départ.

Expérimentation La méthode a été appliquée à la détection des interactions entre gènes dans les textes biologiques. Elle a permis non seulement d'extraire les interactions mais aussi de les caractériser [Cellier *et al.* 2010a]. Il est intéressant de décrire brièvement cette étude de cas pour un aperçu concret du nombre de motifs traités et pour montrer l'aspect

relativement simple de la mise en œuvre de la méthode. Le corpus d'apprentissage se compose d'un peu moins de 5 000 phrases contenant des interactions entre gènes / protéines. Pour la fouille de motifs [6], le seuil a été fixé à 10, ce qui est un seuil bas (0.2% en valeur relative) [7] et le nombre de motifs extraits est élevé (32 millions). Il reste, après l'application des contraintes mentionnées *supra*, 65 000 motifs. On constate aussi que l'utilisation des contraintes réduit très fortement le nombre de motifs, mais ceux-ci restent encore en nombre trop élevé pour une étude manuelle. Le processus de fouille récursif permet encore de réduire ce nombre. Ainsi en fixant le paramètre k à la valeur 4, on conserve pour chaque nom et pour chaque forme verbale (515 en tout) au plus 4 motifs (ce qui correspond à 4 « constructions syntaxiques » pour chacun d'eux, par exemple *activate@vvz*). À l'issue de la fouille récursive effectuée sur chaque sous-ensemble, il reste au total 667 motifs qui ont été analysés par deux utilisateurs : approximativement un tiers des motifs ont été validés comme des patrons d'interaction, un tiers n'a pas été retenu. et le reste des motifs exprimaient des relations caractérisant les interactions. Nous détaillons ci-dessous les trois groupes de motifs.

Les motifs écartés concernent des motifs basés sur des noms ou verbes sémantiquement peu significatifs (*AGENE@np AGENE@np)@) be@vb .@sent / AGENE@np AGENE@np (@(Figure@np)@) .@sent*) ou porteurs de simple occurrence (*AGENE@np AGENE@np appear@vvz to@to .@sent*. Parmi les motifs d'interactions validés, certains représentent explicitement des interactions comme les motifs *AGENE@np bind@vvz to@to AGENE@np .@sent / AGENE@np deplete@vvn AGENE@np .@sent / activation@nn of@in AGENE@np by@in AGENE@np .@sent* qui décrivent des interactions bien connues (liaison, inhibition, activation). D'autres motifs modélisent des interactions entre gènes de façon plus générale, signifiant simplement qu'un gène joue un rôle dans l'activité d'un autre gène comme les motifs *AGENE@np involve@vvn in@in AGENE@np .@sent / AGENE@np play@vvz role@nn in@in the@dt AGENE@np .@sent / AGENE@np play@vvz role@nn in@in of@in AGENE@np .@sent*. Enfin, les motifs porteurs de sens pour caractériser les interactions sont de deux types. Un premier groupe a été qualifié de patrons de modalité associant un degré de certitude sur l'interaction exprimée en tant que connaissance scientifique (par exemple le motif : *demonstrate@vvn that@in/that AGENE@np AGENE@np .@sent*). Ces patrons peuvent être regroupés en quatre types : *Negation, Assumption, Observation, Demonstration, Related Word*. Le second groupe porte sur le contexte biologique (la maladie ou l'organisme impliqué dans l'interaction, ou la situation biologique) : *GENE@np with@in AGENE@np in@in Vitro@np*.

Les motifs séquentiels obtenus forment des patrons linguistiques prêts à être appliqués dans des textes biomédicaux pour détecter des interactions entre gènes. Rappelons que pour être appliqués, ces patrons ne s'appuient sur aucune analyse syntaxique de la phrase. Il suffit de chercher à instancier chaque élément du patron dans la phrase.

Pour tester la qualité des patrons, nous avons extrait 440 000 résumés issus de la base PubMed dont les noms de gène ont été étiquetés par notre équipe avec une méthode par dictionnaire associée à des règles heuristiques [Manguin *et al.* 2009]. Nous avons pris

6. L'extracteur utilisé est *dmt*4 développé par C. Rigotti [Nanni et Rigotti 2007].
7. Des tests ont montré qu'augmenter ce seuil éliminait des motifs jugés pertinents pour l'application.

aléatoirement 200 phrases et testé si les patrons linguistiques (c'est-à-dire les 232 motifs précédemment validés) s'appliquaient. Pour chaque phrase contenant une interaction, nous avons mesuré manuellement les performances des patrons linguistiques pour détecter les interactions et ses caractérisations [8]. L'estimation du F-score est autour de 0,79 pour les interactions et 0,77 pour les caractérisations (avec une précision au-dessus de 0,8 et un rappel autour de 0,7 dans les deux cas) ; ce qui est comparable aux approches de la littérature [9].

4.2.2 Discussion

Bien que l'outil d'analyse morphologique des mots (*TreeTagger*) présente des résultats satisfaisants, il subsiste encore un nombre non négligeable d'erreurs d'étiquetage concernant la lemmatisation ou l'attribution d'une catégorie grammaticale. Notre méthode est assez robuste face à ce phénomène. Ainsi, si une erreur d'étiquetage est suffisamment fréquente, elle sera présente dans un motif extrait. Par exemple, *TreeTagger* ne lemmatise pas le mot *cotransfected* mais des motifs extraits contiennent la forme *cotransfected@vvn* et le motif ainsi « erroné » s'appliquera sur un corpus de test puisqu'il contiendra la même erreur.

Notons que la portée des patrons linguistiques se limite au cadre de la phrase. Une telle portée peut introduire des ambiguïtés et des erreurs dans la détection d'interactions lorsque plus de deux gènes apparaissent dans la phrase. Plusieurs cas sont possibles. Soit plusieurs interactions binaires sont présentes dans la phrase, soit l'interaction est de type n-aire ($n \geq 3$) ou encore on peut trouver une interaction en présence d'une simple énumération de gènes. Le cas des interactions n-aires peut être résolu avec un apprentissage sur un jeux de données contenant des interactions n-aires. Les deux autres cas peuvent être traités en introduisant des règles limitant la portée des patrons, par exemple à l'aide de connecteurs (*but*, *however*, etc.). Les cas de silence sont relatifs à des noms ou verbes d'interactions insuffisamment fréquents pour apparaître dans les motifs : par exemple, le nom *modulation* n'a pas été appris dans les motifs. En revanche, le verbe *modulate* est bien présent dans les motifs. On peut penser que l'utilisation de ressources linguistiques légères, de type lexique ou dictionnaire, devrait permettre d'enrichir les motifs, manuellement ou semi-automatiquement (comme par exemple, les listes de verbes et noms d'interaction de [Hao *et al.* 2005] et [Temkin et Gilder 2003]).

La méthode a été expérimentée sur le problème spécifique des interactions entre gènes, mais est conçue pour le problème général de la détection de relation entre entités nommées, même distantes. L'absence d'analyse syntaxique dans ce travail ne signifie pas que la fouille de motifs permet de s'en affranchir complètement dans les applications du TAL. En effet, la tâche présentée est relativement spécifique sur un corpus très volumineux mais spécialisé. La découverte de motifs permet dans ce cadre de révéler la présence de régularités relatives à un phénomène linguistique (relation sémantique spécifique entre entités nommées par exemple) qui évite une analyse syntaxique fine. De plus, le caractère séquentiel des motifs

8. Il a été préalablement procédé à un étiquetage des mots de la phrase par l'outil *TreeTagger* pour pouvoir appliquer correctement les patrons linguistiques.

9. Cependant, l'estimation du F-score pour les interactions ne tient pas compte des erreurs qu'il peut y avoir sur la reconnaissance des noms de gène. Ainsi par exemple si dans une phrase un gène n'a pas été étiqueté en tant que tel, et si ce gène fait partie d'une interaction, la non reconnaissance de l'interaction n'a pas été comptée comme un faux négatif.

Séquence
$\langle(homme\ NOM)(de\ PRP)(culture\ NOM)\rangle$
$\langle(homme\ NOM)(de\ de\ PRP)(culture\ NOM)(modeste\ ADJ)\rangle$
$\langle(en\ PRP)(vieux\ ADJ)(farceur\ NOM)(misanthrope\ ADJ)\rangle$
$\langle(réputé\ réputer\ VER\ pper)(pour\ PRP)(sa\ son\ DET\ POS)(cruauté\ NOM)\rangle$

TABLE 4.2 – Extrait d'une base de séquences pour les expressions : *homme de culture, homme de culture modeste, en vieux farceur misanthrope, réputé pour sa cruauté.*

allié à l'utilisation d'informations morphologiques permet de repérer le sens de l'interaction entre gènes (*e.g.*, prépositions et voix active/passive des verbes) et supplée dans ce cas l'absence d'analyse syntaxique.

4.3 Exploration de motifs pour l'apprentissage de patrons

Dans la section 4.2, l'apprentissage des patrons est fondée sur des motifs séquentiels dont les éléments sont de simples items. D'un point de vue TAL, cela signifie qu'un mot du patron n'est représenté que par une seule information (son lemme, ou bien sa catégorie grammaticale, ou bien la conjonction des deux sous forme d'un seul item). Il peut être intéressant de pouvoir associer un ensemble d'informations à un mot afin d'apprendre des patrons mêlant par exemple la forme du mot, son lemme et / ou sa catégorie grammaticale $\langle(homme\ NOM)(de)(NOM)\rangle$ [10]. Pour cela, nous avons proposé une méthode fondée sur les motifs séquentiels d'itemsets [Cellier et Charnois 2010]. À la différence de la méthode précédente, les séquences sont composées d'ensemble d'items et non plus de simples items. L'autre originalité de ce travail consiste à mettre en place une solution pour gérer le grand nombre de motifs extraits. L'idée est d'exploiter un ordre partiel entre les motifs pour permettre à l'utilisateur d'explorer les motifs au sein d'une hiérarchie et de les valider aisément. Cette exploration n'est pas liée au domaine linguistique et nous avons montré qu'elle peut se généraliser à tout type de motifs extraits par la fouille de données (motifs séquentiels ou règles d'association) dans le cadre de l'analyse de concepts logique, une variante de l'analyse formelle de concepts [Cellier *et al.* 2011].

4.3.1 Motifs séquentiels d'itemsets

Les notions sur les séquences d'items sont ici généralisées à des séquences d'ensembles :
- une séquence est décrite par une liste ordonnée d'ensembles d'items (ensembles communément appelés *itemsets*) ;
- un motif est simplement une séquence d'itemsets. Dans le tableau 4.2, on a ainsi 4 séquences. La dernière comporte 4 itemsets dont le premier est composé de 4 items (la forme *réputé*, le lemme *réputer*, les informations morpho-syntaxiques *VER* et *pper* pour verbe et participe passé).

10. Ce patron est à lire comme étant composé du lemme *homme* de catégorie NOM, suivi du lemme *de*, et enfin d'un NOM.

- une séquence, $S_1 = \langle a_1 ... a_n \rangle$ est contenue dans une séquence $S_2 = \langle b_1 ... b_m \rangle$ s'il existe des entiers $i_1 < i_2 < ... < i_n$ tels que $a_1 \subseteq b_{i_1}, a_2 \subseteq b_{i_2}, ..., a_n \subseteq b_{i_n}$. S_1 est alors appelée *sous-séquence* de S_2, noté $S_1 \preceq S_2$. On a donc par exemple : $\langle (PRP)(NOM) \rangle \preceq \langle (en\ PRP)\ (vieux\ ADJ)(farceur\ NOM) \rangle$;
- les notions de support et de motif fréquent données sur les séquences d'items s'appliquent de la même manière. Le motif $\langle (homme\ NOM)(de\ PRP)\ (culture\ NOM) \rangle$ a ainsi un support de 2 dans la base du tableau 4.2, le motif $\langle (PRP)(NOM) \rangle$ un support de 4 dans ce tableau.

4.3.2 Apprentissage et exploration de patrons

Nous avons appliqué notre méthode à l'apprentissage de patrons linguistiques dénotant des expressions porteuses de qualification ou de jugement et en position détachée comme décrites en section 3.3.2 page 31.

Représentation condensée Pour le calcul des motifs d'itemsets séquentiels fréquents, deux contraintes relatives aux types de patrons visés – décrivant des expressions en position détachée – ont été fixées : i) les motifs débutent l'expression (*i.e.* la séquence) ; ii) ils sont formés d'éléments contigus ($maxgap = 0$). Afin de réduire le nombre de motifs extraits sans perte d'information, on utilise une représentation condensée [11] des motifs fréquents qui élimine les redondances entre motifs : les *motifs fermés*, et un ordre partiel qui permet d'avoir une énumération des motifs structurée afin de faciliter leur exploration.

Un motif fréquent, S, est un motif fermé fréquent, s'il n'existe pas un motif fréquent S' tel que $S \preceq S'$ et $support(S) = support(S')$. Par exemple, soient les motifs :
s_1=$\langle (homme\ NOM)(de\ de\ PRP)(culture\ NOM) \rangle$
s_2=$\langle (NOM)(de\ PRP)(NOM) \rangle$
s_3=$\langle (homme\ NOM)(PRP)(culture\ culture\ NOM) \rangle$
s_4=$\langle (homme\ NOM)(de\ PRP)\ (culture\ NOM)(modeste\ ADJ) \rangle$

Par rapport à notre base du tableau 4.2, le support de s_1, s_2 et s_3 est 2, celui de s_4 1, de plus : $s_2 \preceq s_1$, $s_3 \preceq s_1$ et $s_1 \preceq s_4$. Les motifs s_1 et s_4 sont donc deux motifs fermés pour cette base.

Exploration et visualisation des motifs Les motifs fermés fréquents extraits sont partiellement ordonnés entre eux par la relation de spécialisation $\preceq$ définie *supra*. Afin de mieux visualiser cette relation d'ordre on peut afficher les motifs dans un *diagramme de Hasse* qui est une représentation graphique d'un ordre partiel [Davey et Priestley 1990]. Un exemple est présenté à la figure 4.2.

Notons que la taille du diagramme de Hasse peut être trop grande pour que celui-ci soit affiché entièrement. Nous avons proposé d'utiliser un outil de navigation permettant de naviguer dans le diagramme des motifs les plus généraux aux plus spécifiques afin de les

11. À partir d'un ensemble de motifs extraits et formant la représentation condensée, il est possible de regénérer la collection complète des motifs avec les valeurs exactes des mesures associées, ici la fréquence.

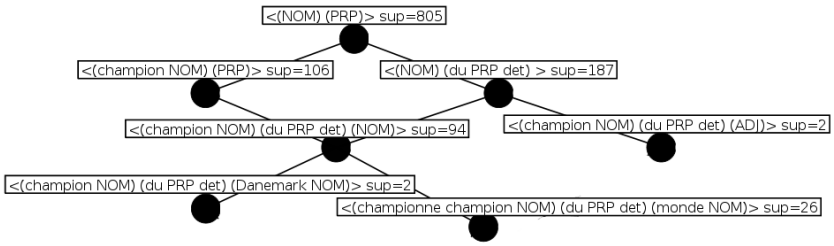


FIGURE 4.2 – Extrait du diagramme de Hasse contenant des motifs séquentiels.

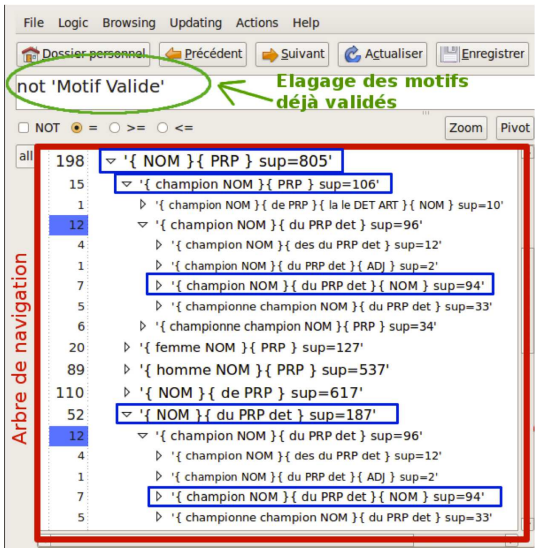


FIGURE 4.3 – Navigation dans les patrons linguistiques pour les valider.

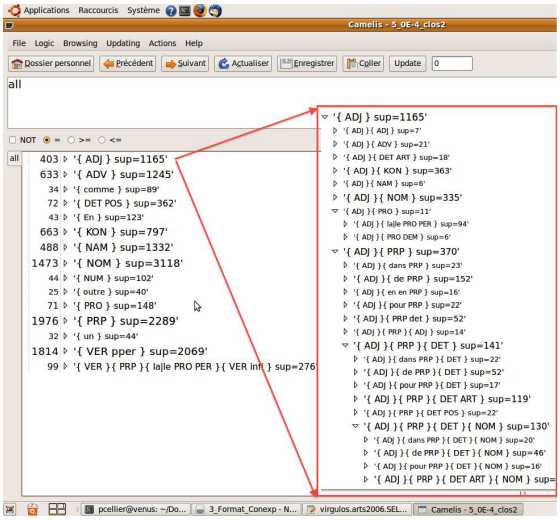


FIGURE 4.4 – Hiérarchie des motifs découverts sur le corpus ARTS : à gauche le haut de la hiérarchie ; à droite les motifs plus spécifiques de *ADJ*.

valider en tant que patrons linguistiques. La figure 4.3 donne un exemple de navigation avec l'outil Camelis [12] [Ferré 2009]. La navigation se fait via « l'arbre de navigation » se trouvant à gauche de l'outil. Remarquons que lorsqu'un motif, M, est sélectionné par l'expert comme patron, tous les motifs dont M est une sous-séquence ne sont plus à examiner. En effet, lors de l'application des patrons, les morceaux de textes qu'ils reconnaissent sont inclus dans les morceaux de texte reconnus par M. Camelis offre la possibilité de marquer les motifs validés comme patrons linguistiques et ainsi de ne plus afficher ceux-ci ni les motifs qui les contiennent (cf « not 'Motif Valide' » dans la figure 4.3), réduisant l'espace des motifs à vérifier et facilitant l'exploration.

Corpus d'apprentissage En l'absence de corpus disponible il est nécessaire de constituer des données d'apprentissage. Deux corpus ont été constitués. Le corpus AXIOLO est formé de 4063 expressions, constituants détachés axiologiques, issues des expériences de [Jackiewicz *et al.* 2009a;b] et générées par des patrons élaborées manuellement. Le corpus ARTS a été généré à partir de règles heuristiques sur les articles de la rubrique « Art » du journal Le Monde, année 2006, soit 3 539 articles. Ces heuristiques sont destinées à filtrer parmi les constituants périphériques ceux qui ne sont a priori pas porteurs de qualification (exemples : proposition de la phrase contenant un verbe conjugué, groupe circonstanciel de temps, espace, but, causalité). Ce corpus est constitué de 13 576 expressions (i.e. séquences), ce qui représente 85 153 mots. Il contient des exemples négatifs. Nous estimons à environ 32% le pourcentage d'expressions non qualificatives (bruit).

Le pré-traitement des corpus consiste en une mise sous forme de séquences comme illustré dans le tableau 4.2 dans lesquelles chaque mot est remplacé par un itemset composé de la forme du mot, son lemme et sa catégorie grammaticale issus de l'outil *TreeTagger*.

Expérimentation Pour le calcul des motifs fermés d'itemsets, nous avons utilisé l'implémentation de Clospan [Yan *et al.* 2003] proposée dans Illimine [13]. Pour chacun des deux corpus nous avons calculé les motifs séquentiels en faisant varier les valeurs du seuil de support, *minsup*, de 50% à $0,05\%$. On constate que des seuils de support élevés (50% ou 25%) ne fournissent que des motifs très généraux où seules les catégories grammaticales apparaissent. Il est donc plus pertinent de choisir des seuils de support très bas pour obtenir des motifs spécifiques et capturer des expressions, ou phénomènes, linguistiques peu fréquents et potentiellement intéressants.

Résultats quantitatifs Pour le corpus AXIOLO, 8264 motifs fermés fréquents sont extraits. Après application des deux contraintes il reste 1789 motifs. Pour le corpus ARTS, environ 8 millions de motifs fermés fréquents sont extraits. Après application des contraintes il reste 7818 motifs.

Résultats qualitatifs et discussion L'analyse des motifs séquentiels d'itemsets extraits du corpus AXIOLO montre la complétude de la méthode. On retrouve en effet les patrons qui avaient été utilisés pour générer ce corpus.

12. http ://www.irisa.fr/LIS/ferre/camelis/index.html
13. http ://illimine.cs.uiuc.edu/

Les expériences réalisées sur le corpus ARTS ont permis de tester la méthode à une échelle relativement importante sur un corpus généré automatiquement et non annoté. L'ensemble des motifs produits comporte des motifs inintéressants dus au bruit présent dans le corpus. Face à ce problème, la navigation au sein de la hiérarchie des motifs est un point fort de la méthode pour sélectionner aisément les patrons linguistiques. Nous montrons figure 4.4 le haut de la hiérarchie des motifs extraits sur ce corpus. Une première analyse met en lumière de nouveaux patrons au regard de ceux conçus manuellement dans [Jackiewicz *et al.* 2009a]. C'est le cas du motif $\langle(ADJ\ pour\ DET\ NOM)\rangle$ (*célèbre pour son monastère, "baroque pour une histoire d'amour ...*) et de ses variantes ou extensions : $\langle(ADV)(ADJ)(pour)\rangle$ (*très célèbre pour ...*), $\langle(ADJ)(pour)(VER)\rangle$ (*indispensable pour assurer ...*).

Notons que l'outil utilisé pour lemmatiser, *TreeTagger*, associe parfois un mauvais lemme ou une mauvaise catégorie grammaticale à un mot. Lorsque l'erreur est suffisamment fréquente, un motif « erroné » va être extrait, comme $\langle(NAM)(ADJ)(pour)\rangle$ où NAM est l'étiquette de la catégorie nom propre au lieu de l'adverbe [14]. Enfin, soulignons le choix d'un seuil de support suffisamment faible pour obtenir des motifs spécifiques et capturer des expressions, ou phénomènes, linguistiques peu fréquents et pouvant être intéressants.

4.4 Les motifs LSR pour la reconnaissance d'entités nommées

Nous avons vu en section 3.2 que la reconnaissance d'entités nommées était un champ de recherche actif [15]. Dans cette section, nous présentons une méthode fondée sur un nouveau type de motifs intégrant une séquence et son contexte mais relâchant la relation d'ordre entre les mots du contexte sur cette problématique [Plantevit *et al.* 2009].

4.4.1 Aperçu de la méthode

L'utilisation de l'extraction de séquences fréquentes, comme présentée dans les deux sections précédentes, semble être une solution naturelle pour la reconnaissance d'entités nommées [16]. Les motifs séquentiels peuvent être utilisés de deux façons pour réaliser cet objectif :

- les motifs séquentiels fréquents qui contiennent au moins une entité nommée ($AGENE$) peuvent être recherchés. Ensuite, ces motifs peuvent être appliqués dans de nouvelles phrases afin de découvrir des noms de gènes. Par exemple, le motif $\langle w_1, w_2, AGENE, w_3, w_4\rangle$ peut être utilisé dans des textes. Si $\langle w_1, w_2\rangle$ et $\langle w_3, w_4\rangle$ peuvent s'apparier dans une phrase, alors le segment de phrase entre w_1, w_2 et w_3, w_4 est reconnu comme étant un nom de gène ;
- les règles séquentielles, construites à partir des motifs séquentiels fréquents, ont l'avantage d'être associées à une mesure d'intérêt, indiquant la fiabilité de la règle. Ainsi, les règles séquentielles doivent satisfaire à la fois une contrainte de fréquence

14. Erreur fréquente de notre version de l'outil *TreeTagger*. pour les mots en début de phrase commençant par une majuscule

15. Voir cette section 3.2 et l'introduction du présent chapitre pour un état de l'art et positionnement de nos travaux.

16. Remarquons les travaux en cours de [Nouvel *et al.* 2010] sur l'utilisation des motifs séquentiels pour enrichir un système de reconnaissance d'entités nommées à base de transducteurs.

et une contrainte de confiance. La confiance d'une règle $X \rightarrow Y$ correspond à la probabilité $P(Y|X)$ qu'une phrase qui contient la séquence X contienne également la séquence Y, apparaissant après X. Dans le cadre de la découverte de noms de gènes, il est intéressant de découvrir des règles composées d'une séquence de mots concluant sur un nom de gène, comme la règle *the overexpression of* $\rightarrow$ *AGENE*. De telles règles permettent d'identifier le contexte gauche d'un nom de gène. En inversant la relation d'ordre, d'autres règles séquentielles peuvent être découvertes, identifiant le contexte droit des noms de gènes. Ainsi une paire de règles peut être appliquée pour détecter des noms de gènes. Par exemple, les règles $R_l = \langle w_1, w_2, w_3 \rangle \rightarrow AGENE$ et $R_r = AGENE \leftarrow \langle w'_1, w'_2, w'_3 \rangle$ peuvent s'appliquer dans la phrase $\ldots w_1\ w_2\ w_3\ X\ Y\ Z\ w'_3\ w'_2\ w'_1 \ldots$ où $X\ Y\ Z$ est alors identifié comme étant un nom de gène.

Toutefois, des expérimentations montrent les limites des deux approches présentées ci-dessus [Plantevit *et al.* 2009]. En effet, les motifs séquentiels offrent une excellente couverture des phrases qui contiennent des noms de gènes (fort rappel) mais leur utilisation pour la détection de noms de gènes engendre la reconnaissance d'un nombre trop important de faux positifs (faible précision). Ceci est dû à l'absence de mesure de confiance associée aux motifs. Les règles séquentielles offrent, quant à elles, de bons résultats, notamment grâce à l'utilisation de la mesure de confiance, en ce qui concerne la précision, mais leur faible rappel constitue une limite non négligeable.

Idée clé L'idée clé de notre approche consiste à combiner le bon rappel des motifs séquentiels avec la bonne précision des règles séquentielles. Autrement dit, nous visons à améliorer sensiblement la précision des motifs séquentiels sans altérer leur fort rappel. Pour cela, nous proposons un nouveau type de motif, appelé motif LSR, qui permet de considérer le voisinage des motifs séquentiels fréquents – sans imposer de condition sur l'ordre des mots dans ce voisinage – afin de contextualiser les motifs séquentiels dans les séquences de données. Ce voisinage conduit à améliorer l'utilisation du motif séquentiel associé lors de la détection d'entités nommées. Notre approche a donc l'originalité de définir une méthode de fouille « hybride » portant à la fois sur des données séquentielles et ensemblistes.

4.4.2 Motifs LSR : un nouveau type de motif

Les motifs LSR permettent de caractériser une séquence à l'aide d'itemsets modélisant le voisinage de la séquence. Leur caractéristique majeure est de combiner l'information séquentielle portée par une séquence avec son voisinage gauche et droit où la relation d'ordre est relâchée en étant modélisée sous forme d'itemsets. Les motifs LSR, plus qu'une simple combinaison entre une séquence et deux itemsets, permettent de contextualiser une séquence grâce à son voisinage.

Définition 1 (LSR)
Un motif LSR x *est un triplet* $x = (l, s, r)$ *où :*
- s *est un motif séquentiel ;*
- l *et* r *sont des itemsets.*

Les itemsets l et r ont pour but de modéliser le voisinage d'une séquence s. Par exemple, un motif LSR est $x_1 = (\{\}, \langle$"the","$AGENE$"$,\rangle\{$"$gene$","$with$","$associated$"$\})$ où $l = \{\}$ et $r = \{$"$gene$","$with$","$associated$"$\}$, ce qui signifie que ces mots apparaissent dans le voisinage droit de la séquence "$the\ AGENE$".

La contrainte induite par la relation d'ordre est relâchée autour des motifs séquentiels fréquents dans les séquences de données afin d'extraire les itemsets fréquents qui modélisent le voisinage des motifs séquentiels et les contextualisent dans les séquences de données. Avant de montrer l'utilisation des motifs LSR fréquents, nous devons introduire les définitions suivantes.

Contrairement aux itemsets qui apparaissent au plus une fois dans une transaction, une séquence peut apparaître plusieurs fois dans une séquence de données. Par exemple, pour la séquence de données [17] $S = \langle(1, The), (2, results), (3, reported), (4, here), (5, support), (6, the), (7, evidence), (8, of), (9, the), (10, mutational), (11, heterogeneity), (12, of), (13, the), (14, IDS), (15, gene)....\rangle$ il existe 3 occurrences de la séquence $s = \langle the, gene\rangle$:

1. $\langle(1, the), (15, gene)\rangle$;
2. $\langle(6, the), (15, gene)\rangle$;
3. $\langle(13, the), (15, gene)\rangle$.

Il existe donc plusieurs façons d'identifier le voisinage d'une séquence au sein d'une séquence de données. Dans le précédent exemple, on voit que seul le dernier motif est intéressant pour l'identification de l'entité nommée et extraire les itemsets les *plus représentatifs*. De façon générale, l'idée est de rassembler prioritairement ce qui est proche. Dans ce but, nous introduisons la notion d'*occurrence compacte* d'une séquence s dans une séquence de données S.

Définition 2 (occurrence compacte)
Étant donnés un motif séquentiel $s = \langle i_1, i_2, \ldots, i_k\rangle$, un ensemble de contraintes $\mathcal{C}$ et une séquence de données S, une occurrence o_c de s dans S où $o_c = \langle(t_1, i_1), (t_2, i_2), \ldots, (t_k, i_k)\rangle$ est une occurrence compacte de s dans S si les conditions suivantes sont respectées :

- *o_c satisfait $\mathcal{C}$;*
- *il n'existe pas d'occurrence $o' = \langle(t'_1, i_1), (t'_2, i_2), \ldots, (t'_k, i_k)\rangle$ de s dans S telle que $o' \neq o_c$ et o' satisfait C et $t_1 \leq t'_1$ et $\forall\alpha \in \{2, \ldots, k\}, t'_\alpha \leq t_\alpha$.*

Cette définition permet de se concentrer sur les plus petits segments de la séquence de données S qui contiennent la séquence s. En effet, une séquence de données S peut contenir plusieurs occurrences compactes d'une séquence s. Par exemple, étant donné $\mathcal{C} = \emptyset$, la séquence de données
$S = \langle(1, a), (2, c), (3, b), (4, d), (5, a), (7, a), (8, b), (10, e), (12, f), (14, a), (15, g), (16, h), (18, b), (20, c)\rangle$
contient trois occurrences compactes de $s = \langle a, b\rangle$:

1. $\langle(1, a), (3, b)\rangle$;
2. $\langle(7, a), (8, b)\rangle$;

17. Correspondant à la phrase *The results reported here support the evidence of the mutational heterogeneity of the IDS gene...*

3. $\langle(14, a), (18, b)\rangle$.

Notons que la séquence $\langle(1, a), (18, b)\rangle$ n'est pas une occurrence compacte de s dans S puisqu'elle n'est pas minimale.

Si nous ajoutons la contrainte de temps « maximal gap » $max_gap = 2$ à $\mathcal{C}$, ce qui signifie que l'écart temporel entre deux items consécutifs dans la séquence s est 2, alors S contient deux occurrences compactes de s : $\langle(1, a), (3, b)\rangle$ et $\langle(7, a), (8, b)\rangle$.

Extraction des motifs LSR Je donne les principes généraux du processus d'extraction des motifs LSR qui est détaillé dans [Plantevit *et al.* 2009]. L'extraction se réalise en deux étapes. La première extrait l'ensemble des motifs séquentiels vérifiant un ensemble de contraintes (support, maxgap, *etc.*). La deuxième consiste à fouiller le voisinage des séquences fréquentes pour extraire leurs contextes gauche et droit. Ces contextes sont formés des items fréquents situés dans un voisinage N_R de l'occurrence compacte de la séquence fréquente considérée (N_R est un nombre entier correspondant à la différence absolue entre l'estampille temporelle de l'item considéré et l'estampille de l'élément de l'occurrence compacte le plus proche, ou si l'on considère des mots consécutifs, N_R s'exprime en nombre de mots).

4.4.3 Utilisation des motifs LSR pour la détection d'entités nommées

L'extraction des motifs LSR est effectué sur un corpus lemmatisé. Ces motifs possèdent un caractère générique et peuvent être utilisés pour de multiples tâches, mais il est simple d'y ajouter des contraintes pour les spécialiser sur une tâche donnée. Ainsi, pour la détection d'entités nommées, les séquences doivent contenir une entité nommée, par exemple l'item $AGENE$ dans le cas de la reconnaissance de noms de gènes. De plus, des contraintes de temps doivent être imposées pour considérer seulement les événements consécutifs ($maxgap = 0$).

Nous proposons d'associer une mesure de confiance au motif séquentiel s de chaque motif LSR. Celle-ci correspond au rapport entre la fréquence d'un motif séquentiel s contenant n'importe quelle instance d'une entité nommée et la fréquence de ce même motif séquentiel s'appliquant dans le texte indifféremment de la présence ou non d'une entité nommée ; autrement dit, est calculé le rapport entre le nombre de fois où le motif s reconnaît l'entité nommée par rapport au nombre de fois où il s'applique dans le texte. L'objectif de cette mesure est de déterminer si le motif séquentiel s peut être appliqué seul ou s'il est nécessaire de prendre en compte son voisinage (itemsets l et r) afin de l'utiliser de façon fiable lors de la reconnaissance d'entités nommées. La confiance d'un motif séquentiel s pour une entité nommée E est donc égale au support de s divisé par le support de la séquence s (notée $s_{[E/*]}$) dans laquelle les items correspondant à une entité nommée (*e.g.*, $AGENE$) sont remplacés par une valeur joker $*$. La définition du support s'applique à de telles séquences, où la valeur $*$ s'apparie avec n'importe quel groupe de mots.

Définition 3 (confiance)

Étant donnée une entité nommée E, la confiance d'un motif séquentiel s contenant E est :

$$Confiance_E(s) = \frac{support(s)}{support(s_{[E/*]})}$$

Cette mesure permet de déterminer si l'occurrence de l'entité E est fortement reliée à la présence des autres items de la séquence. Par exemple, si le support de la séquence $\langle the\ gene\ AGENE\ interacts\ with \rangle$ est similaire à celui de $\langle the\ gene\ *\ interacts\ with \rangle$ (confiance $\simeq 1$), alors quand une phrase contient d'abord « *the gene* » et plus loin « *interacts with* », il y a très certainement un nom de gène entre eux.

Détection des entités nommées

À partir des motifs LSR et des confiances calculés précédemment, il devient possible de détecter des entités nommées dans un texte. Étant donnée une nouvelle phrase écrite en langue naturelle, intialement lemmatisée, la première étape consiste à trouver les motifs LSR qui peuvent s'appliquer dans la phrase. Si une séquence s d'un motif (l, s, r) est susceptible de s'appliquer (tous les items de s, qui caractérisent une entité nommée, sont présents dans la phrase), la confiance de s est alors prise en compte. Si cette confiance est supérieure ou égale à un seuil fixé, alors la séquence s est considérée comme étant suffisamment en rapport avec la présence d'une entité nommée et peut directement être appliquée pour identifier cette entité au sein de la phrase. Si tel n'est pas le cas, s ne possède pas une confiance suffisante pour être appliquée seule, et c'est alors le voisinage de s dans la phrase qui est examiné pour vérifier si l'application de s est pertinente. Si un nombre suffisant d'items, par rapport à un seuil W_{min}, de l et r sont présents dans les contextes gauche et droit alors l'utilisation de s pour l'identification de l'entité nommée dans la phrase considérée est jugée pertinente. Enfin, notons qu'un motif séquentiel s peut s'appliquer plusieurs fois dans une phrase. Il est donc nécessaire de considérer toutes les occurrences compactes de s dans la phrase.

Pour ce traitement, l'algorithme 1 est utilisé afin de déterminer où un motif (l, s, r) peut être appliqué dans une phrase S. Pour cela, il considère toutes les occurrences compactes o_c de $s_{[E/*]}$ (recherche de s dans laquelle l'entité nommée E est remplacée par le joker $*$) dans la phrase S. Pour chacune de ces occurrences, si la confiance de s est suffisante (c'est-à-dire $\geq minconf$) alors la partie appariée avec $*$ est considérée comme correspondant à l'entité nommée E. Sinon, le second critère utilisant l et r est appliqué. Pour cela, l'algorithme détermine le nombre d'éléments du contexte gauche (préfixe composé des N_R éléments précédents o_c dans S) qui sont aussi dans l. Le même calcul est effectué pour le contexte droit (suffixe de N_R éléments suivants o_c dans S) par rapport à r. Si le nombre total d'éléments de l ou r, présents respectivement dans le contexte gauche et droit, est supérieur ou égal à W_{min}, alors dans ce cas aussi la partie appariée avec $*$ est étiquetée comme entité nommé E. L'algorithme termine après avoir considéré toutes les occurrences compactes de $s_{[E/*]}$ dans S.

Expérimentation et résultats

Des expérimentations menées sur le corpus issu du challenge *BioCreative* [Yeh *et al.* 2005], (*cf.* figure 4.5) ont pour objectif de montrer l'intérêt des motifs LSR pour la décou-

Algorithme 1 : Utilisation d'un motif LSR pour la détection d'entités nommées.

Data : Phrase S, motif LSR $x = (l, s, r)$, seuil de confiance $minconf$, rayon de voisinage N_R, nombre minimal de mots W_{min}, entité nommée E

begin

  for *chaque occurrence compacte* o_c *de* $s_{[E/*]}$ *dans* S **do**

    if *Confidence(s)*$\geq minconf$ **then**

      Étiqueter avec E la partie de o_c correspondant à $*$ dans $s_{[E/*]}$;

    else

      if $|prefix(o_c, S, N_R) \cap l| + |suffix(o_c, S, N_R) \cap r| \geq W_{min}$ **then**

        Étiqueter avec E la partie de o_c correspondant à $*$ dans $s_{[E/*]}$;

      else

        Ne pas appliquer s ;

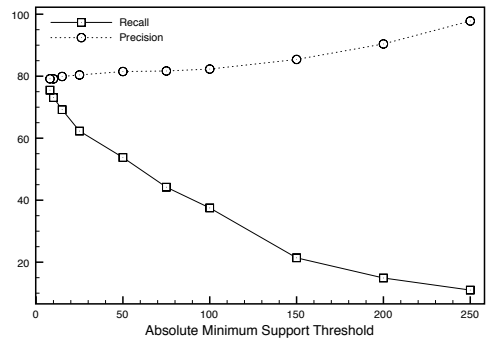
end

verte de noms de gènes dans des textes biomédicaux où ils représentent un bon compromis entre le fort rappel des motifs séquentiels et la forte confiance des règles séquentielles. Le corpus a été lemmatisé et chaque phrase représente une séquence de données. La base de séquences SDB regroupe alors l'ensemble des phrases du corpus. Nous avons utilisé une validation croisée (partition en 10 sous-ensembles) pour estimer la fiabilité de notre approche. Les temps d'exécution sont par ailleurs négligeables : environ 1 000 phrases sont traitées par seconde pour la détection des entités nommées (application de l'algorithme 1).

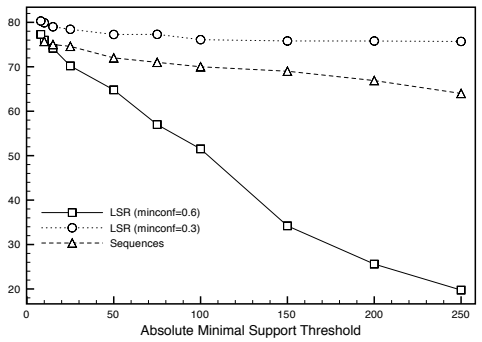
Les motifs LSR permettent d'exploiter pleinement les motifs séquentiels qui ne présentent pas une garantie de confiance suffisante pour cette tâche. Par exemple, la séquence $\langle$*that AGENE*$\rangle$ a une très mauvaise confiance, mais certains mots apparaissent fréquemment dans son voisinage gauche (*indicated, revealed, demonstrate, evidence*) et droit (*binds, expressed, activity, protein,* etc.) et il peut être utilisé combiné avec ces mots. Ainsi de tels motifs séquentiels qui seraient inutilisables sinon, peuvent-ils être appliqués grâce à la prise en compte de leur voisinage.

Le but des ces expérimentations est d'évaluer la qualité de l'application des motifs LSR pour la reconnaissance des entités nommées biomédicales. Nous étudions également le comportement des motifs LSR en fonction du seuil de support, du seuil de confiance et de W_{min}. La valeur du rayon de voisinage N_R a été fixée à 5 : l'idée étant que les mots du contexte qui sont linguistiquement intéressants sont ceux situés dans le voisinage proche de l'entité nommée, c'est-à-dire soit le syntagme dans lequel figure l'entité nommée, soit un syntagme contigu, (par exemple, lorsque l'entité est l'argument d'un prédicat verbal ou nominal comme on peut le voir dans les exemples de motifs LSR donnés dans la figure 4.6). Les expérimentations menées avec des valeurs inférieures ou supérieures à 5 confortent le choix de cette valeur.

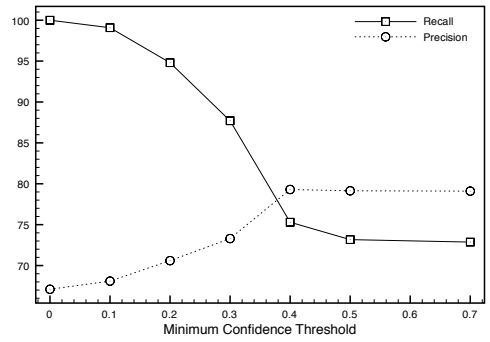
Au final, les expérimentations montrent que la prise en compte du voisinage fournit des résultats intéressants qui s'approchent des résultats donnés par les meilleures ap-



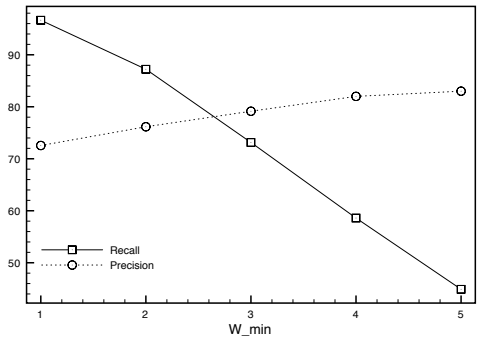
Précision et rappel des motifs LSR en fonction du seuil de support *supmin* ($W_{min} = 3$, $N_R = 5$). *Lorsque supmin diminue, un plus grand nombre de motifs* LSR *est extrait, produisant une meilleure couverture (rappel), mais une précision en baisse.*



F-score des motifs LSR et des motifs séquentiels ($W_{min} = 3$, $N_R = 5$). *Les motifs* LSR *offrent de meilleures résultats.*



Précision et rappel des motifs LSR en fonction du seuil de confiance *minconf* (*minsupp* = 10, *vmin* = 3, $N_R = 5$)*Plus le seuil de confiance est faible, plus le nombre de faux positifs augmente.*



Précision et rappel des motifs LSR en fonction de W_{min} ($minconf = 0.6$, $N_R = 5$). *La précision augmente et le rappel diminue lorsque* W_min *(nombre de mots devant apparaître dans le voisinage du motif) devient plus contraignant.*

FIGURE 4.5 – Expérimentations menées sur le corpus *BioCreative*

$\langle\{a, of, member\}, \langle the, AGENE\rangle, \{of, family\}\rangle$
$\langle\{is, a\}, \langle member, of, the, AGENE\rangle, \{family\}\rangle$
$\langle\{mutations\}, \langle the, AGENE, gene\rangle, \{\}\rangle$
$\langle\{interaction\}, \langle with, the, AGENE\rangle, \{\}\rangle$
$\langle\{expression\}, \langle of, AGENE\rangle, \{cancer\}\rangle$
$\langle\{the\}, \langle AGENE, gene\rangle, \{with, associated\}\rangle$

FIGURE 4.6 – Exemples de motifs LSR

proches sur les mêmes données. De plus, les motifs LSR sont facilement interprétables. Par exemple, le motif $\langle\{\}, \langle AGENE, expression, in\rangle, \{cells\}\rangle$ signifie que le mot *cells* apparaît fréquemment dans le voisinage droit de la séquence fréquente $\langle AGENE\ expression\ in\rangle$. Les exemples de la figure 4.6 illustrent un autre intérêt de notre approche : l'utilisation des motifs LSR par des linguistes et/ou comme ressource linguistique de par leur caractére

qualitatif.

Comme signalé pour les motifs d'interaction, il serait intéressant d'utiliser des ressources linguistiques pour enrichir les motifs LSR. Notons toutefois qu'un pré-traitement sur le corpus qui éliminerait les mots communs (*stop-lists*) ne nous paraît pas, dans ce cas, pertinent, car certains mots communs dénomment aussi des entités nommées (par exemple *she* est aussi un nom de gène), ou peuvent être utiles pour délimiter les entités.

Chapitre 5

Bilan et perspectives

L'accès à l'information dans les textes constitue le fil directeur des travaux présentés dans ce mémoire. Parce que l'objet d'étude est la langue, cette thématique est abordée sous un angle linguistique en TAL et en fouille. Notre objectif est de concevoir des systèmes performants aussi bien en termes de résultats quantitatifs qu'au niveau de la compréhension des phénomènes en jeu pour permettre un retour qualitatif sur le modèle linguistique.

La thématique de l'accès à l'information a été abordée sous différents angles, de la compréhension automatique à l'accès au contenu ciblé, sur différents phénomènes linguistiques, temps, espace et expression de l'opinion, et pour différentes applications. Sous l'angle de la compréhension globale de textes, nous avons montré l'intérêt d'un formalisme comme les graphes conceptuels pour obtenir une représentation sémantique formelle de l'énoncé permettant des inférences. Pour le traitement de corpus de grande taille, et sur des domaines de spécialité, nous avons mis en place des analyses locales et modulaires, produisant des annotations sémantiques des expressions et permettant une analyse discursive (segmentation thématique, analyse de l'opinion). Ces analyses ont été développées dans le cadre de projets de recherche et pour différentes applications (recherche documentaire, extraction d'information et BioNLP). L'ensemble de ces travaux a pour autre particularité d'être fondé sur des modèles descriptifs issus de la linguistique et/ou élaborés en collaboration avec les linguistes (théorie du temps et de l'aspect, encadrement du discours, opinion).

Pour acquérir automatiquement les ressources linguistiques nécessaires à ces analyses, les approches issues de la fouille de données nous ont semblé particulièrement appropriées de par leur caractère symbolique et non supervisé. L'originalité de nos travaux a consisté à concevoir des méthodes de fouille intégrant des spécificités du traitement de la langue et, réciproquement, à utiliser l'apport de la fouille pour le TAL, tel que l'apprentissage automatique de ressources linguistiques. Nous avons en effet montré l'intérêt d'utiliser les motifs issus des méthodes de fouille de données dans le domaine du TAL appliqué à la biologie médicale et génétique, et plus particulièrement dans les tâches d'extraction d'information. Nous avons proposé une approche pour apprendre les patrons linguistiques par une méthode de fouille de données basée sur les motifs séquentiels et sur une fouille dite récursive des motifs eux-mêmes. Pour la reconnaissance d'entités nommées comme les noms de gènes, nous avons conçu un nouveau type de motifs intégrant le contexte d'une séquence mais relâchant la relation d'ordre entre les mots du contexte. Nous avons aussi présenté une

méthode pour apprendre des patrons linguistiques mêlant différents types d'information et produisant des patrons ordonnés en une hiérarchie (des plus généraux aux plus spécifiques). Sur le plan de la fouille, deux types d'approches ont été conçues pour apporter une solution au problème posé par l'abondance des motifs produits et de la sélection des motifs intéressants. L'une consiste à intégrer les connaissances linguistiques dans le processus de fouille sous formes de contraintes et à utiliser la fouille récursive. La seconde permet d'explorer les motifs et les valider aisément. Qu'il s'agisse des motifs LSR, de la fouille de séquences d'itemsets ou encore de la fouille récursive, les méthodes que nous avons conçues sont adaptées aux textes mais ne sont pas dédiées à ces seuls types de données. Elles sont utilisables dans différents contextes de la fouille de données, comme par exemple les flux de données séquentielles dans lesquels des événements arrivent simultanément.

Nos perspectives de recherche résident bien sûr dans le prolongement de nos travaux sur l'accès à l'information et la fouille de données.

Analyse d'opinion Un premier travail concerne la poursuite du travail entamé sur la thématique de l'expression de l'opinion. La méthode proposée pour acquérir des patrons constituants détachés (section 3.3.2) ne permet pas de distinguer leur caractère évaluatif ou qualificatif. L'objectif est de produire des règles pour opérer cette distinction. Nous pensons que les motifs LSR pourraient être utilisés dans ce sens pour étudier les contextes dans lesquels apparaissent ces constituants. Plus exploratoire, nous proposons d'aborder le problème de la détermination de la cible de l'opinion (dans un premier temps au niveau intra-phrastique) par l'utilisation de motifs séquentiels d'itemsets pour prendre en compte ce type de relation qui peut être distant et pour s'appuyer sur diverses marques préalablement annotées dans les textes. L'idée est de découvrir des combinaisons de ces annotations qui restent à définir (par exemple : entités nommées, caractères évaluatifs de certains termes, groupes constituants de la phrase...).

Fouille pour la linguistique Sur un plan plus linguistique, une collaboration avec Dominique Legallois (linguiste au CRISCO, Caen) a été récemment amorcée [Legallois *et al.* 2011]. Cette collaboration se poursuit cette année avec le recrutement d'un post-doctorant (Solen Quiniou) que nous co-encadrons. Ce travail, inspiré du modèle de [Hoey 1991] présente une méthode de navigation dans les textes, fondée à la fois sur la « réduction textuelle » et sur la répétition lexicale. Issu de l'analyse formelle de concepts, l'outil de navigation proposé calcule et affiche des « réseaux d'appariements phrastiques ». Cette première implémentation du modèle permet ainsi une exploration de corpus de grande taille en réduisant le texte aux phrases informationnellement les plus pertinentes et en faisant apparaître la cohésion interphrastique du texte à travers les appariements entre phrases. L'objectif est de permettre au linguiste de mieux comprendre l'organisation textuelle. L'appariement se fonde actuellement sur la simple répétition morphologique des lexèmes. Une première amélioration va consister à prendre en compte les formes dérivées (*stabiliser/stable*) et à étudier une possible corrélation entre mots fréquents et zones denses des réseaux phrastiques. Une expérimentation va être entreprise sur un corpus thématique diachronique, par exemple dans un ensemble d'articles du Monde consacrés à un même thème sur une période de 20 ans. La navigation intertextuelle peut se révéler intéressante :

elle permettrait par exemple, d'observer des rapports entre textes, les formations idéologiques, la constitution de clichés et des collocations, le développement des argumentations, la rupture des positions, *etc.* (*cf.* [Legallois 2011]). Plus largement, cette collaboration vise l'exploitation des méthodes de fouille par des chercheurs en sciences humaines, et sur des textes à valeurs littéraires, sociologiques ou historiques pour l'analyse stylistique par exemple. Cet objectif nous conduira à adapter les outils de fouille à l'utilisation non experte, afin de les mutualiser pour des disciplines différentes, ce qui nécessite de concevoir des méthodes et outils adéquats et pertinents pour l'exploration textuelle en sciences humaines, ainsi que des représentations de résultats facilement interprétables.

Extraction d'information et capitalisation de connaissances La thèse en cours de Laurie Serrano que je co-encadre a démarré en janvier dernier. Le sujet porte sur l'extraction d'information et la capitalisation de connaissances à partir des informations repérées dans les textes. Le premier axe s'intéresse à l'extraction dans les dépêches de presse des événements de type militaire (comme les attentats), les entités impliquées et les circonstants spatio-temporels. Il s'agit ici de l'événement principal de la dépêche (tel qu'il peut apparaître dans le titre). Différentes méthodes d'extraction (règles linguistiques conçues manuellement, champs aléatoires conditionnels, motifs séquentiels) vont être comparées selon les tâches. L'extraction sera guidée par une ontologie du domaine. Ce travail va en outre s'attacher à évaluer la qualité de l'information extraite par l'attribution d'un indice de confiance aux techniques ou règles à l'origine de l'information. Le second axe vise à définir des méthodes pour transformer les informations en connaissances, à savoir les relier, les structurer, les fusionner *etc.* Ces traitements aboutissent à la construction de fiches, alimentant une base de connaissances, et doivent permettre leur mise à jour lors du traitement de nouveaux documents mais aussi des mécanismes de raisonnement et d'inférence afin d'en déduire de nouvelles connaissances. De plus, un utilisateur devra pouvoir ajouter, modifier ou valider une information, et un mécanisme de rétro-propagation permettra de réévaluer la confiance accordée aux règles afin d'améliorer les performances du système.

Vers une hybridation fouille/TAL Nous avons présenté dans ce mémoire une première tentative d'utilisation des méthodes de fouille en TAL pour l'apprentissage de ressources linguistiques, et symétriquement, la donnée textuelle nous a conduits à revisiter les méthodes de fouille. Au-delà de l'apprentissage de ressources, notre objectif est de tirer parti des deux types d'approche sur la problématique générale de l'extraction de connaissances à partir de textes. Cet axe de travail trouve une concrétisation dans le projet ANR Hybride (programme blanc) dont je suis le coordinateur pour le GREYC et qui démarre le 1er décembre de cette année.

Ce projet de recherche a pour ambition de développer de nouvelles méthodes et outils pour guider la découverte de connaissances à partir de textes en combinant des méthodes de TAL et des méthodes de fouille. Une idée maîtresse est de concevoir un processus convergent qui met en interaction les deux approches, où les méthodes de TAL guident les méthodes de fouille de données et réciproquement, pour l'analyse et la fouille de documents textuels par leur contenu. Par exemple, les méthodes de TAL appliqués au texte identifient des informations textuelles qui sont utilisées comme contraintes pour focaliser les méthodes de fouille sur des motifs intéressants. En parallèle, les méthodes de fouille permettent

d'extraire des motifs qui sont utilisés pour guider l'extraction d'information et l'analyse des textes. Cette combinaison peut se voir comme un cercle vertueux où chacune des familles de méthodes est en interaction avec l'autre pour produire le meilleur sur la base des résultats que l'une fournit à l'autre et réciproquement.

Un système intégrant ces principes et les opérations associées à la boucle TAL/fouille va être mis en place dans la cadre du réseau Orphanet relatif aux maladies orphelines pour l'analyse de textes scientifiques et médicaux et la documentation nouvelle de ces maladies rares. L'application ciblera deux types de connaissances dans les textes : i) l'extraction des relations entre gènes et maladies rares et les contextes modaux de validation relatifs (permettant d'associer un certain degré de certitude) ; ii) l'identification de l'ordre chronologique des symptômes cliniques associés aux maladies rares.

À court terme, un axe de recherche va consister à définir et à élaborer un modèle formel des contraintes linguistiques à intégrer au sein même des algorithmes de fouille de données pour guider la découverte de motifs (avec les outils existants, les contraintes de ce type que nous avons utilisées, et présentées au chapitre 4, ne peuvent être appliquées qu'en phase de post-traitement). Dans le cadre d'un co-encadrement d'un post-doctorant (Nicolas Béchet), ce travail est actuellement amorcé par la réalisation d'un extracteur de fouille de données séquentielles prenant en compte des itemsets et intégrant des contraintes syntaxiques classiques (gap, longueur, *etc.*), mais aussi des contraintes textuelles comme l'appartenance ou non d'un item d'un type particulier dans les motifs extraits, ce type pouvant être spécifié par une information linguistique (exemple : présence de la catégorie *verbe*, ou encore *verbe d'interaction...*). L'objectif applicatif est d'associer des informations linguistiques variées (morphologiques, syntaxiques, sémantiques...) à des unités textuelles de différents niveaux (mot, proposition, phrase...) : les unités textuelles sont par nature ordonnées donc séquentielles, alors que les informations linguistiques sont de type ensembliste (itemsets).

À terme, l'objectif sera de concevoir et développer des méthodes de fouille adaptées à l'analyse discursive. Un des enjeux majeurs est la découverte de motifs considérant plusieurs niveaux de granularité dans un document : la granularité est une notion centrale en analyse du discours, elle permet aussi d'appréhender le texte selon plusieurs points de vue. Nous pensons que le paradigme de la fouille de graphes est particulièrement adapté, et il serait intéressant d'envisager la fouille sur les graphes conceptuels, afin de tirer parti des hiérarchies de types et concepts, et plus généralement de la relation de généralisation, définis sur ce formalisme. À plus long terme, il s'agit de dégager des dimensions structurelles sur les données comme leur granularité, le langage (*e.g.*, itemsets, séquences, graphes) pour produire des méthodes de fouille reposant sur ces dimensions et les combinant.

Annexe : présentation du modèle des graphes conceptuels

Je donne dans cette annexe les principales notions sur les GC – graphes conceptuels – en suivant l'exposé très rigoureux de [Chein et Mugnier 1992] et [Mugnier et Chein 1996] qui proposent une formalisation couvrant le modèle de [Sowa 1984].

Un GC est un graphe fini, qui possède deux types de sommets, les sommets concepts (ou c-concepts) et les sommets relations (ou r-relations) qui représentent les liens entre les concepts. Un sommet d'un type ne peut être lié qu'à un sommet de l'autre type :
$[concept] \rightarrow (relation) \rightarrow [concept]$

On se donne un support pour définir le vocabulaire nécessaire à la définition d'un graphe conceptuel.

Définition 4 *Un support est un quintuplet $S = (T_C, T_R, \sigma, I, \tau)$ tel que :*
- *T_C est l'ensemble des types de concepts. C'est un ensemble fini partiellement ordonné par la relation sorte-de possédant un plus petit élément (type absurde noté $\bot$) et un plus grand élément (type universel noté $\top$),*
- *T_R est l'ensemble des types de relations. C'est un ensemble fini, partiellement ordonné par une relation de subsomption et partitionné en sous-ensembles de types relations de même arité.*
  $T_R = T_{R_1} \cup \cdots \cup T_{R_p}$ où T_{R_i} est l'ensemble des types de relation d'arité i, $i \neq 0$. Tout T_{R_i} admet un plus grand et un plus petit élément et forme un treillis.
- *σ est l'application qui associe à tout type de relation t_r (où il existe j tel que $t_r \in T_{R_j}$) une signature, c'est-à-dire le type maximal (au sens de la relation définie sur T_C) de chacun de ses arguments. Pour $1 \leq i \leq j$, on note $\sigma(i)$ le $i^{ème}$ argument de t_r.*
- *I est l'ensemble des marqueurs individuels. C'est un ensemble dénombrable, non nécessairement fini. Il existe en outre un marqueur dit générique, noté $*$ qui désigne un individu non spécifié. L'ensemble des marqueurs $I \cup \{*\}$ est muni de l'ordre suivant : $*$ est plus grand que tous les marqueurs individuels, et ces derniers sont deux à deux incomparables.*
- *τ est une relation appelée relation de conformité de I dans $T_C \setminus \{\bot\}$ qui à tout marqueur individuel associe un type de concept.*

Définition 5 *Un graphe conceptuel simple $G = (R, C, U, Etiq)$ défini sur un support est un multigraphe*[1]*, non orienté, biparti, non nécessairement connexe, où :*

1. On appelle multigraphe un graphe qui peut avoir plusieurs arêtes entre deux sommets (ici entre

- *R et C sont les deux classes de sommets, appelés sommets concepts et sommets relations.*
- *U est l'ensemble des arêtes ; l'ensemble des arêtes adjacentes à tout sommet relation r est totalement ordonné, ce que l'on représente en numérotant les arêtes (de 1 à degré de r). On note $G_i(r)$ le ième voisin de r dans G et $G(r)$ l'ensemble des voisins de r.*
- *étiq est une application, qui à tout sommet associe une étiquette :*
  - *Si $r \in R$, $étiq(r) \in T_R$, c'est le type du sommet relation. Il existe un ordre partiel défini sur les étiquettes des relations de même arité.*
  - *Si $c \in C$, $étiq(c) \in (Tc \setminus \{\bot\}) \times (I \cup \{*\})$. L'étiquette d'un sommet concept est un couple (type(c), marqueur(c)). Si c possède un marqueur générique, c est dit sommet concept générique, sinon c'est un sommet concept individuel. Il existe un ordre partiel sur les étiquettes des sommets concepts. Etant données deux étiquettes $e = (t, m)$ et $e' = (t', m')$, $e \leq e'$ si et seulement si $t \leq t'$ et $m \leq m'$.*
- *étiq obéit aux contraintes fixées par les applications σ et τ :*
  - *Pour tout $r \in R$, $type(Gi(r)) \leq \sigma_i(type(r))$.*
  - *Pour tout $c \in C$, si $marqueur(c) \in I$, alors $type(c) \geq \tau(marqueur(c))$.*

Dans la pratique, si on ne considère que des relations binaires entre concepts, comme c'est le cas en traitement du langage naturel où les relations codent généralement des relations sémantiques, on peut remplacer un GC, dont la représentation graphique est un graphe non orienté, par un graphe orienté en supprimant le marquage des arêtes par un arc sortant pour le premier argument et rentrant pour le second.

Les concepts sont étiquetés par un couple (définition 5) : $[type(c)marqueur(c)]$. Le type représente la classe à laquelle se rattache le concept et le marqueur (nommé aussi référent) indique une occurrence de la classe (l'individu de la classe) qui peut être :

- une instanciation du concept, c'est-à-dire un marqueur, par exemple : $[usager : `M.Dupond`], [usager : \#x]$ ou $[usager : x]$;
- un individu quelconque de la classe du concept, (marqueur générique *) par exemple : $[usager]$ ou $[usager : *]$.

On peut aussi nommer le marqueur générique par exemple : $[usager : *x]$.

Règles de spécialisation / généralisation

Des règles de dérivation sont définies de la manière suivante :

Soit G un graphe conceptuel défini sur un support donné. On peut obtenir H à partir de G, par :

- *Simplification : si G possède deux sommets relations de même type et ayant les mêmes voisins dans le même ordre, H s'obtient en supprimant l'un des jumeaux.*
- *Restriction de relation : soit r un sommet relation de G. H s'obtient en diminuant l'étiquette de r : le type de r est remplacé par un type plus petit. Cette opération doit respecter les contraintes fixées par la signature de la relation.*

un sommet relation et un sommet concept). Dans nos travaux, les graphes n'ont qu'une arête entre deux sommets.

- *Restriction de concept : soit c un sommet concept de G. H s'obtient en diminuant l'étiquette de c : on remplace étiq(c)=(t,m) par e'=(t',m') avec t'≤t (restriction du type) et m'≤m (restriction du référent).*
- *Jointure interne : soient c1 et c2 deux sommets concepts de G ayant même étiquette, H s'obtient en fusionnant c1 et c2.*

La jointure externe est une opération définie sur deux graphes. Joindre G1 et G2 consiste à construire un graphe G3 contenant les sommets de G1 et de G2, puis à effectuer une ou plusieurs jointures internes.

Ces règles permettent de définir une relation de spécialisation (ou subsomption) sur l'ensemble des GC.

Définition 6 *: Si un GC u défini sur un support donné est dérivable d'un GC v par l'application d'une séquence de règles de formation canonique alors u est appelé une spécialisation de v, u≤v, et v une généralisation de u.*

La relation de spécialisation est à la base de tout raisonnement sur les GC. Pour vérifier qu'un graphe est plus spécialisé qu'un autre, Sowa introduit l'opération de projection (figure 5.1). C'est une opération fondamentale du formalisme : toutes les autres opérations peuvent s'exprimer en termes de projection.

Définition 7 *: Une projection d'un GC G1=(R1,C1,U1,Etiq1) dans un GC G2=(R2,C2,U2,Etiq2) est un couple d'applications* $\pi = (f, g)$,
f : R1 → R2,
g : C1→C2, qui :

- *conserve les arêtes et la numérotation des arêtes : Pour tout arête rc de U1, f(r)g(c) est une arête de U2. De plus, si la ième arête de r est liée à c, la ième arête de f(r) est liée à g(c).*
- *peut restreindre les étiquettes des sommets :*
  pour tout sommet r de R2, $étiq_{G2}(f(r)) \leq étiq_{G1}(r)$ *;*
  pour tout sommet c de C2, $étiq_{G2}(g(c)) \leq étiq_{G1}(c)$.

Théorème : *Soient G et H deux GC, il existe une projection de H dans G si et seulement si G≤H.*

L'isojoint maximal (ou jointure maximale)

Sowa a introduit la jointure maximale qu'il définit comme une jointure sur les projections compatibles étendues au maximum. Cette opération permet l'appariement de deux graphes ; en d'autres termes elle permet de combiner deux graphes en un seul en fusionnant le « maximum » de nœuds (figure 5.2). L'algorithme donné par Sowa est le suivant :

- Commencer par joindre les deux graphes sur un simple concept,
- Etendre la jointure sur une relation conceptuelle à la fois en cherchant des sommets candidats au voisinage immédiat de la partie qui a déjà été jointe,

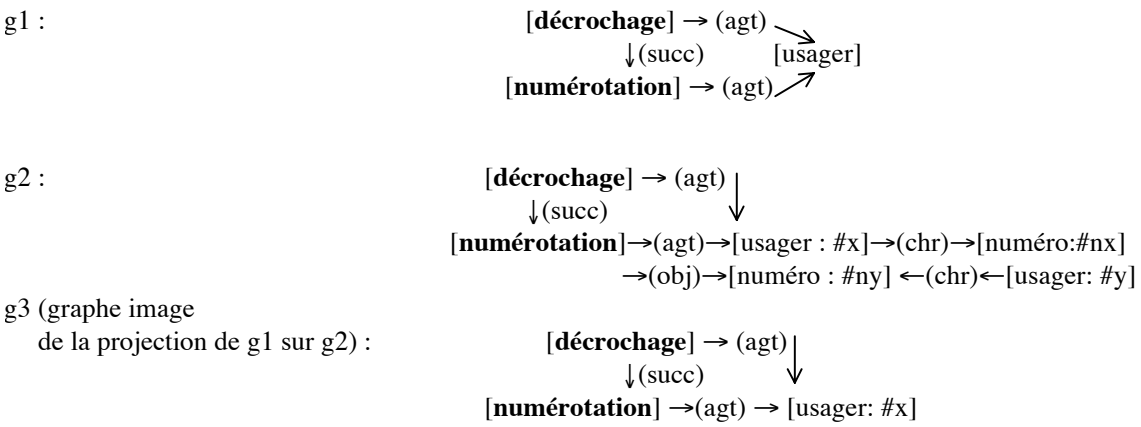


FIGURE 5.1 – Un exemple de projection : la projection de $g1$ sur $g2$ permet de vérifier la présence d'un usager agent du décrochage et de la numérotation (l'usager x)

G1 : [numérotation] →(agt) → [usager :*x] →(chr)→[numéro :*nx]
G2 : [restreindre] -
→(agt) →[usager :#a] → (chr)→[numéro : #na]
→(moyen)→ [analyse] → (obj) → [numéro : #nb]
Grésultat : [usager : #a]-
→ (chr)→[numéro :#na]
←(agt)←[numérotation]
←(agt)←[restreindre]→ (moyen)→[analyse] → (obj) → [numéro :#nb]

FIGURE 5.2 – Un exemple de jointure maximale

– Si aucun candidat ne peut plus être trouvé la jointure est alors localement maximale.

Le sens de maximalité n'est pas vraiment défini par Sowa. Chein et Mugnier proposent trois acceptations du terme. Nous donnons d'abord quelques définitions.

Définition 8 *Deux sommets d'étiquette e et e' sont fusionnables s'il existe un sommet différent du type absurde et d'étiquette e'' telle que $e'' \leq e$ et $e'' \leq e'$.*

Définition 9 *La fusion de deux sommets d'étiquette e et e' et pour borne inférieure un sommet d'étiquette e'' consiste à les identifier en l'unique sommet d'étiquette e''.*

Cette définition tirée de [Chein et Mugnier 1992] s'appliquait aux sommets concepts, nous l'étendons aux sommets relations pour tenir compte du treillis des relations.

Définition 10 *G'_1 et G'_2 deux CG sont dits fusionnables si et seulement si il existe un GC F et deux projections surjectives f et g tels que :*
$f : G'_1 \rightarrow F$
$g : G'_2 \rightarrow F$

Définition 11 *Une jointure étendue entre deux GC $G1$ et $G2$ consiste à fusionner deux de leur sous-GC G'_1 et G'_2 qui sont fusionnables .*

Nous intéresse plus particulièrement le cas correspondant à la jointure maximale donnée par Sowa où les deux sous-graphes à fusionner sont isomorphes et qui est définie par Chein et Mugnier comme un isojoint de la manière suivante [2] :

Définition 12 *Soient $G = (R, C, U, eti)$ et $G' = (R', C', U', eti')$ deux CG, α une bijection de $A \subseteq R$ sur $A' \subseteq R'$ et β une bijection de $B \subseteq C$ sur $B' \subseteq C'$. Le couple de bijections $\gamma = (\alpha, \beta)$ est dit compatible si l'on a :*

1. *$\forall c \in B, c$ et $\beta(c)$ sont fusionnables,*
2. *$\forall r \in R,\ r$ et $\alpha(r)$ sont fusionnables, $G(r) \subseteq B,$ et $\forall i, 1 \leq i \leq degré(r)$, $\beta(G_i(r)) = G'_i(r')$ et de plus, si r'' est le sommet relation résultant de la fusion de r et de $\alpha(r)$ alors les sommets arguments de la relation r'' doivent respecter les contraintes fixées par la signature de la relation r'',*
3. *$G[A \cup B]$ et $G'[A' \cup B']$ sont connexes [3].*

On dira que $G[A \cup B]$ et $G'[A' \cup B']$ sont isofusionnables et la fusion associée est appelée isojoint entre G et G' suivant le couple de bijections compatibles γ.

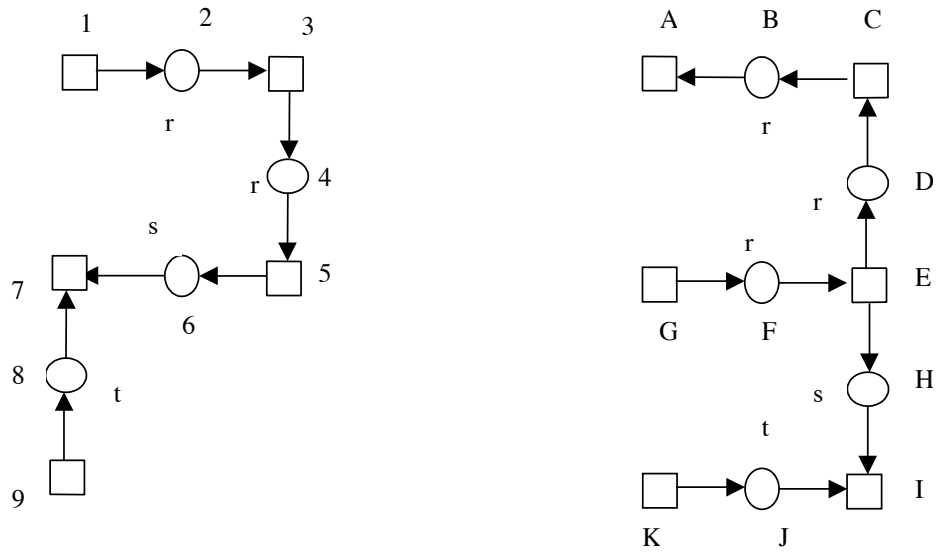
On peut maintenant exposer les trois sens possibles de la notion de maximalité (proposés par Chein et Mugnier) pour la jointure étendue telle qu'elle est définie ci-dessus :

- ou bien le cardinal de $A \cup B$ (ou de γ) est maximal ;
- Ou bien il n'existe pas de sous-graphe de G qui soit solution et contienne strictement $A \cup B$ (maximalité de $A \cup B$ au sens de l'inclusion) ;
- Ou bien il n'existe pas d'extension de γ qui conserve la connexité du sous-graphe fusionné (maximalité de γ au sens de l'inclusion).

La figure 5.3 montre un exemple des différentes maximalités d'un isojoint (ou jointure). Dans nos travaux (chapitre 2), nous utilisons le premier sens de la maximalité (maximalité du cardinal de $A \cup B$).

2. Nous adaptons cette définition pour prendre en compte la fusion des relations. Dans [Chein et Mugnier 1992], la jointure sur les sommets relations n'est réalisable qu'en cas d'égalité.

3. $G[A \cup B]$ désigne le GC restreint aux sommets de $A \cup B$

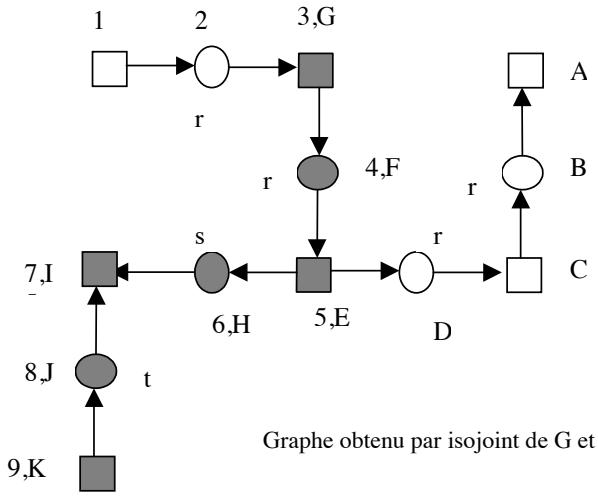


Tous les sommets concepts sont supposés fusionnables deux à deux.

γ maximal au sens de l'inclusion :
γ1 : α1={(2,B)} et β1={(1,C),(3,A)}

A∪B maximal au sens de l'inclusion :
γ2 : α2={(2,F),(4,D)} et β2={(1,G),(3,E),(5,C)}

A∪B ou γ maximal au sens du cardinal :
γ3 : α3={(4,F),(6,H),(8,J)} et β3={(3,G),(5,E),(7,I),(9,K)}



Graphe obtenu par isojoint de G et G ' suivant γ3

FIGURE 5.3 – Exemple d'isojoints maximaux entre deux GC [Chein et Mugnier 1992]

Bibliographie

Agrawal R., Imielinski T., Swami A., « Mining association rules between sets of items in large databases », *ACM SIGMOD Intl. Conf. Management of Data*, May, 1993.

Agrawal R., Srikant R., « Mining sequential patterns », *Proc. of the 11th Int. Conf. on Data Engineering (ICDE'95)*, p. 3-14, 1995.

Allen J., « Maintaining knowledge about temporal intervals », *Communication of the ACM*, vol. 26, n° 11, p. 832-843, 1983.

Amghar T., Battistelli D., Charnois T., « Reasoning on French Aspecto-Temporal Information in French within Conceptual Graphs », *Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'2002)*, Washington D.C., USA, p. 315-324, November, 2002.

Amghar T., Charnois T., Battistelli D., « Aspecto-Temporal Data and Lexical Representations in French within Simple Conceptual Graphs on the Basis of Semantico-Cognitive Schemes », *in* H. S. Delugach, G. Stumme (eds), *Conceptual Structures : Broadening the Base, 9th International Conference on Conceptual Structures, ICCS 2001*, vol. 2120 of *Lecture Notes in Computer Science*, Springer, Stanford, CA, USA, p. 29-43, July, 2001.

Amghar T., Gayral F., Levrat B., « Table 10 left without paying the bill ! A Good Reason to Treat Metonymy with Conceptual Graphs », *in* G. Ellis, R. Levinson, W. Rich, J. F. Sowa (eds), *Conceptual Structures : Applications, Implementation and Theory — Proceedings of the Third International Conference on Conceptual Structures (ICCS'1995)*, Santa Cruz, USA, p. 129-143, 1995.

Azé J., Heitz T., Mela A., Mezaour A., Peinl P., Roche M., « Présentation de DEFT'06 (DEfi Fouille de Textes) », *Actes de l'atelier DEFT'06, SDN'06 (Semaine du Document Numérique)*, p. 1-10, 2006.

Battistelli D., Passer du texte à une séquence d'images : analyse spatio-temporelle de textes, modélisation et réalisation informatique (système SPAT), PhD thesis, Université Paris IV - Sorbonne, 2000.

Battistelli D., Vazov N., « Building of semantico-cognitive representation of different types of narratives in French », *Proceedings of the Second International Conference on Recent Advances in Natural Language Processing (RANLP 97)*, Tzigov Chark, Bulgaria, 1997.

Bestougeff H., Ligozat G., *Outils logiques pour le traitement du temps, de la linguistique á l'intelligence artificielle*, Masson, 1989.

Bilhaut F., Charnois T., Enjalbert P., Mathet Y., « Passage Extraction in Geographical Documents », *Proceedings of New Trends in Intelligent Information Processing and Web Mining (IIPWM'03)*, Zakopane, Pologne, p. 121-130, June, 2003a.

Bilhaut F., Charnois T., Mathet Y., Enjalbert P., « Geographic Reference Analysis for Geographic Document Querying », *in* K. A., S. B. (eds), *Proceedings of the Workshop on the Analysis of Geographic References, Human Language Technology Conference (NAACL-HLT)*, Association for Computational Linguistics, Edmonton, Canada, p. 55-62, May, 2003b.

Bilhaut F., Dumoncel F., Enjalbert P., Hernandez N., « Indexation sémantique et recherche d'information interactive », *Actes CORIA 2007*, Université de Saint-Étienne, p. 65-76, 2007.

Bilhaut F., Ho Dac L., Borillo A., Charnois T., Enjalbert P., Le Draoulec A., Mathet Y., Miguet H., Péry-woodley M.-P., Sarda L., « Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique », *Actes de la 10e Conférence Traitement Automatique du Langage Naturel (TALN'03)*, Batz-sur-Mer, France, p. 315-320, 2003c.

Bilhaut F., Widlöcher A., « LinguaStream : An Integrated Environment for Computational Linguistics Experimentation », *Proceedings of the 11th Conference of the European Chapter of the Association of Computational Linguistics (EACL'06), Companion Volume*, Trento, Italy, p. 95-98, April, 2006.

Bourigault D., « Extraction et structuration automatiques de terminologie pour l'aide à l'acquition des connaissances à partir de textes », *9ème Congrès Reconnaissance des Formes et Intelligence Artificielle, AFCET-RFIA*, Paris, France, p. 397-408, January, 1994.

Brachman R. J., « On the Epistemological Status of Semantic Networks », *in* N. V. Findler (ed.), *Associative Networks : Representation and Use of Knowledge by Computers*, Academic Press, New York, p. 3-50, 1979.

Bérard-Dugourd A., Fargues J., Landau M.-C., « Natural language analysis using conceptual graphs », *Proceedings of the International Computer Science Conference '88*, Honk-Kong, p. 265-272, 1988.

Califf M. E., Mooney R. J., « Relational Learning of Pattern-Match Rules for Information Extraction », *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, AAAI, p. 328-334, 1999.

Cellier P., Charnois T., « Fouille de données séquentielle d'itemsets pour l'apprentissage de patrons linguistiques », *Traitement Automatique des Langues Naturelles (short paper)*, Montréal, Canada, 2010.

Cellier P., Charnois T., Plantevit M., « Sequential Patterns to Discover and Characterise Biological Relations », *in* A. F. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing (CICLing 2010)*, LNCS 6008, Springer, p. 537-548, 2010a.

Cellier P., Charnois T., Plantevit M., Crémilleux B., « Recursive Sequence Mining to Discover Named Entity Relations », *in* P. R. Cohen, N. M. Adams, M. R. Berthold (eds), *Advances in Intelligent Data Analysis IX, 9th International Symposium (IDA 2010)*, LNCS 6065, Springer, p. 30-41, 2010b.

Cellier P., Ferré S., Ducassé M., Charnois T., « Partial Orders and Logical Concept Analysis to Explore Patterns Extracted by Data Mining », *Conceptual Structures for Discovering Knowledge - 19th International Conference on Conceptual Structures, ICCS 2011*, vol. 6828 of *Lecture Notes in Computer Science*, Springer, p. 77-90, 2011.

Chang C.-H., Kayed M., Ramzy M., Shaalan K. F., « A Survey of Web Information Extraction Systems », *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, n° 10, p. 1411-1428, October, 2006.

Charnois T., Détection d'anomalies sémantiques dans les textes de spécifications – Application au problème des interactions de services de télécommunication, PhD thesis, Université Paris XIII, 1999.

Charnois T., « Maximal Isojoin for Representing Software Textual Specifications and Detecting Semantic Anomalies », *in* G. Stumme (ed.), *Working with Conceptual Structures, 8th International Conference on Conceptual Structures, ICCS'2000*, Darmstadt, Germany, p. 189-200, July, 2000.

Charnois T., Doucet A., Mathet Y., Rioult F., « Trois approches du GREYC pour la classification de textes », *Défi fouille de texte 2008*, p. 171-180, 2008.

Charnois T., Durand N., Kléma J., « Automated Information Extraction from Gene Summaries », *Proceedings of the ECML/PKDD Workshop on Data and Text Mining for Integrative Biology*, Berlin, Germany, p. 4-15, September, 2006.

Charnois T., Enjalbert P., « Compréhension automatique », *in* P. Enjalbert (ed.), *Sémantique et traitement automatique du langage naturel*, Traité IC2, série Cognition et traitement de l'information, Hermès Sciences, Lavoisier, chapter 7, p. 267-308, 2005.

Charnois T., Plantevit M., Rigotti C., Crémilleux B., « Fouille de données séquentielles pour l'extraction d'information dans les textes », *TAL*, vol. 50, n° 3, p. 59-87, 2009.

Charolles M., « L'encadrement du discours - Univers, champs, domaines et espaces », *Cahier de recherche linguistique*, 1997.

Chein M., Mugnier M., « Conceptual graphs : fundamental notions », *Revue d'Intelligence Artificielle*, vol. 6, n° 4, p. 365-406, 1992.

Cohen A. M., Hersh W. R., « A survey of current work in biomedical text mining. », *Brief Bioinform*, vol. 6, n° 1, p. 57-71, March, 2005.

Cohen K. B., Hunter L., « Natural Language Processing and Systems Biology », *Artificial Intelligence Methods and Tools for Systems Biology*, Springer, p. 147-173, 2004.

Consortium Menelas, *Menelas final report, Technical Report Deliverable report AIM-Menelas 17*, DIAM-SIM/INSERM U.194, 1995.

Couto J., Ferret O., Grau B., Hernandez N., Jackiewicz A., Minel J.-L., Porhiel S., « RÉGAL, un système pour la visualisation sélective de documents. », *Revue d'Intelligence Artificielle*, vol. 18, n° 4, p. 481-514, 2004.

Davey B. A., Priestley H. A., *Introduction To Lattices And Order*, Cambridge University Press, 1990.

Desclés J.-P., Jouis C., OH H., Maire-Reppert D., « Exploration contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte », *in* J. R. D. Herin-Aime, R. Dieng, J. Angoujard (eds), *Knowledge modeling and expertise transfer*, IOS Press, p. 371-400, 1991.

Desclés J.-P., « State, Events, Process and Topology », *General Linguistics*, vol. 29, n° 3, p. 159-200, 1989.

Desclés J.-P., *Langages applicatifs, Langues naturelles et Cognition*, Hermès, 1990.

Draoulec A. L., Pery-Woodley M.-P., « Corpus-based identification of temporal organisation in discourse », *in* P. Rayson, A. Wilson, T. McEnery, A. Hardie, S. Khoja (eds), *Proceedings of the Corpus Linguistics 2001 Conference*, Lancaster, 2001.

Enjalbert P., Bilhaut F., « L'accès assisté à l'information documentaire », *in* P. Enjalbert (ed.), *Sémantique et traitement automatique du langage naturel*, Traité IC2, série Cognition et traitement de l'information, Hermès Sciences, Lavoisier, chapter 9, p. 335-370, 2005.

Esuli A., Sebastiani F., « SentiWordNet : A Publicly Available Lexical Resource for Opinion Mining », *Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation*, Genova, Italy, 2006.

Ferrari S., *Etudes pour le traitement automatique des langues – De la rhétorique des figures à la rhétorique du discours*, Habilitation à Diriger des Recherches, Université de Caen, décembre, 2010.

Ferrari S., Bilhaut F., Widlöcher A., Laignelet M., « Une plate-forme logicielle et une démarche pour la validation de ressources linguistiques sur corpus : application à l'évaluation de la détection automatique de cadres temporels », *in* G. Williams (ed.), *Actes des 4èmes Journées de la Linguistique de Corpus (JLC'05)*, Lorient, France, septembre, 2005.

Ferrari S., Charnois T., Jackiewicz A., Gardin P., Widlöcher A., « Jugements d'évaluation : des constituants détachés à d'autres formes régulières », *Actes des 6èmes Journées de la Linguistique de Corpus*, septembre, Université de Bretagne Sud, Lorient, France, 2009a.

Ferrari S., Charnois T., Mathet Y., cois Rioult F., Legallois D., « Analyse de discours évaluatif, modèle linguistique et applications », *Revue des Nouvelles Technologies de l'Information*, vol. E-17, p. 71-93, 2009b. numéro spécial Fouille de données d'opinions.

Ferret O., Grau B., Minel J.-L., Porhiel S., « Repérage de structures thématiques dans des textes », *Traitement Automatique des Langues Naturelles (TALN'01)*, Tours, France, juillet, 2001.

Ferré S., « Camelis : a logical information system to organize and browse a collection of documents », *Int. J. General Systems*, 2009.

Fillmore C., « The Case for Case », *in* E. Bach, R. Harms (eds), *Universals in Linguistic Theory*, Holt, Rinehart and Winston, p. 1-90, 1968.

Fukuda K., Tamura A., Tsunoda T., Takagi T., « Toward Information Extraction : Identifying Protein Names from Biological Papers », *Pacific Symposium Biocomputing (PSB'98)*, Hawaii, p. 362-373, Jan., 1998.

Gaizauskas R., Demetriou G., Artymiuk P. J., Willett P., « Protein structures and information extraction from biological texts : the PASTA system. », *Bioinformatics*, vol. 19, n° 1, p. 135-143, January, 2003.

Garofalakis M., Rastogi R., K. S., « SPIRIT : Sequential Pattern Mining with Regular Expression Constraints », *Proc. of the 25th Int. Conf. on Very Large Databases (VLDB'99)*, p. 223-234, 1999.

Gosselin L., *Sémantique de la temporalité en français*, Duculot, Louvain-la-Neuve, 1996.

Gosselin L., « Temporalité », *in* P. Enjalbert (ed.), *Sémantique et traitement automatique du langage naturel*, Traité IC2, série Cognition et traitement de l'information, Hermès Sciences, Lavoisier, chapter 5, p. 173-213, 2005.

Halliday M., Hasan R., *Cohesion in English*, English language series, Longman, 1976.

Han J., *Data Mining : Concepts and Techniques*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

Hao Y., Zhu X., Huang M., Li M., « Discovering Patterns to Extract Protein- Protein Interactions from the Literature », *Bioinformatics*, vol. 21, n° 15, p. 3294-3300, 2005.

Harb A., Dray G., Plantié M., Poncelet P., Roche M., Trousset F., « Détection d'Opinion : Apprenons les bons Adjectifs ! », *in* M. Roche, P. Poncelet (eds), *INFORSID'08 : Informatique des Organisations et Systèmes d'Information et de Décision - Atelier FODOP'08 (FOuille des Données d'OPinions)*, Fontainebleau, France, INFORSID, p. 59-66, 2008.

Hearst M., « Multi-paragraph Segmentation of Expository Text », *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, 1994.

Hearst M. A., « Automatic Acquisition of Hyponyms from Large Text Corpora », *In Proceedings of the 14th International Conference on Computational Linguistics*, p. 539-545, 1992.

Hobbs J. R., Riloff E., « Information Extraction », *in* N. Indurkhya, F. J. Damerau (eds), *Handbook of Natural Language Processing, Second Edition*, CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010. ISBN 978-1420085921.

Hoey M., *Patterns of Lexis in Text*, Describing English Language, Oxford University Press, 1991.

Hu M., Liu B., « Mining opinion features in customer reviews », *in* D. L. McGuinness, G. Ferguson (eds), *Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence AAAI 2004*, San Jose, California, USA, AAAI Press / The MIT Press, p. 755-760, 2004.

Hurault-Plantet M., Berthelin J.-B., El Ayari S., Grouin C., Loiseau S., Paroubek P., « Résultats de l'édition 2008 du DÉfi Fouille de Textes », *Actes de JEP–TALN–RECITAL 2008*, Avignon, France, 2008. 13 juin 2008.

Hérin R., *Atlas de la France Scolaire de la Maternelle au Lycée*, RECLUS - La Documentation Française, Collection Dynamiques du Territoire, 1994.

Jackiewicz A., « Structures avec constituants détachés et jugements d'évaluation », *Document Numérique*, vol. 13, n° 3, p. 11-40, 2010.

Jackiewicz A., Charnois T., Ferrari S., « Jugements d'évaluation et constituants périphériques », *Actes de la 16ème conférence sur le traitement automatique des langues naturelles (TALN'09)*, juin, Senlis, France, 2009a.

Jackiewicz A., Vigier D., Charnois T., Ferrari S., « Vers une analyse automatique des discours évaluatifs. Le cas des constituants détachés "en N <exp>" », *in* L. Sarda (ed.), *Linguistic and Psycholinguistic Approaches to Text Structuring*, September, ENS, Paris, France, 2009b.

Janetzko D., Cherfi H., Kennke R., Napoli A., Toussaint Y., « Knowledge-Based Selection of Association Rules for Text Mining. », *in* R. L. de Mántaras, L. Saitta (eds), *ECAI*, IOS Press, p. 485-489, 2004.

Kamp H., « Events, Instants, and Temporal Reference », *in* R. Bauerle, U. Egli, A. von Stechow (eds), *Semantics from different points of view*, Springer Verlag, p. 373-417, 1979.

Kayser D., « Traitement du langage naturel : un regard sur son évolution et ses perspectives », *Technique et Science Informatiques*, vol. 19, n° 1-2-3, p. 291-301, 2000.

Klema J., Blachon S., Soulet A., Crémilleux B., Gandrillon G., « Constraint-Based Knowledge Discovery from SAGE Data », *In Silico Biology*, 2008.

Krallinger M., Leitner F., Rodriguez-Penagos C., Valencia A., « Overview of the protein-protein interaction annotation extraction task of BioCreative II », *Genome Biology*, 2008.

Leclère M., Les connaissances de niveau terminologique du modèle des graphes conceptuels : construction et exploitation, PhD thesis, Université Montpellier II, 1995.

Legallois D., *Système, norme, parole : application au lexique, au texte, à la phrase et aux constructions grammaticales*, Habilitation à Diriger des Recherches, Université de Caen, novembre, 2011.

Legallois D., Cellier P., Charnois T., « Calcul de réseaux phrastiques pour l'analyse et la navigation textuelle », *Actes de la 18e conférence sur le traitement automatique des langues naturelles (TALN ?11)*, Montpellier, juillet, 2011.

Legallois D., Ferrari S., « Vers une grammaire de l'évaluation des objets culturels », *Actes d'ISDD06, colloque international Discours et Document, Schedae, 2006, fascicule n° 1*, Caen, 15 et 16 juin 2006, Presses universitaires de Caen, p. 57-68, 2006. prépublication n° 8.

Leleu M., Rigotti C., Boulicaut J.-F., Euvrard G., « Constraint-Based Mining of Sequential Patterns over Datasets with Consecutive Repetitions », *in* N. Lavrac, D. Gamberger, H. Blockeel, L. Todorovski (eds), *PKDD*, vol. 2838 of *Lecture Notes in Computer Science*, Springer, p. 303-314, 2003.

Leser U., Hakenberg J., « What makes a gene name ? Named entity recognition in the biomedical literature. », *Brief Bioinform*, vol. 6, n° 4, p. 357-369, December, 2005.

Liquière M., Brissac O., « A class of conceptual graphs with polynomial iso-projection », *in* J. P. D. William M. Tepfenhart, J. F. Sowa (eds), *Second International Conference on Conceptual Structures— Proceedings Supplement*, p. 278-292, 1994.

Manguin J.-L., Plantevit M., Charnois T., « Constitution d'un corpus pré-annoté de résumés d'articles en biologie génétique », *in* G. Williams (ed.), *Actes des 6èmes Journées de la Linguistique de Corpus*, Université de Bretagne-Sud, Lorient, France, septembre, 2009. Poster.

Mathet Y., Widlöcher A., « Une approche holiste et unifiée de l'alignement et de la mesure d'accord inter-annotateurs », *Actes TALN 2011*, Montpellier, juin, 2011.

Moulin B., « The representation of linguistic information in an approach used for modelling temporal knowledge in discourses », *in* M. G.W., M. B., S. J.F. (eds), *Conceptual Graphs for knowledge Representation*, n° 699 in *Lecture Notes in Artificial Intelligence*, Springer Verlag, p. 182-204, 1993.

Moulin B., Coté D., « Extending the Conceptual Graph Model for Differentiating Temporal and Non-Temporal Knowledge », *in* T. E. Nagle, J. A. Nagle, L. L. Gehorlz, P. W. Eklund (eds), *Conceptual Structures and Practice*, p. 381-390, 1992.

Mugnier M., « On generalisation / specialisation for conceptual graphs », *Journal of Experimental and Theorical Artificial Intelligence*, vol. 7, n° 3, p. 325-344, 1995.

Mugnier M., Chein M., « Représenter des connaissances et raisonner avec des graphes », *Revue d'Intelligence Artificielle*, vol. 10, n° 1, p. 7-56, 1996.

Muller P., Tannier X., « Annotating and measuring temporal relations in texts », *Proceedings of Coling 2004*, vol. I, Genève, p. 50-56, 2004.

Nanni M., Rigotti C., « Extracting Trees of Quantitative Serial Episodes », *Knowledge Discovery in Inductive Databases 5th Int. Workshop KDID'06, Revised Selected and Invited Papers*, Springer-Verlag LNCS 4747, p. 170-188, 2007.

Nazarenko A., « Representing Natural Language Causality in Conceptual Graphs : the Higher Order Conceptual Relation Problem », *in* G. W. Mineau, B. Moulin, J. F. Sowa (eds), *Conceptual Graphs for Knowledge Representation — Proceedings of the First International Conference on Conceptual Structures*, Québec City, Canada, p. 205-222, 1993.

Nazarenko A., « Le point sur l'état actuel des connaissances en traitement automatique des langues », *in* G. Sabah (ed.), *Compréhension des langues et interaction*, Traité IC2, série Cognition et traitement de l'information, Hermès Sciences, Lavoisier, p. 31-70, 2006.

Nédellec C., Bessieres P., Bossy R., Kotoujansky A., Manine A.-P., « Annotation Guidelines for Machine Learning-Based Named Entity Recognition in Microbiology », *Proceedings of the Data and text mining in integrative biology workshop, ECML/PKDD*, Berlin, p. 40-54, Sept., 2006.

Nouvel D., Soulet A., Antoine J.-Y., Friburger N., Maurel D., « Reconnaissance d'entités nommées : enrichissement d ?un système à base de connaissances à partir de techniques de fouille de textes », *Traitement Automatique des Langues Naturelles (TALN'10)*, 2010.

Pang B., Lee L., « A Sentimental Education : Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts », *Proceedings of the 42nd Annual Meeting of the ACL*, Barcelona, Spain, ACL, p. 271-278, 2004.

Paroubek P., Berthelin J.-B., El Ayari S., Grouin C., Heitz T., Hurault-Plantet M., Jardino M., Khalis Z., Lastes M., « Résultats de l'édition 2007 du DÉfi Fouille de Textes », *Actes de l'atelier de clôture du 3ème Défi Fouille de Textes*, Grenoble, France, p. 9-17, 2007. 3 juillet 2007.

Pei J., Han J., Wang W., « Mining sequential patterns with constraints in large databases », *CIKM*, ACM, p. 18-25, 2002.

Plantevit M., Charnois T., Kléma J., Rigotti C., Crémilleux B., « Combining Sequence and Itemset Mining to Discover Named Entities in Biomedical Texts : A New Type of Pattern », *International Journal of Data Mining, Modelling and Management*, vol. 1, n° 2, p. 119-148, 2009.

Read J., Hope D., Carroll J., « Annotating Expressions of Appraisal in English », *Proceedings of the Linguistic Annotation Workshop*, Prague, Czech Republic, ACL, p. 93-100, 2007.

Reichenbach H., *Elements of symbolic logic*, McMillan, 1947.

Riloff E., « Automatically Generating Extraction Patterns from Untagged Text », *AAAI/IAAI, Vol. 2*, p. 1044-1049, 1996.

Roche M., Azé J., Matte-Tailliez O., Kodratoff Y., « Mining texts by association rules discovery in a technical corpus. », *in* M. A. Klopotek, S. T. Wierzchon, K. Trojanowski (eds), *Intelligent Information Systems*, Advances in Soft Computing, Springer, p. 89-98, 2004.

Sabah G., « Sens et traitements automatiques des langues », *in* J.-M. Pierrel (ed.), *Ingénierie des langues*, Information - Commande - Communication, Hermès Sciences, chapter 3, p. 77-108, 2000.

Sabah G., « Compréhension automatique des langues ; où va-t-on ? Où pourrait-on aller ? », *in* G. Sabah (ed.), *Compréhension des langues et interaction*, Traité IC2, série Cognition et traitement de l'information, Hermès Sciences, Lavoisier, p. 375-393, 2006.

Sallaberry C., Baziz M., Lesbegueries J., Gaio M., « Une approche d'extraction et de recherche d'information spatiale dans les documents textuels - évaluation. », *Actes CORIA 2007*, Université de Saint-Étienne, p. 53-64, 2007.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *Proceedings of International Conference on New Methods in Language Processing*, September, 1994.

Schneider G., Kaljurand K., Rinaldi F., « Detecting Protein-Protein Interactions in Biomedical Texts Using a Parser and Linguistic Resources », *in* A. F. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing, 10th International Conference, CICLing 2009*, Mexico, p. 406-417, March, 2009.

Smith L., Tanabe L., Ando R., Kuo C., Chung I., Hsu C., Lin Y., Klinger R., Friedrich C., Ganchev K., Torii M., Liu H., Haddow B., Struble C., Povinelli R., Vlachos A., Baumgartner W., Hunter L., Carpenter B., Tsai R., Dai H., Liu F., Chen Y., Sun C., Katrenko S., Adriaans P., Blaschke C., Torres R., Neves M., Nakov P., Divoli A., Maña López M., Mata J., Wilbur W., « Overview of BioCreative II gene mention recognition », *Genome Biology*, 2008.

Soulet A., « Résumer les contrastes par l'extraction récursive de motifs. », *Actes de CAP'07, Conférence francophone sur l'apprentissage automatique - 2007, Grenoble, France*, 2007.

Sowa J. F., *Conceptual Structures : Information Processing in Mind and Machine*, Addison-Wesley, 1984.

Tantzen S., *« Ein Prologparser für temporale und lokale Ausdrücke in einem GeoSem-System für das Deutsche »*, Master's thesis, Fakultät II (Sprach- und Literaturwissenschaften), der Friedrich-Alexander-Universität Erlangen-Nürnberg, 2004.

Tellier I., « Préface », *TAL*, vol. 50, n° 3 (numéro spécial Apprentissage Automatique pour le TAL), p. 7-21, 2009.

Temkin J., Gilder M., « Extraction of protein interaction information from unstructured text using a context-free grammar. Bioinformatics », *Bioinformatics*, vol. 19, n° 16, p. 2046-2053, 2003.

Tsuruoka Y., Tsujii J.-I., « Probabilistic term variant generator for biomedical terms », *SIGIR*, p. 167-173, 2003.

Turmel L., Lucas N., B. C., « Extraction d'associations pour la caractérisation de segments de textes en anglais avec et sans faute », *Conférence internationale sur le document électronique (CIDE'03)*, Europia, Caen, p. 221-237, novembre, 2003.

Turney P., « Thumbs Up or Thumbs Down ? Semantic Orientation Applied to Unsupervised Classification of Reviews », *Proceedings of the 40th Annual Meeting of the ACL (ACL'02)*, Philadelphia, Pennsylvania, USA, ACL, p. 417-424, 2002.

Vaid S., Jones C. B., Joho H., Sanderson M., « Spatio-textual Indexing for Geographical Search on the Web », *9th International Symposium on Spatial and Temporal Databases SSTD 2005*, vol. 3633 of *Lecture Notes in Computer Science*, p. 218-235, 2005.

Vernier M., Ferrari S., « Tracking evaluation in discourse », *ASOS07, Workshop on Applications of Semantics, Opinions and Sentiments*, Eurolan summer school, July 23 - August 3, 2007, University of Iaşi, Romania, 2007.

Vernier M., Mathet Y., Rioult F., Charnois T., Ferrari S., Legallois D., « Classification de textes d'opinions : une approche mixte n-grammes et sémantique », *Troisième DÉfi de Fouille de Textes (DEFT'07), plate-forme AFIA 2007*, AFIA, Assiciation FranÃ§aise d'Intelligence Artificielle, Grenoble, France, p. 95-109, juillet, 2007.

Whitelaw C., Garg N., Argamon S., « Using Appraisal Taxonomies for Sentiment Analysis », *Proceedings of CIKM-05, the ACM SIGIR Conference on Information and Knowledge Management*, 2005.

Widlöcher A., Bilhaut F., Hernandez N., Rioult F., Charnois T., Ferrari S., Enjalbert P., « Une approche hybride de la segmentation thématique : collaboration du traitement automatique des langues et de la fouille de texte », *Deuxième DÉfi de Fouille de Textes (DEFT'06), Semaine du Document Numérique (SDN'2006)*, 18-22 septembre, Fribourg, Suisse, 2006.

Wiebe J., Wilson T., Cardie C., « Annotating expressions of opinions and emotions in language », *Language Resources and Evaluation*, vol. 39, n° 2-3, p. 165-210, 2005.

Willems M., « Projection and unification for conceptual graphs », *in* W. R. Gerard Ellis, Robert Levinson, J. F. Sowa (eds), *Conceptual Structures : Applications, Implementation and Theory — Proceedings of the Third International Conference on Conceptual Structures*, vol. 954 of *Lecture Notes in Artificial Intelligence*, Springer-Verlag, Santa Cruz, USA, p. 278-292, 1995.

Wonsever D., Minel J.-L., « Contextual Rules for Text Analysis », *CICLing '01 : Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing*, Springer, London, UK, p. 509-523, 2001.

Woodruff A., Plaunt C., « GIPSY : Automated Geographic Indexing of Text Documents », *JASIS*, vol. 45, n° 9, p. 645-655, 1994.

Woods W. A., « What's in a Link : Foundation for Semantic Networks », *in* D. G. Bobrow, A. M. Collins (eds), *Representation and Understanding : Studies in Cognitive Science*, Academic Press, London, p. 35-82, 1975.

Yan X., Han J., Afshar R., « CloSpan : Mining Closed Sequential Patterns in Large Databases », *in* D. Barbará, C. Kamath (eds), *SDM*, SIAM, 2003.

Yeh A., Morgan A., Colosimo M., Hirschman L., « BioCreAtIvE Task 1A : gene mention finding evaluation », *BMC Bioinformatics*, vol. 6, p. 10, 2005.

Yu H., Hatzivassiloglou V., « Towards Answering Opinion Questions : Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences », *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, Sapporo, Japan, ACL, p. 129-136, 2003.

Zweigenbaum P., Demner-Fushman D., Yu H., Cohen K. B., « Frontiers of biomedical text mining : current progress », *Brief Bioinform*, vol. 8, n° 5, p. 358-375, October, 2007.