# Un système de question-réponse dans le domaine médical : le système Esculape

Mehdi Embarek

Université de Paris-Est

Ecole doctorale : Information, Communication, Modélisation et Simulation (ICMS)

THÈSE

pour obtenir le grade de

Docteur de l'Université Paris-Est

Spécialité : INFORMATIQUE

présentée et soutenue publiquement par

Mehdi EMBAREK

le : 04 juillet 2008

# Un système de question-réponse dans le domaine médical

## Le système Esculape

## A question answering system in the medical domain

### The Esculape system

Directeur de thèse
Christian FLUHR

Jury
Brigitte Grau (Rapporteur)
Pierre Zweigenbaum (Rapporteur)
Christian Fluhr (Directeur)
Patrice Bellot (Examinateur)
Olivier Ferret (Examinateur)

# Remerciements

Je tiens en premier lieu à remercier Olivier Ferret pour avoir assuré le suivi de ma thèse et pour l'attention qu'il y a portée. Sa patience, sa disponibilité, ses conseils et ses qualités scientifiques ont été très précieux pour mener à bien cette thèse. Qu'il trouve dans ces quelques mots l'expression de ma profonde gratitude.

Je remercie mon directeur de thèse Christian Fluhr pour m'avoir donné la possibilité de réaliser cette thèse au sein du laboratoire LIC2M. Les remarques et corrections qu'il a prodiguées ont été d'une aide précieuse.

Je remercie Brigitte Grau et Pierre Zweigenbaum pour avoir accepté d'être rapporteur de ce travail. Leurs commentaires et leurs suggestions m'ont permis d'améliorer la qualité de ce manuscrit.

Je remercie Patrice Bellot pour avoir accepté d'examiner cette thèse et de faire partie de mon jury.

J'adresse mes sincères remerciements aux personnes qui ont accepté de relire la première version de ma thèse et qui ont contribué à faire de ce document ce qu'il est aujourd'hui : Delphine Lagarde et Laurent Gillard.

Je remercie respectivement Rodolph Gelin et Arnauld Leservot pour m'avoir accueilli au sein de leur service.

Je remercie tous mes amis et collègues du LIC2M pour leur soutien, leurs encouragements et… les croissants du matin. Ils ont dû supporter mon humeur et mes blagues durant tous ce temps. Merci à mes colocataires de bureau : Benoît Mathieu, Delphine Lagarde et Faïza Gara, pour leur bonne humeur et les fous rires. Merci à Pierre-Alain Moellic pour ces parties de Squash et de Tennis. J'espère que le prochain challenger continuera à enchaîner les victoires. Bien sûr merci à Halima Dahmani et Nasredine Semmar pour leurs conseils ainsi qu'à Meriama Laib-Boukhari et son légendaire Tiramisu. Sans oublier Olivier Mesnard, Gregory Grefenstette, Romaric Besançon, Gaël De Chalendar, Patrick Hède, Hervé Le Borgne, Bertrand Delezoïde, Sofiane Souidi, Christophe Millet, Adrian Popescu et Marc Mergy. Vous avez été une seconde famille pour moi ! Encore une fois merci.

Je remercie tous les thésards et anciens stagiaires du LIC2M pour leur soutien, nos discussions et nos pauses.

Enfin, une pensée particulière à mes parents qui m'ont toujours encouragé et soutenu pour mener à terme ce travail.

Ces remerciements ne seraient pas complets sans mes pensées pour les êtres qui me sont les plus chers. Ainsi, je dédie cette thèse :

**À mes grands parents,**

**À mes parents Mohamed et Nadia,**
Nulle dédicace ne serait vous exprimer toute ma reconnaissance et tout mon amour. Vous m'avez particulièrement encouragé et aidé durant toutes mes années d'études. Ma reconnaissance vous est éternelle pour l'éducation et les principes que vous m'avez inculqués.
Que ce travail soit preuve de mon éternelle reconnaissance.

**À mon frère et ma sœur,**
À Walid et Amel

**À Baghdadi Laalaouna,**
L'admiration et l'estime qu'impose votre qualité humaine, m'ont poussé et incité pour mener à terme ce travail. Merci pour votre encouragement et votre soutien.
Veuillez trouver dans ce travail l'expression de mon profond respect.

**À Delphine,**
Tu m'as remarquablement encouragé et réconforté dans les moments difficiles. Ton aide morale et ton soutien m'ont été d'un immense soutien dans l'élaboration de ce manuscrit. Merci d'avoir toujours cru en moi et pour tout le bonheur que tu me procures.
Que ce mémoire soit le témoignage de ma sincère gratitude.

**À tous mes collègues de MED POINT DZ,**
Vous m'avez soutenu et veillé à mon succès pendant ces années d'étude loin de vous.
J'ai pour vous l'estime et l'admiration qu'imposent vos grandes qualités humaines.
Veuillez trouver dans ce travail l'expression de mon profond respect.

Enfin, merci à tous mes proches et amis, pour leur soutien et leurs encouragements…

# Résumé

Le domaine médical dispose aujourd'hui d'un très grand volume de documents électroniques permettant ainsi la recherche d'une information médicale quelconque. Cependant, l'exploitation de cette grande quantité de données rend la recherche d'une information précise complexe et coûteuse en termes de temps. Cette difficulté a motivé le développement de nouveaux outils de recherche adaptés, comme les systèmes de question-réponse. En effet, ce type de système permet à un utilisateur de poser une question en langage naturel et de retourner une réponse précise à sa requête au lieu d'un ensemble de documents jugés pertinents, comme c'est le cas des moteurs de recherche. Les questions soumises à un système de question-réponse portent généralement sur un type d'objet ou sur une relation entre objets. Dans le cas d'une question telle que « Qui a découvert l'Amérique ? » par exemple, l'objet de la question est une personne. Dans des domaines plus spécifiques, tel que le domaine médical, les types rencontrés sont eux-mêmes plus spécifiques. La question « Comment rechercher l'hématurie ? » appelle ainsi une réponse de type examen médical.

L'objectif de ce travail est de mettre en place un système de question-réponse pour des médecins généralistes portant sur les bonnes pratiques médicales. Ce système permettra au médecin de consulter une base de connaissances lorsqu'il se trouve en consultation avec un patient. Ainsi, dans ce travail, nous présentons une stratégie de recherche adaptée au domaine médical. Plus précisément, nous exposerons une méthode pour l'analyse des questions médicales et l'approche adoptée pour trouver une réponse à une question posée. Cette approche consiste à rechercher en premier lieu une réponse dans une ontologie médicale construite à partir de ressources sémantiques disponibles pour la spécialité. Si la réponse n'est pas trouvée, le système applique des patrons linguistiques appris automatiquement pour repérer la réponse recherchée dans une collection de documents candidats. L'intérêt de notre approche a été illustré au travers du système de question-réponse « Esculape » qui a fait l'objet d'une évaluation montrant que la prise en compte explicite de connaissances médicales permet d'améliorer les résultats des différents modules du processus de traitement.

Mots-clés : systèmes de question-réponse, domaine médical, ontologie, patrons linguistiques.

# Abstract

The medical domain has currently a very high volume of electronic documents facilitating the search of any medical information. However, the exploitation of this large quantity of data makes the search of specific information complex and time consuming. This difficulty has prompted the development of new adapted research tools, as question-answering systems. Indeed, this type of system allows a user to ask a question in natural language and send a specific answer to its request instead of a set of documents deemed pertinent, as is the case with search engines. The questions submitted to a question-answering system concern generally a type of object or a relationship between objects. In the case of a question such as “Who discovered America?” the object of question is a person. In more specific areas, such as the medical domain, the types are themselves more specific. The question “How to Search the hematuria?” waiting for an answer type medical examination.

This dissertation studies the development of a question-answering system for physicians on good medical practices. This system will allow the doctor to consult a knowledge base when he is in consultation with a patient. Thus, we present an adapted research strategy to medical domain. Specifically, we will present a method for analyzing medical questions and the approach to find an answer to a submitted question. This approach consists to find an answer first in a medical ontology built from semantic resources available for the domain. If the answer is not found, the system applies linguistic patterns learned automatically to identify the answer in a collection of documents. The interest of our approach has been illustrated through the question answering system “Esculape” which has been the subject of an evaluation showing that the incorporation of explicit medical knowledge can improves the results of the different modules of the treatment processes.

Keywords: question-answering systems, medical domain, ontology, linguistic patterns.

# Table des matières

# Liste des tableaux

# Liste des figures

# Liste des annexes

# Introduction

L'expansion constante du nombre de documents électroniques, notamment grâce à Internet, a rendu l'accès à l'information plus aisée et rapide. De nos jours, rechercher une information ou un document sur le Web est devenu une activité quotidienne et prépondérante pour les internautes. Cette explosion du nombre de documents s'accompagne d'un accroissement du nombre d'utilisateurs interrogeant les différents moteurs de recherche devenus très populaires tels que Google (http://www.google.com) et Yahoo! Search (http://www.yahoo.com). Selon les chiffres de la société de mesure d'audience Comscore Networks (http://www.comscore.com), le moteur de recherche Google a ainsi traité, en novembre 2006, 5,6 milliards de requêtes (+ 9,1% par rapport à novembre 2005).

Cependant, cette masse documentaire est devenue de plus en plus difficile à exploiter et à gérer. L'exploitation de cette grande quantité de données a rendu la recherche complexe et coûteuse en termes de temps. Désormais, l'utilisateur éprouve beaucoup de difficultés à trouver l'information correspondant à son besoin. Deux facteurs en sont essentiellement responsables : le nombre de documents retournés par les moteurs de recherche d'une part ; l'hétérogénéité des informations disponibles sur le Web d'autre part. De plus, parmi tous les documents retournés par les moteurs, la plupart d'entre eux ne sont pas pertinents. De ce fait, un nouveau besoin a émergé : les futurs systèmes de recherche d'information doivent pouvoir répondre, en un minimum de temps, à des besoins plus précis que les systèmes actuels pour mieux satisfaire les utilisateurs.

Les systèmes de Question/Réponse (Q/R) sont une extension des systèmes de recherche documentaire allant dans ce sens. Ce type de système permet à un utilisateur de poser une question en langage naturel et de retourner une réponse à cette question au lieu d'un ensemble de documents jugés pertinents, comme c'est le cas des moteurs de recherche. En effet, face à une question donnée, les moteurs de recherche renvoient tous les documents jugés pertinents par rapport à la question, et c'est à l'utilisateur que revient la tâche d'explorer ces documents afin de trouver la réponse à sa question. Répondre à des questions précises requiert une analyse plus en profondeur des documents sélectionnés afin d'en extraire l'information recherchée.

De ce fait, les systèmes de question-réponse se distinguent, par rapport aux autres systèmes de recherche d'information, par la complexité de leur architecture. Cette dernière repose sur un enchaînement de plusieurs traitements incluant des modules de recherche documentaire et de

traitement automatique de la langue. L'architecture d'un système classique conduit à distinguer trois phases principales dans le processus de recherche. Une première phase consiste à analyser la question posée par l'utilisateur syntaxiquement et sémantiquement. Cette étape permet de déterminer le type de la question suivant une classification définie au préalable, de détecter le type de la réponse attendue, en particulier lorsqu'il s'agit d'une entité nommée[1], et de mettre en évidence les termes de la question les plus importants du point de vue de la recherche d'une réponse. Cette phase est suivie par une étape de recherche de documents réalisée en interrogeant un ou plusieurs moteurs de recherche, étape qui débouche, en faisant appel à des traitements plus élaborés, à la sélection des passages susceptibles de contenir une réponse. Enfin, la dernière étape consiste à extraire des réponses candidates de ces passages en s'appuyant sur les informations issues de l'analyse de la question et la façon dont elles se retrouvent au niveau des passages. Il est à noter que certains systèmes de question-réponse, plus évolués, comportent une ultime couche leur permettant, par exemple en sollicitant les moteurs de recherche du Web avec comme mots-clés la réponse et les mots importants de la question, de justifier et de valider les réponses extraites.

La plupart des systèmes de question-réponse actuels affichent une certaine pertinence sur les questions factuelles, c'est-à-dire les questions portant sur un fait précis et dont la réponse attendue est une entité nommée. À titre d'exemple, pour la question « Qui a écrit Germinal ? » le type de la réponse attendue est « personne ». Ce type de questions est généralement plus facile à traiter car les entités nommées sont facilement repérables dans les textes, au contraire d'autres questions, classées non factuelles, pour lesquelles les réponses sont moins directement identifiables dans les textes.

Les questions soumises à un système de question-réponse portent généralement sur un type d'objet ou sur une relation entre objets. Dans le cas d'une question telle que « Qui a découvert l'Amérique ? » par exemple, l'objet de la question est une « personne ». Dans des domaines plus spécifiques, tel que le domaine médical, les types rencontrés sont eux-mêmes plus spécifiques. La question « Comment rechercher l'hématurie ? » appelle ainsi une réponse de type « examen médical ».

[1] Les noms propres désignant les noms de personnes, lieux, organisations, etc.

L'objectif de ce travail est de mettre en place un système de question-réponse pour des médecins généralistes sur les bonnes pratiques médicales. Le but est de définir une stratégie de recherche adaptée au domaine médical. Ce système permettra aux professionnels de la santé de consulter une base de connaissances lorsqu'ils se trouvent en consultation avec un patient, ce qui impose une grande efficacité. Le système doit ainsi pouvoir trouver la réponse à la question posée en un nombre minimum de requêtes. En outre, comme toutes les réponses n'apparaîtront pas explicitement dans les documents, la prise en compte par le système d'un niveau minimal de connaissances médicales est indispensable pour pouvoir réaliser certaines inférences.

Depuis plusieurs années, grâce à l'émergence des nouvelles technologies de l'information et de la communication, l'information médicale est devenue de plus en plus disponible et accessible. Le domaine médical dispose aujourd'hui d'une grande quantité de documents électroniques et de multiples ressources linguistiques et terminologiques. Toutefois, ce vaste domaine présente certaines particularités. Il est caractérisé par la richesse et la complexité de son vocabulaire spécialisé. Cette dynamique contribue largement à la fréquence d'accès à l'information médicale et à la nécessité de la mise à jour de cette dernière.

La disponibilité de ces bases documentaires médicales, bien que contenant l'information, ne garantit pas la qualité de cette dernière. C'est un souci majeur dans un domaine spécialisé comme la médecine où la précision et la validité des informations recherchées sont des critères importants. De ce fait, le recours à des bases de connaissances médicales certifiées, comme les thésaurus, s'impose. En effet, ces bases de connaissances peuvent aider les systèmes de recherche d'information à trouver l'information souhaitée. En pratique, il existe plusieurs ressources sémantiques conçues explicitement pour le domaine médical. Les plus notables de ces ressources, souvent accessibles sur le Web, comptent le thésaurus[2] MeSH (*Medical Subject Heading*) (cf. Section 2.2.1), utilisé principalement pour l'indexation des documents médicaux, l'UMLS (*Unified Medical Language System*) (cf. Section 2.2.5) (Lindberg et al., 1993), qui centralise plus d'une centaine de thésaurus de différentes langues ou encore ORPHANET (cf. Section 2.2.4), qui répertorie tous les noms de maladies rares et leur définitions.

[2] Un thésaurus est une sorte de dictionnaire hiérarchisé, un vocabulaire normalisé sur la base de termes génériques et de termes spécifiques à un domaine. (source Wikipédia : http://fr.wikipedia.org/wiki/Thesaurus).

À l'image des réseaux lexicaux de même type mais plus généraux, tels que WordNet (Fellbaum, 1998), la plupart de ces ressources, très riches en terminologie, contiennent majoritairement des relations d'hyperonymie ou de synonymie et sont beaucoup moins riches en relations que l'on peut qualifier de syntagmatiques, comme celles caractérisant le fait qu'une maladie M peut être soignée par le traitement T ou que l'examen E permet de diagnostiquer la maladie M. Cependant, l'UMLS dispose d'un réseau sémantique constitué de 134 types sémantiques hiérarchisés par le lien « is-a » (Delbecque et al., 2005 ; McCray, 1989).

À la fois la contrainte d'une grande précision et l'existence d'importantes ressources font qu'un système de question-réponse dans le domaine médical doit être fortement guidé par les connaissances sur le domaine, que nous désignerons ici de façon générique sous le vocable d'ontologie[3]. Par leur degré important de structuration et la validation dont elles ont généralement fait l'objet, les ontologies offrent aux systèmes de question-réponse les moyens de remplir les contraintes de précision et de fiabilité que nous avons identifiées comme particulièrement importantes dans le contexte du domaine médical.

Dans le cadre de ce travail, notre démarche a consisté dans un premier temps à définir une ontologie du domaine de la médecine générale permettant de faire apparaître les entités caractérisant ce domaine ainsi que les relations existantes entre ces entités. Cette ontologie a été définie à la fois en sollicitant directement des médecins et par l'analyse des questions typiquement posées par des médecins généralistes (Ely et al., 1999 ; Ely et al., 2000). Notre étude s'est plus spécifiquement centrée sur un sous-ensemble représentatif de cette ontologie, défini autour des cinq entités suivantes : Maladie, Traitement, Examen, Médicament et Symptôme. Cette restriction n'est cependant pas limitative quant à l'approche développée pour mettre en œuvre un système de question-réponse permettant de répondre aux questions auxquelles sont confrontés quotidiennement les professionnels de la santé.

Dans une seconde étape, notre intérêt s'est focalisé sur la construction d'une base de connaissances médicales portant sur des relations plus spécifiques du domaine médical telles

[3] Une ontologie est une hiérarchie conceptuelle arborescente, fondée sur une structure terminologique et basée sur des principes linguistiques. Cette terminologie représente une organisation des connaissances propre à un domaine spécifique et à une tâche particulière dans ce domaine (Malaisé, 2005, page xi).

que la relation « Traite » entre une maladie et un traitement, à partir des documents électroniques médicaux disponibles sur Internet. Cette étape peut également être vue comme le peuplement de l'ontologie du domaine médical que nous avons définie. Elle commence par la reconnaissance des concepts du domaine dans les textes, réalisée dans le cas présent par l'application de règles de reconnaissance d'entités nommées. Ces règles, écrites manuellement, s'appuient sur des ressources obtenues à partir du Web mais permettent aussi de reconnaître de nouvelles entités non présentes dans la base de connaissances. Cette dernière s'en trouve ainsi améliorée et complétée au fur et à mesure. Le second aspect de ce processus de peuplement concerne les relations. À la différence des autres ressources, notre but a été de constituer une base de connaissances contenant des relations entre les types médicaux retenus pour notre étude de nature surtout syntagmatique, c'est-à-dire portant sur des relations sémantiques différentes des relations hiérarchiques.

Notre étude s'est ensuite portée sur l'une des étapes les plus importantes et déterminantes de la chaîne de traitement d'un système de question-réponse, en l'occurrence l'analyse de la question. Cette procédure consiste à classer la question et à déterminer le type de la réponse attendue (entité nommée ou autre), ce qui détermine ensuite la stratégie de recherche adoptée pour trouver une réponse dans un passage de document. De plus, outre le type de la réponse, cette étape permet de repérer le ou les entités nommées médicales présentes dans la question et éventuellement la relation entre l'objet de la question et l'objet de la réponse attendue. Parmi les relations auxquelles nous nous sommes attaché, on note : la relation « Traite » entre l'entité Maladie et l'entité Traitement, la relation « Soigne » entre Maladie et Médicament, la relation « Détecte » entre Maladie et Examen et enfin la relation « Signe » entre Maladie et Symptôme. À noter que, bien qu'il soit possible de considérer le concept « Médicament » comme un traitement, nous l'avons traité indépendamment du concept « Traitement » puisqu'il représente une classe sémantique importante dans une consultation de médecine générale. Pour réaliser cette analyse, nous avons adopté le même principe que pour l'identification des entités nommées, c'est-à-dire la définition de règles de reconnaissance.

Enfin, la dernière partie de notre travail s'est concentrée sur l'extraction de la réponse dans les documents médicaux, ou plus exactement, sur la proposition d'une démarche à adopter pour trouver une réponse à une question posée. Pour cela, nous avons défini une méthode se fondant sur deux approches complémentaires. Une première approche repose sur la construction et l'enrichissement d'une base de connaissances du domaine ainsi que sur la

recherche des réponses directement dans cette base. La seconde approche, qui n'est utilisée que lorsque la première échoue ou n'a pas été mise en œuvre, consiste à rechercher des réponses dans une source de textes. Dans le contexte d'un domaine fortement structuré par des ontologies, ces deux approches se déclinent de la façon suivante :

- identification des concepts médicaux et extraction des relations sémantiques entre deux concepts différents, cette phase contribuant à la constitution et l'enrichissement de la base de connaissances ;
- identification des relations sous-jacentes aux questions et extraction des réponses sur la base de ces relations.

Les deux approches reposent sur les mêmes outils : identification des concepts de l'ontologie selon une vision « entités nommées » et utilisation de patrons [4] lexico-syntaxiques caractéristiques des relations de l'ontologie, appris automatiquement à partir d'exemples. Ces patrons servent à valider la présence d'une relation, ce qui permet dans le premier cas d'en acquérir de nouvelles et dans le second cas, de s'assurer que la relation dans laquelle se trouve impliquée la réponse candidate est compatible avec celle sous-tendant la question.

**Organisation de l'exposé**

Ce mémoire s'articule en six chapitres. Le premier chapitre présente un état de l'art des systèmes de recherche d'information en montrant l'évolution des moteurs de recherche vers des systèmes plus performants tels que les systèmes de question-réponse. Nous présentons ensuite l'architecture typique d'un système de question-réponse et détaillons les différents modules intervenant dans la chaîne de traitement, de l'analyse de la question jusqu'à l'extraction de la réponse. Dans une deuxième partie, nous donnons quelques exemples de systèmes de question-réponse et précisons les approches adoptées par chacun d'entre eux pour extraire les réponses. Enfin, dans la dernière partie de ce chapitre, nous exposons la problématique et les limites actuelles de ces systèmes dans un domaine restreint, et plus particulièrement dans le domaine médical.

[4] Dans le cadre de ce travail, un patron représente une formule linguistique qui reflète une relation sémantique entre deux termes.

Dans le chapitre 2, nous nous intéressons à la présentation de quelques ressources terminologiques existantes dans le domaine médical, telles que le MeSH ou l'UMLS. La grande majorité de ces ressources contiennent essentiellement des relations sémantiques de type paradigmatique (comme l'hyperonymie). Elles manquent en revanche de relations de type syntagmatique, c'est-à-dire des relations plus spécialisées comme « X est un traitement de Y » ou encore « Y est un symptôme de X ». Ce constat nous a amené à proposer une ontologie du domaine regroupant des concepts médicaux et les relations sémantiques qui les unissent.

Le chapitre 3 porte sur le peuplement à partir de textes de l'ontologie définie. Pour ce faire, dans une première partie, nous nous intéressons à l'identification des concepts médicaux. Nous exposons plus particulièrement comment les concepts retenus pour notre étude sont reconnus dans les textes en utilisant des règles de reconnaissance d'entités nommées écrites manuellement et une ressource sémantique construite à partir de bases de connaissances existantes du domaine. Dans une deuxième partie, nous abordons l'extraction des relations sémantiques entre les concepts médicaux en évoquant dans un premier temps quelques travaux sur l'extraction de relations sémantiques, en particulier à base de patrons. Nous détaillons ensuite notre méthode d'acquisition des patrons lexico-syntaxiques et l'application de ces derniers pour identifier de nouvelles relations. Enfin, dans la dernière partie, nous présentons les résultats d'évaluations menées à la fois pour l'identification des concepts sémantiques et l'extraction de relations.

Le chapitre 4 est dédié exclusivement à la présentation du système de question-réponse, développé initialement et sur lequel nous avons travaillé, Œdipe. Le but de ce chapitre est de décrire l'architecture du système Œdipe ainsi que les principes des différents modules qui composent cette architecture, en particulier le module d'analyse des questions. Ce système repose sur la combinaison de modules de recherche documentaire et de traitement automatique de la langue. Nous présentons à la fin de ce chapitre, l'analyseur linguistique LIMA (LIc2m[5] Multilingual Analyzer) (Besançon et al., 2004), qui représente une des briques de base du système Œdipe.

[5] Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue.

Le chapitre 5 présente quant à lui le système Esculape, l'extension du système Œdipe que nous avons développée pour le domaine médical. Cette extension se caractérise par l'exploitation d'une ontologie afin de répondre à des questions portant sur le domaine de la médecine générale. Pour cela, nous commençons par présenter la méthode utilisée pour la classification et l'analyse des questions médicales. Cette étape permet, à partir d'une question, de déterminer le type de la réponse attendue (entité nommée médicale ou autre), l'objet important de la question (focus) et la relation sémantique sous-jacente à la question. De cette phase résulte une représentation de la question sous l'angle du schéma [concept]-(relation)-[concept]. Dans un second temps, nous décrivons la méthode adoptée pour l'apprentissage des patrons d'extraction de réponses. Puis, nous montrons comment ces patrons linguistiques sont exploités et appliqués pour extraire les réponses à partir de textes. Pour finir, dans la troisième et dernière partie, nous exposons l'évaluation des deux méthodes abordées dans ce chapitre, *i.e.* l'analyse des questions et l'extraction des réponses.

Ce manuscrit s'achève par le chapitre 6, qui présente les résultats des évaluations pour le français relatives aux systèmes Œdipe et Esculape. Ces évaluations ont été réalisées à la fois sur les corpus de la campagne d'évaluation CLEF QA (http://www.clef-campaign.org/) et sur les corpus de la tâche générale de la campagne EQueR (http://www.technolangue.net) dans le cas d'Œdipe, et sur les corpus de la tâche médicale d'EQueR pour Esculape.

# Premier chapitre
# Systèmes de question/réponse : problématique et état de l’art

# 1. Systèmes de question-réponse : problématique et état de l'art

*Ce chapitre dresse un état de l'art du domaine des systèmes de question-réponse, le domaine de notre travail. Il commence par décrire l'émergence de ce domaine et son évolution pour ensuite aborder l'architecture typique d'un système de question-réponse et présenter les différents modules qui le composent. Il présente aussi quelques exemples de systèmes existants pour exposer les différentes stratégies adoptées pour trouver les réponses attendues. Enfin, il expose les limites actuelles de ce type de systèmes et les difficultés rencontrées en domaine restreint, notamment le domaine médical, le domaine de notre étude.*

## 1.1 Introduction

La quantité de documents électroniques mise à disposition, notamment grâce aux réseaux informatiques, a largement modifié la notion de recherche d'information. Les utilisateurs ont en effet un accès de plus en plus direct à l'information. Cependant, pour accéder plus facilement à une information pertinente, des systèmes de recherche d'information se révèlent incontournables. Bien que les moteurs de recherche constituent une solution efficace pour trouver des documents correspondant à une requête utilisateur, ils s'avèrent moins performants concernant la recherche d'une donnée précise. De ce fait, il est primordial de faire appel à des systèmes plus élaborés capables de retourner une information fiable à un besoin d'information précis. C'est l'ambition des systèmes de question-réponse.

Les systèmes de question-réponse peuvent se définir comme étant des systèmes de recherche d'information évolués qui permettent de retourner une réponse précise, ou un passage contenant la réponse, à une requête utilisateur, au contraire d'un moteur de recherche qui renvoie un ensemble de documents jugés pertinents. Ils offrent la possibilité aux utilisateurs de poser une question en langage naturel sans aucune restriction sur le vocabulaire. La question est analysée et traitée afin d'extraire automatiquement, à partir d'une base documentaire, une réponse directe à la question posée. Cette extraction, à la différence des moteurs de recherche, ne nécessite pas d'intervention manuelle.

La majorité des systèmes de question-réponse actuels affichent une certaine pertinence sur les questions factuelles, c'est-à-dire les questions dont la réponse attendue est une entité nommée. Toutefois, de nos jours, les systèmes ont tendance à se focaliser sur le traitement d'autres types de questions plus complexes, à savoir, les questions non factuelles, dont les réponses ne sont généralement pas aussi évidentes à trouver dans les corpus. Ce type de questions nécessite une analyse en profondeur de la question afin d'en extraire tous les éléments indispensables pouvant intervenir dans le processus de recherche. Pour ce faire, les systèmes de question-réponse utilisent différentes techniques pour améliorer l'analyse des questions comme les outils issus du traitement automatique des langues. L'idée consiste à déterminer non seulement le type de la réponse recherchée, mais aussi les entités nommées présentes et l'objet sur lequel porte la question. Par ailleurs, pour étendre leurs performances, les systèmes ont recourt à des ressources sémantiques, structurées et/ou semi-structurées, éventuellement extraites du Web. Cette utilisation de bases de connaissances existantes telles que le réseau lexico-sémantique de WordNet[6] (Harabagiu et al., 1999 ; Plamondon et al., 2002) ou encore les ontologies d'un domaine précis (Vargas-Vera et al., 2004 ; Lopez Garcia et al., 2004) dans le cas d'un système de question-réponse en domaine restreint, permet aux systèmes d'augmenter la précision des réponses proposées.

Dans ce chapitre, nous exposerons le fonctionnement des systèmes de question-réponse et détaillerons par la suite les différents modules intervenant dans la chaîne de traitement, soit de l'analyse de la question jusqu'à l'élaboration de la réponse souhaitée en passant par la recherche des documents candidats. Nous présenterons aussi quelques systèmes existants dans le but de montrer les différentes techniques utilisées et les démarches adoptées pour rechercher les réponses. Enfin, nous terminerons ce chapitre en exposant les lacunes de ces systèmes de question-réponse en domaine restreint, en particulier dans le domaine médical.

## 1.2 Du moteur de recherche au système de question-réponse

L'émergence des nouvelles technologies de l'information et de la communication a largement contribué à la naissance d'un nouveau besoin qui est « la recherche d'information ». Le

[6] Base de données lexicale organisée en ensemble de synonymes reliés entre eux par des relations sémantiques.

domaine de la recherche d'information, plus exactement de l'accès à l'information, suscite depuis plusieurs années un intérêt particulier. Cet intérêt est motivé en premier lieu par le besoin de définir des stratégies appropriées et performantes afin d'exploiter et de gérer l'extraordinaire base documentaire disponible sur le Web. En effet, de nos jours, trouver une information précise reste indéniablement difficile à réaliser, notamment en raison de la structure des documents électroniques et de l'hétérogénéité des informations disponibles sur la Toile. La recherche d'information consiste donc à donner à un individu la possibilité de consulter une base documentaire et à lui retourner les éléments correspondant à sa recherche. De ce besoin, ont émergé les systèmes de recherche d'information qui représentent un intermédiaire permettant aux utilisateurs d'interroger des ressources documentaires. Le but de ces systèmes, appelés aussi moteurs de recherche, est de faire correspondre d'une façon intelligente les mots-clés exprimés dans la requête en langage naturel par l'utilisateur avec les documents existants dans la base de documents afin de ne lui fournir que les éléments susceptibles de contenir l'information recherchée. Cet appariement consiste généralement à effectuer une comparaison entre les mots de la requête et les documents.

Cependant, la principale difficulté à laquelle sont confrontés les systèmes de recherche d'information traditionnels concerne l'interprétation et la compréhension de la requête formulée par un utilisateur. La polysémie, c'est-à-dire le fait qu'un terme de la requête peut être interprété de différentes manières au niveau sémantique, est un exemple de ces difficultés. Un autre cas de difficulté, rencontrée par les systèmes, concerne la présence des éléments clés de la requête dans des documents pertinents sous une forme différente de celle employée dans la requête initiale mais sémantiquement liée à la forme originelle. Ces phénomènes ont un impact négatif sur la performance des systèmes de recherche d'information entraînant la récupération de documents non pertinents ou étiquetant des documents comme non pertinents bien que porteurs de l'information désirée. De ce fait, l'utilisation des outils de traitement automatique des langues s'avère indispensable pour une meilleure compréhension de la question afin de permettre aux systèmes d'être plus efficaces dans la recherche documentaire (Jacquemin et al., 2000b).

Les moteurs de recherche ont été surtout développés pour retourner une liste de documents jugés pertinents organisée par ordre de pertinence par rapport au thème de la requête exprimée par l'utilisateur comme une suite de mots-clés. Mais c'est à l'utilisateur que revient la tâche

de parcourir l'ensemble des documents retournés pour rechercher l'information désirée. Cette tâche peut s'avérer fastidieuse et engendrer une perte de temps, surtout si le document contenant l'information recherchée n'apparaît pas en tête de la liste, ce qui incite parfois l'usager à modifier sa requête ou rajouter des mots-clés à celle-ci afin d'augmenter ses chances de trouver un document pertinent. De ce fait, ces systèmes se révèlent moins performants pour répondre aux attentes des utilisateurs désirant rechercher des informations précises, plus exactement des requêtes portant sur un fait particulier, comme répondre à des questions. C'est en revanche l'objectif principal des systèmes de question-réponse.

Les systèmes de question-réponse constituent une avancée importante des systèmes de recherche d'information. Ils sont dotés d'une architecture complexe et s'appuient sur des techniques de recherche plus élaborées. Leur domaine de recherche se situe à l'intersection de deux domaines de recherche, à savoir la recherche d'information et le traitement automatique des langues. Les premiers systèmes de question-réponse sont apparus dès les années 60 en introduisant une approche fondée sur le dialogue Homme-Machine. Le but de ces systèmes consistait exclusivement à consulter des bases de données d'un domaine spécifique. L'approche utilisée dans ces systèmes reposait sur la transformation d'une question posée en langage naturel en une requête afin de récupérer une réponse courte à partir de la base de données interrogée. Parmi les systèmes les plus connus adoptant ce procédé, on note les deux systèmes BASEBALL (Green et al., 1961) et LUNAR (Woods, 1973).

## 1.3 Architecture d'un système de question-réponse

La notion de système de question-réponse fut introduite à la fin des années 70 avec le système QUALM (QUestion Answering Mechanism) développé par Lehnert en 1977 (Lehnert, 1977). La conception de ce système a largement contribué au développement des systèmes de question-réponse. Le processus de recherche débute par la catégorisation de la question posée ; le but est ici de délimiter le contexte de la question afin de déterminer la stratégie de recherche à employer pour extraire la réponse. Cette dernière est extraite en appliquant des heuristiques. Cependant, il a fallu attendre la première campagne d'évaluation pour les systèmes de question-réponse, à savoir la piste Question Answering de TREC (Text Retrieval and Evaluation Conference : http://trec.nist.gov) en 1999 (Voorhees, 1999), pour constater

l'intérêt de la communauté de la recherche d'information pour ce domaine et voir émerger, depuis lors, un grand nombre de systèmes.

Bien que les techniques diffèrent d'un système à l'autre, la plupart des systèmes de question-réponse reposent sur une architecture classiquement fondée sur quatre modules complémentaires que nous détaillerons dans la suite de ce chapitre (voir Figure 1.1). Le premier de ces quatre modules concerne l'analyse de la question. Il vise plus précisément à extraire d'une question les informations permettant de repérer la réponse dans les documents comme le type de la question posée, l'objet sur lequel porte cette question, appelé aussi «focus», le type de la réponse attendue et les mots importants de la question. Le deuxième module a quant à lui pour objectif de sélectionner un ensemble de documents ou d'extraits de documents facilitant ainsi les traitements de la suite de la chaîne. Le troisième module se charge d'analyser les documents sélectionnés et d'en extraire les passages candidats susceptibles de contenir la réponse. Enfin, le quatrième et dernier module permet de rechercher dans les passages sélectionnés la réponse qui, selon la question et la particularité des systèmes, se présente sous la forme d'une entité nommée ou d'un passage contenant la réponse. Ces quatre modules s'appuient principalement sur des techniques de traitement automatique de la langue et de recherche d'information. Les outils de recherche d'information servent plus particulièrement à la recherche des documents et des passages les plus pertinents, tandis que les techniques de traitement de la langue permettent d'améliorer les procédures d'extraction d'information en offrant la possibilité d'effectuer une analyse plus en profondeur de la question et des documents.

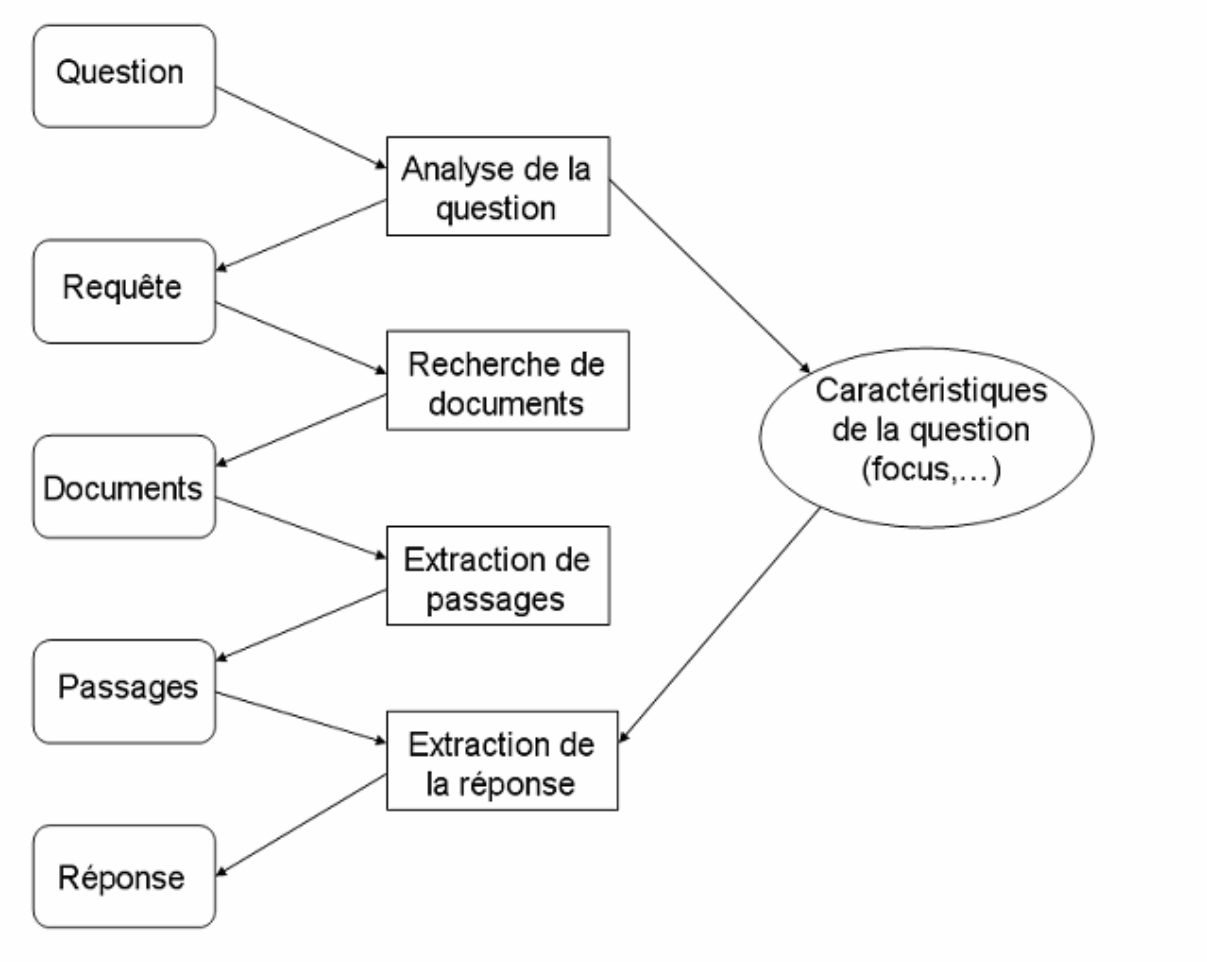


Figure 1.1 Architecture d'un système de question-réponse

### 1.3.1 Analyse des questions

L'analyse de la question est une étape importante dans la chaîne de traitement d'un système de question-réponse (Mendes et al., 2004), outre le fait qu'elle est la première. En effet, il est primordial pour un système d'analyser une question aussi soigneusement que possible car cette analyse conditionne la stratégie de recherche à appliquer.

L'objectif principal de l'analyse de la question est à la fois de déterminer ce que le système doit chercher et de mettre en évidence les éléments informatifs permettant de sélectionner une réponse. Ainsi, l'analyse de la question doit déterminer :

- **le typage de la question :** il permet d'attribuer à la question une catégorie selon une classification prédéfinie (Définition, Factuelle, Booléenne[7]). Par exemple la question suivante : « *Quelle est la définition du paludisme ?* » est une question définitoire, tandis que la question « *Citer sept pays membres de l'Union européenne ?* » se verra attribuer la catégorie factuelle de type liste ;

- **les entités nommées de la question :** il s'agit de repérer toutes les entités nommées présentes dans la question. Cela revient à repérer par exemple l'entité personne « *Pablo Picasso* » dans la question « *Dans quelle ville est né Pablo Picasso ?* » ;

- **le type de la réponse attendue :** ce type est généralement formalisé sous la forme d'un type d'entité nommée (personne, date, lieu, …) ou d'un type d'entité plus élargi (maladie, traitement, évènement, …). Ainsi, pour la question « *Qui a écrit Harry Potter ?* », le type de la réponse attendue est une entité nommée PERSONNE ; pour la question « Quel est le traitement de la cirrhose ? », le type attendu est l'entité TRAITEMENT. Ce type de questions est souvent plus facile à traiter que les questions portant sur des définitions ou des explications où le type sémantique de la réponse est plus complexe et moins facilement identifiable ;

- **le focus de la question :** il s'agit d'extraire l'objet sur lequel porte la question, c'est-à-dire un élément susceptible d'être présent dans le passage réponse. Pour la question « En quelle année est né Alexandre Pouchkine ? », le focus est ainsi Alexandre Pouchkine.

Parallèlement, les mots-clés présents dans la question sont extraits pour composer une requête d'interrogation permettant à un système de recherche documentaire de retourner un ensemble de documents jugés pertinents. Ces mots sont considérés comme des éléments importants ayant un rapport direct avec la réponse permettant ainsi de restreindre le contexte de la question. Par exemple, pour la question : « Combien d'oscars a reçu le film Titanic ? », les mots-clés à extraire sont : « oscars, film, Titanic » et la réponse à rechercher est une entité numérique de type quantité (en oscars).

[7] Questions attendant une réponse de type oui/non.

Afin de classer les questions, les systèmes de question-réponse utilisent des approches différentes mais s'appuyant généralement sur des critères linguistiques. Ils utilisent pour ce faire différents outils de traitement automatique des langues allant de l'étiqueteur morpho-syntaxique jusqu'à l'analyseur syntaxique (Hermjakob, 2001 ; Graesser et al., 1992) en passant par le reconnaisseur d'entités nommées.

Certains systèmes de question-réponse effectuent une analyse plus en profondeur des questions allant jusqu'à une véritable analyse sémantique et une reformulation. Il s'agit dans ce cas d'extraire la ou les relations sémantiques sous-jacentes à la question et d'en construire une représentation sémantique, à la manière du système JAVELIN (Nyberg et al., 2002). Enfin, il est important de souligner que dans le cas de la reformulation d'une question ou de l'extension d'une requête, la plupart des systèmes se fondent sur une approche utilisant des connaissances sémantiques comme le réseau WordNet pour obtenir les différentes variations sémantiques des termes constituant la question.

### 1.3.2 Recherche des documents

Dans un système de question-réponse, la recherche des documents se fait par l'interrogation d'un système de recherche d'information. Cette étape se révèle particulièrement capitale et complémentaire à l'analyse de la question pour la recherche de la bonne réponse car les systèmes de question-réponse ne peuvent trouver une réponse à une question que si elle est présente dans les documents sélectionnés. Cette tâche consiste donc à interroger un moteur de recherche classique pour récupérer une sélection de documents ou de passages restreints potentiellement porteurs de la réponse. Pour ce faire, les systèmes de question-réponse se reposent sur l'analyse de la question qui permet de générer une requête, souvent de nature booléenne, dédiée à l'interrogation d'une base textuelle. Dans un contexte des systèmes de question-réponse en domaine restreint, la recherche documentaire se fait sur un ensemble généralement limité de documents alors que pour les systèmes en domaine ouvert, la recherche d'information s'effectue sur une grande collection de textes couvrant presque tous les domaines tels que les sources de données existantes sur le Web. De plus, utiliser le Web comme source de connaissances permet aux systèmes de question-réponse de bénéficier de la redondance informationnelle (Lin, 2007), cependant, la fiabilité de ces informations est mise en cause.

La requête d'interrogation est constituée principalement des termes importants de la question tels que les noms, verbes et adjectifs. Elle permet à la fois de restreindre le contexte de la recherche d'information et d'identifier les documents jugés pertinents par le moteur de recherche pour l'extraction de la réponse. Ces mêmes documents sont utilisés non seulement pour extraire la réponse recherchée mais aussi pour la justification de celle-ci. Cependant, l'exploitation d'un mot-clé d'une question ne permet pas nécessairement de repérer la réponse dans un document. En effet, la signification d'un mot peut être représentée ou interprétée de différentes manières. Aussi, grâce à l'apport de techniques du traitement automatique de la langue, les systèmes de question-réponse évolués effectuent des transformations de la requête. Ces transformations consistent essentiellement à étendre la requête par l'ajout de termes en relation avec les mots-clés constituant la requête. L'idée est d'orienter le comportement des systèmes de recherche d'information afin de sélectionner non pas des documents qui traitent du sujet de la question mais plutôt des documents porteurs de la réponse. Il est ainsi possible de récupérer plus de documents pertinents contenant la réponse. Les termes ajoutés sont en pratique des mots proches des mots-clés de la question et entretiennent avec eux des relations sémantiques telles que les relations d'hyperonymie ou de synonymie. L'expansion de requête se base donc sur l'enrichissement de la requête initiale par des variations sémantiques (comme les synonymes, hyperonymes…) des termes qui la composent (Harabagiu et al., 2001), ou encore en exploitant les liens sémantiques entre les noms et les verbes, comme dans (Claveau et al., 2004). Pour extraire les différentes variantes linguistiques des mots, les systèmes utilisent des ressources lexicales et des bases de connaissances sémantiques spécialisées comme dans (Voorhees, 1994) qui exploite le thésaurus WordNet.

### 1.3.3 Analyse des documents candidats

Les techniques avancées de traitement automatique de la langue, souvent utilisées pour l'extraction de réponse, demeurent trop lourdes pour être utilisées sur une grande quantité de textes. C'est ce qui amène les systèmes de question-réponse à faire appel aux systèmes de recherche d'information pour restreindre le nombre de documents à analyser. Les documents retournés par le moteur de recherche sont généralement en relation directe avec le thème de la question et sont censés apporter la réponse à la question initiale. Dans la même perspective et en vue de réduire le temps d'extraction des réponses, les documents candidats sont ensuite

classés par pertinence. Cette tâche consiste à ordonner les documents selon un poids calculé sur la base de la présence des mots-clés de la question dans les textes.

L'analyse des documents candidats a pour objectif principal de parcourir les documents sélectionnés pour rechercher les meilleurs passages de textes ou les phrases correspondant à la réponse recherchée en s'appuyant principalement sur les éléments issus de l'analyse de la question. La stratégie pour ce faire consiste le plus souvent à extraire des documents les passages ou les phrases comportant au moins un mot de la question ou une entité du même type sémantique que la réponse attendue. De même que pour la sélection des documents candidats, ces passages ou ces phrases sont hiérarchisés par ordre de pertinence. Leur choix est réalisé par des approches différentes spécifiques à chaque système. La méthode la plus utilisée consiste à repérer les mots de la question dans les documents pour n'extraire que les passages ou les phrases ayant le plus de mots en commun avec la question (Gillard et al., 2005). Un certain nombre de systèmes adoptent une stratégie plus avancée fondée sur le calcul d'une mesure de proximité entre les mots de la question dans les passages (Nyberg et al., 2003), c'est-à-dire qu'ils font l'hypothèse que dans les documents censés contenir une réponse, les termes de la question et le type la réponse attendue sont proches. D'autres approches, améliorant la performance des systèmes de question-réponse dans la sélection des passages pertinents ont été proposées et appliquées comme celle de (Gillard et al., 2006) qui repose sur la densité des mots de la question dans les passages. Le calcul de cette densité est tout d'abord déterminé par l'extraction des objets de la question : les lemmes des mots, les types d'entités nommées présentes et le type de la réponse à rechercher. Ensuite, pour chaque élément, une distance moyenne est calculée entre l'objet courant et les autres objets de la question. Cette distance est utilisée par la suite pour le calcul du score de densité afin d'identifier le passage le plus en relation avec la question, *i.e.* le passage censé contenir la réponse souhaitée. Pour réduire la perte d'information, le passage candidat est composé d'un bloc de trois phrases regroupant la phrase réponse complétée par la phrase précédente et la phrase suivante.

Parallèlement au découpage des documents sélectionnés en passages, les méthodes d'analyse des documents permettent de réaliser un enrichissement de chaque passage candidat. Parmi les enrichissements les plus fréquents, les entités nommées présentes dans la phrase sont identifiées et les variations terminologiques des mots de la question reconnues. La

reconnaissance des entités nommées consiste à extraire les différents types d'entités nommées que contient le passage, les plus communes étant les entités nommées de type MUC (Message Understanding Conferences) (Grishman et al., 1995) : les noms de personnes, d'organisations, les lieux, les unités de mesures ainsi que les dates. Cette tâche est effectuée en respectant une hiérarchie de classes et de sous-classes définie au préalable qui peut varier d'un système à un autre (voir Figure 1.2 pour un exemple d'une telle hiérarchie). Enfin, pour compléter cette analyse des passages, la plupart des systèmes de question-réponse ont recourt à des bases de connaissance leur permettant d'identifier les variantes lexicales des mots de la question dans les passages (Yang et al., 2002 ; Ferret et al., 2001a). À ce niveau, les systèmes font généralement intervenir des connaissances morphologiques et sémantiques existantes issues de dictionnaires électroniques ou des ressources lexicales plus évoluées telles que WordNet.

D'autres systèmes plus sophistiqués vont encore plus loin dans l'analyse en utilisant des méthodes spécifiques visant à désambiguïser le sens de certains termes présents dans les passages pouvant receler des indices nécessaires à l'extraction de la réponse recherchée. Par exemple (Crestan et al., 2004) ont développé un module spécialisé de résolution d'anaphores. Ce module n'est utilisé que dans leur système en anglais en raison d'une difficulté rencontrée pour le français. En effet, ce module connaît des difficultés pour distinguer les formes impersonnelles (le pronom « il » par exemple) dans les textes.

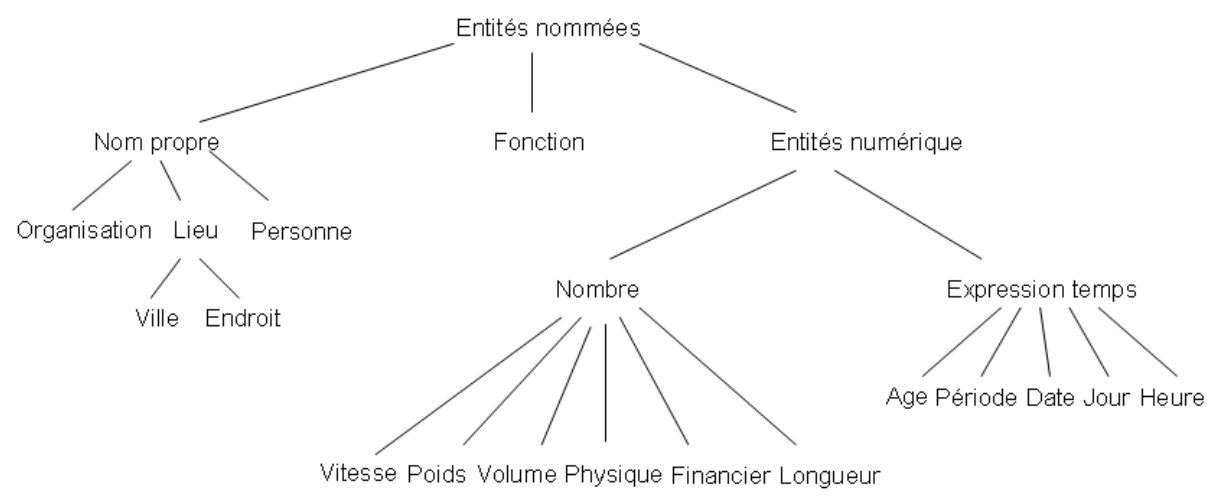


Figure 1.2 Exemple de la hiérarchie des entités nommées du système QALC (Ferret et al., 2001a)

### 1.3.4 Extraction des réponses

Le module d'extraction de réponses constitue le dernier maillon de la chaîne de traitement d'un système de question-réponse. Cette fonction symbolise la différence majeure d'un tel type de systèmes par rapport aux systèmes de recherche d'information traditionnels. Rechercher une réponse à une question revient à fouiller les passages candidats sélectionnés par l'analyse des documents choisis afin d'identifier et extraire le passage réponse correspondant à la question formulée. Cette notion de « passage réponse », qui caractérise la réponse supposée correcte retournée par le système, peut être présentée sous différentes formes suivant le système. Dans la majorité des systèmes de question-réponse, la réponse retournée est une liste de réponses organisée selon un indice de confiance ou bien leur fréquence d'apparition dans les documents candidats tandis que pour certains, la réponse retournée est une réponse unique courte ou un extrait d'un document contenant la bonne réponse avec son contexte.

La fonction d'extraction de réponses concentre l'intérêt principal des systèmes de question-réponse. Cette phase constitue l'un des points caractéristiques permettant d'individualiser les différents systèmes de question-réponse. En pratique, elle est le résultat d'un appariement réalisé entre la représentation de la question et les portions de textes sélectionnées à l'issue de l'analyse des documents candidats. La représentation d'une question peut prendre différentes formes et peut être plus ou moins riche en connaissances. (Monceaux et al., 2002) exploite par exemple les connaissances syntaxiques des mots de la question tandis que (Mendes et al,. 2004) va jusqu'à s'appuyer sur la transformation des éléments de la question en prédicats logiques. Le but de cette représentation est d'exploiter au maximum les contraintes syntaxiques et sémantiques des questions afin d'effectuer certaines inférences pour retrouver les réponses. Pour ce faire, les systèmes performants se fondent sur des outils élaborés de traitement automatique des langues tels que l'analyse sémantique, dont l'apport s'avère primordial pour réaliser une meilleure analyse des questions (Poibeau et al., 2003) et déterminer des stratégies de recherche adaptées.

La façon d'extraire les réponses est dépendante du type de la réponse attendue. Lorsqu'il s'agit d'une entité nommée, une approche commune est de repérer les entités correspondant au type sémantique de la réponse désirée dans les passages pertinents puis de les classer selon

leur fréquence d'apparition (Ferret et al., 2001a). Cette fréquence est généralement calculée sur l'ensemble des documents renvoyés par le moteur de recherche, ou parfois pour certains systèmes, elle peut même être étendue sur une grande quantité de documents comme le Web pour profiter de la redondance de l'information (Berthelin et al., 2003). Dans le cas où la question n'attend pas une entité nommée en réponse, les systèmes font appel à des motifs d'extraction prédéfinis (Soubbotin et al., 2002 ; Malaisé et al., 2005), appelés aussi patrons d'extraction. Ces patrons linguistiques exprimés sous la forme d'expressions régulières sont habituellement écrits manuellement mais sont parfois appris automatiquement *a priori* à partir de corpus de textes (Ravichandran et al., 2002).

Une autre technique permettant de sélectionner une réponse consiste à utiliser des sources de connaissances existantes (Katz et al., 2002). En effet, pour bien répondre à certains types de question, plus particulièrement aux questions portant sur des définitions, il est parfois très utile de disposer d'une ressource *a priori*. Cette dernière peut permettre à un système de question-réponse de trouver directement une réponse correcte[8]. L'utilisation de telles ressources offre la possibilité de vérifier et de valider les réponses extraites retournées par le système et ainsi permettre à ce dernier d'ordonner l'ensemble des réponses candidates. Pour certains systèmes, la validation de la réponse exploite des connaissances sémantiques appropriées afin de s'assurer que la réponse sélectionnée correspond au bon type d'information recherchée tandis que pour d'autres, elle repose sur la fréquence d'apparition de la réponse dans une base documentaire restreinte ou à partir du Web, comme dans (Berthelin et al., 2003).

Enfin, l'objectif des systèmes de question-réponse actuels et à venir va au-delà de l'identification de réponses (Burger et al., 2003). Plus explicitement une des ambitions futures des systèmes est de parvenir à leur justification. La grande majorité des systèmes renvoie des réponses avec les contextes dans lesquels elles ont été extraites et c'est à l'utilisateur que revient la tâche de vérifier la validité des réponses proposées. Cependant, un certain nombre de systèmes élaborés sont actuellement capables d'accomplir cette fonctionnalité automatiquement, à l'instar du système PowerAnswer (Harabagiu et al., 2005) (voir Section 1.4.5) qui repose sur un raisonneur logique, appelé COGEX (Moldovan et al., 2003a),

[8] Pour certains systèmes de question-réponse, l'encyclopédie Wikipédia (http://wikipedia.org) est utilisée comme une base de réponses possibles.

permettant d'associer la réponse trouvée à la question articulée. Une autre ambition pour les systèmes de question-réponse est de construire une réponse en réalisant la fusion de plusieurs réponses lorsque c'est nécessaire. En effet, pour trouver une réponse à certaines questions complexes, il est parfois indispensable d'effectuer des inférences entre des réponses candidates résultant de sources documentaires différentes afin de constituer la réponse exacte à renvoyer à l'utilisateur. Par exemple, pour la question « Quel footballeur brésilien a remporté le ballon d'or en 2002 ? », la réponse « Ronaldo » est retournée à partir de la fusion des éléments de réponse provenant de différents documents justifiant les éléments de la question initiale (voir Figure 1.3 ci-dessous).

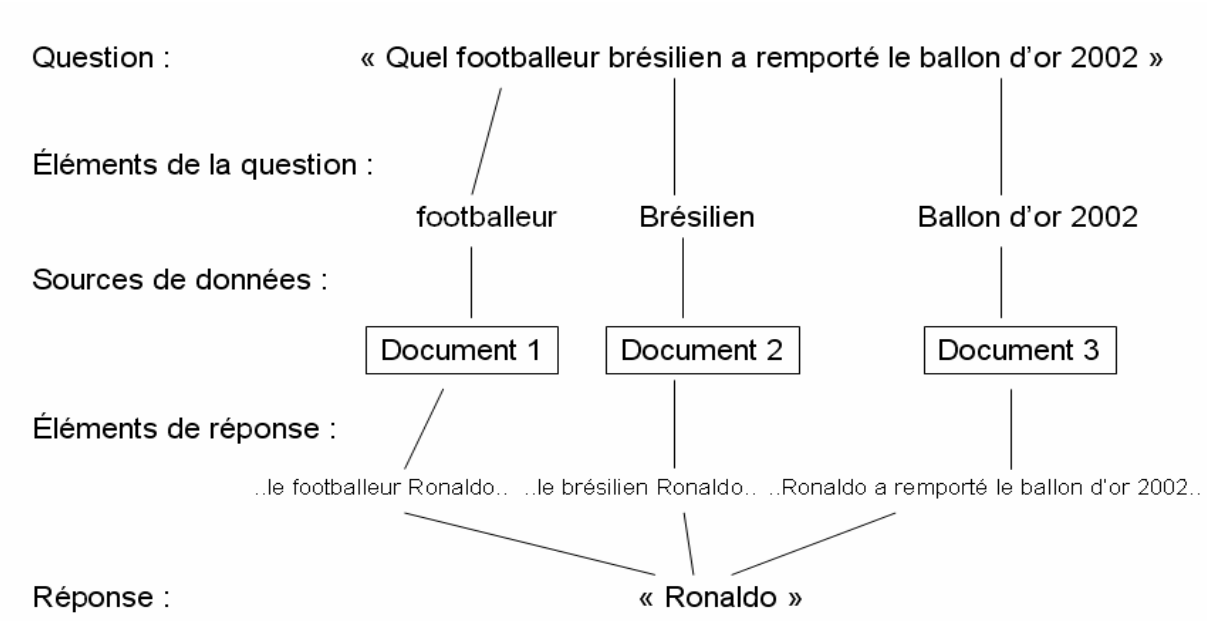


Figure 1.3 Exemple sur la fusion de réponses

## 1.4 Présentation de quelques systèmes de question-réponse

Les premiers systèmes de question-réponse ont émergé à partir des années 60 avec la naissance de systèmes permettant d'interroger des bases de données dans un domaine précis en langage naturel. Le système BASEBALL (Green et al., 1961) fut l'un des premiers dans cette voie, avec l'objectif de répondre à des questions correspondant aux résultats du championnat américain de baseball. Puis, d'autres systèmes de ce type ont vu le jour : le système LUNAR (Woods, 1973) pour répondre à des questions concernant la Lune, le système LIFER (Hendrix, 1977) pour produire des statistiques sur des employés et le système STUDENT (Winograd, 1973) pour répondre à des questions portant sur des problèmes

mathématiques. Ces systèmes offraient une interface en langage naturel pour pouvoir consulter directement des bases de données contenant des connaissances codées manuellement. Cette interrogation se faisait selon une approche visant à traduire la question posée en langage naturel en une requête d'interrogation de base de données. Cependant, ces systèmes affichaient une certaine limite en ne pouvant pas facilement s'étendre à d'autres domaines car s'appuyant sur des bases de connaissances dédiées uniquement à un domaine très précis.

Vinrent ensuite des systèmes de question-réponse reposant sur des techniques plus évoluées comme le système QUALM (Lehnert, 1978) proposé par Lehnert qui appliquait une approche fondée sur la compréhension de textes. Un autre système, SHRDLU (Winograd, 1972), permettait de gérer un dialogue entre un robot et un humain à propos d'un monde d'objets et mettait en œuvre une approche permettant au robot d'interpréter le texte soumis afin d'exécuter les instructions ordonnées par l'utilisateur. Le système UC (Unix Consultant) (Wilensky, 1982) quant à lui permettait de répondre à des questions concernant le système d'exploitation Unix.

Le développement du Web et l'amélioration considérable des outils de traitement automatique du langage naturel ont largement contribué à la possibilité de développer des systèmes de question-réponse ayant pour objectif de répondre à tout type de questions. Cependant, l'intérêt porté pour les systèmes de question-réponse n'a connu son plein essor qu'après l'apparition de la tâche « Question/Réponse » dans des conférences d'évaluation des systèmes de recherche d'information, et principalement la conférence TREC, qui fut la première à introduire cette tâche en 1999 pour les systèmes de question-réponse en anglais. Le but de la tâche Question/Réponse de la conférence TREC-8 (Voorhees, 1999) consistait à évaluer les différents systèmes participant dans leur capacité à trouver les réponses à une liste de questions en domaine ouvert dans un corpus de textes constitué par le NIST (National Institute of Standards and Technology). Cette évaluation, qui portait sur le jugement du passage réponse retourné, a permis de constater les avancées réalisées dans le domaine du Question/Réponse. Depuis, la tâche Question/Réponse fut introduite dans différentes compétitions dédiées aux systèmes de recherche d'information comme la campagne NTCIR (Test Collection for IR Systems) (http://research.nii.ac.jp/ntcir) depuis 2003. Cette évaluation s'intéresse à la recherche d'information avec un axe multilingue asiatique. Lancée en 2000, la

compétition CLEF (Cross Language Evaluation Forum) proposait une évaluation Question/Réponse dès sa quatrième édition (CLEF 2003). Elle est aussi axée sur l'aspect multilingue pour les langues européennes. En 2004, la première campagne d'évaluation des systèmes de question-réponse en français, la campagne EQueR (Ayache, 2006), a vu le jour. La campagne proposait deux tâches : une première dans le domaine général et une deuxième plus spécialisée qui concernait le domaine médical. Les systèmes candidats étaient évalués sur les bonnes réponses courtes trouvées ou sur des courts extraits censés contenir la réponse correcte. L'objectif de ces campagnes d'évaluation est principalement d'améliorer la performance des systèmes de question-réponse car elles fournissent un contexte d'application et d'évaluation pour ces derniers. Elles permettent également d'aborder plusieurs sous-tâches dans le domaine des questions-réponses par exemple : retourner des réponses à une suite de questions enchaînées sur le même thème ou répondre à des questions comprenant un contexte temporel. Les campagnes d'évaluation actuelles tendent à diversifier les sources textuelles au-delà d'une simple collection de textes journalistiques en considérant le Web ou des ressources textuelles plus structurées comme Wikipédia ; c'est le cas par exemple de la campagne CLEF-QA (www.clef-campaign.org).

Les systèmes de question-réponse peuvent être différenciés selon les stratégies de recherche employées. Dans ce qui suit, nous présentons quelques approches caractéristiques qui ont obtenu les meilleurs résultats dans les tâches de question-réponse lors des récentes campagnes d'évaluation TREC, CLEF et EQueR : les indices terminologiques (le système QALC), le Traitement Automatique des Langues (le système QRISTAL), statistiques (le système PIQUANT), l'interaction avec l'utilisateur (le système JAVELIN), le raisonnement logique (le système PowerAnswer), les inférences (le système WEBCOOP) et enfin les patrons d'extraction (le système d'InsightSoft).

### 1.4.1 Le système QALC

Le système QALC (Question Answering program of the Language and Cognition group) (Ferret et al., 2000 ; Ferret et al., 2001a) a été le premier système de question-réponse développé pour l'anglais au sein du LIMSI dans le cadre de la campagne d'évaluation TREC en 1999. Il a constitué la base des systèmes suivants ayant participé à d'autres campagnes d'évaluation comme EQueR 2004 et CLEF-QA. Ce système s'appuie sur un ensemble de

modules de traitement automatique des langues intervenant en aval d'un moteur de recherche opérant sur une vaste sélection de documents. Le système a été initialement conçu pour répondre à des questions factuelles portant sur n'importe quel domaine. Il est composé d'un module d'analyse des questions, de sélection des documents, de reconnaissance des entités nommées et d'extraction de réponses.

QALC effectue un premier traitement sur la question, réalisé par un analyseur syntaxique partiel dédié, qui permet de déterminer un certain nombre de caractéristiques de la question qui seront utilisées dans la suite de la chaîne de traitement, dont en particulier : le type de la réponse attendu, la catégorie de la question, les entités nommées de la question et le focus de la question. Après interrogation du moteur de recherche, une sélection de documents candidats est effectuée sur la base de la présence des termes de la question ou de leurs variantes dans les documents. Cette identification est réalisée par l'analyseur FASTR [9] (Jacquemin, 1999). Les documents sont ensuite découpés en phrases pour ne conserver que les phrases contenant au moins un mot de la question ou une variante d'un mot. Enfin, pour l'extraction de la réponse, deux stratégies différentes sont appliquées, en fonction du type attendu de la réponse. Si la question attend une entité nommée en réponse, le système choisira l'entité nommée la plus proche du barycentre des variantes des mots de la question, pondérées par leur poids FASTR. Sinon, des patrons d'extraction écrits manuellement sont appliqués. Ces patrons permettent d'étiqueter des constituants comme réponse s'ils sont entourés de constituants comprenant une caractéristique de la question, et séparés de ceux-ci par des séparateurs prédéfinis.

Une version du système QALC, le système FRASQUES, a été adaptée au français pour participer à l'évaluation EQueR en 2004 (Grau et al., 2006a). Pour ce faire, les outils sur lesquels repose l'analyse des questions (étiqueteur morphosyntaxique et analyseur syntaxique) ont été modifiés. Les sorties de ces outils ont été projetées sur des formats communs en français et en anglais, afin que le module d'analyse des questions puisse être le même dans les deux langues.

[9] Outil linguistique dédié au repérage des termes et de leurs variantes

Depuis 2005, une version crosslingue a également été élaborée pour participer aux campagnes d'évaluation CLEF-QA en 2005 et 2006 (Grau et al., 2006b). Ce système, MUSCLEF, prend en entrée des questions en français, et recherche leurs réponses dans des documents en anglais. Cela correspond en réalité à deux sous-systèmes, utilisant deux stratégies parallèles pour passer d'une langue à l'autre. La première stratégie s'appuie sur la traduction de la question par un traducteur automatique qui est ensuite passée en entrée du système QALC. La deuxième stratégie quant à elle consiste à traduire chaque terme de la question. L'ensemble des termes traduits est implémenté dans un système particulier appelé MUSQUAT. Enfin, les résultats des deux sous-systèmes sont ensuite combinés, et les réponses résultant de cette fusion sont celles du système MUSCLEF.

### 1.4.2 Le système QRISTAL

QRISTAL (Questions-Réponses Intégrant un Système de Traitement Automatique des Langues) (Laurent et al., 2005) est un système de question-réponse multilingue (français, anglais, portugais, italien et polonais), développé par Synapse Développement, pour extraire des réponses dans une base documentaire locale ou à partir du Web. Le système se compose de plusieurs modules de traitement automatique des langues, à savoir une analyse syntaxique, une désambiguïsation sémantique, une recherche des référents des anaphores, une détection des métaphores, un repérage des entités nommées et enfin une analyse conceptuelle et thématique. Cette utilisation massive des outils du traitement automatique des langues a largement contribué aux bons résultats obtenus par le système lors de l'évaluation EQueR 2004 puisque le système s'est classé premier sur sept systèmes participant. En novembre 2004, QRISTAL est devenu le premier système de question-réponse commercialisé pour la plate-forme Windows.

L'originalité de ce système réside dans son moteur d'indexation. En effet, chaque document d'une source de données est découpé en blocs de textes de longueur fixe de un kilo-octet. Ces blocs sont ensuite analysés syntaxiquement et sémantiquement afin de générer plusieurs index tels que : l'index des noms propres, l'index des expressions idiomatiques[10], l'index des entités nommées, l'index des domaines, l'index des types de question-réponse (définition,

[10] Expression qui a un sens dans une langue mais qui ne peut pas être traduite dans une autre langue mot à mot

distance…), l'index des mots-clés du texte, etc. Ces différents index offrent des possibilités intéressantes pour la recherche de la réponse. Ce même processus est identique pour chacune des langues traitées par QRISTAL.

Le système procède à une analyse syntaxique et sémantique de la question pour identifier le type de question-réponse parmi ses 86 types factuels (dimension, surface, pourcentage, etc.) et non factuels (comparaison, causalité, opinion, etc.). L'extraction de la réponse quant à elle se fait tout d'abord par l'analyse des blocs sélectionnés en rapport avec la question de l'utilisateur. Ce traitement repose principalement sur le calcul d'un poids pour chaque phrase candidate dans le but de déterminer l'ordonnancement des réponses. Ce poids concerne le nombre de mots et d'entités nommées repérés dans la phrase ainsi que la présence du type de la réponse attendue. Les phrases sélectionnées sont ensuite triées avant la phase d'extraction des entités nommées ou des groupes de mots correspondant aux réponses. Selon les auteurs, sur une sélection restreinte de textes, le temps de réponse est d'environ 3 secondes, tandis que sur le Web, les premières réponses sont retournées au bout de 2 secondes.

### 1.4.3 Le système PIQUANT

Le système PIQUANT d'IBM (Chu-Carroll et al., 2002) se fonde sur l'utilisation de plusieurs systèmes de question-réponse selon le type de la question et, par conséquent, bénéficie d'une meilleure pertinence grâce à la pluralité et à la redondance des réponses trouvées. Ainsi, PIQUANT s'appuie sur différents agents indépendants pour rechercher une réponse. Parmi ces « agents réponses », on note un agent fondé sur des outils statistiques et d'autres, sur des outils de traitement automatique des langues.

Le typage de la question repose sur une analyse syntaxique et permet de déterminer le type de la question, le type de la réponse désirée, les mots-clés et une forme sémantique de la question. Pour extraire les réponses, le système utilise plusieurs sources de connaissances comme WordNet, pour produire les synonymes des termes, ou encore CYC (http://www.cyc.com) (Lenat, 1995), pour les connaissances de sens commun qui permet au système de réaliser des inférences. Le choix de la réponse se fait en considérant un agent parmi d'autres suivant le type de la question.

### 1.4.4 Le système JAVELIN

(Nyberg et al., 2002) ont proposé un système de question-réponse JAVELIN (Justification-based Answer Valuation through Language INterpretation) fondé sur une interaction avec l'utilisateur dont l'intérêt est d'élucider la question et de déterminer une stratégie de recherche adaptée pour trouver la réponse. L'analyse de la question est réalisée par un analyseur dédié qui permet de déterminer certaines caractéristiques de la question : le type de la question selon une classification prédéfinie propre au système, le type de la réponse attendue, les mots-clés de la question avec leurs différentes variantes grâce au réseau sémantique de WordNet, la méthode de recherche à adopter et enfin une représentation sémantique de la question.

La recherche documentaire exploite plusieurs bases de données semi-structurées. L'idée est de sélectionner, suivant le type de la question, la base de données à interroger en utilisant une interface entre le système et les bases de données. Par exemple, une base de données biographiques est consultée pour les questions portant sur des dates de naissance. L'extraction de la réponse se fonde quant à elle sur une sélection de passages candidats à partir des documents retournés par le moteur de recherche. Cette sélection consiste dans un premier temps à filtrer les passages en éliminant tous les passages jugés non-pertinents sur la base du calcul d'un indice de confiance par rapport à la présence de la réponse dans le passage. Dans un second temps, il s'agit de classer les réponses sélectionnées selon les scores obtenus lors de l'étape précédente.

Le système offre aussi à l'utilisateur la possibilité de définir la procédure de recherche à sa question si celle-ci s'avère non fructueuse. En effet, JAVELIN permet un retour à tous les niveaux du processus de recherche grâce aux évaluations effectuées sur l'efficacité des différents modules utilisés. Le but de ce mécanisme est de permettre à l'utilisateur d'apporter plus de précision à sa requête et ainsi de guider la stratégie de recherche de la réponse. Enfin, JAVELIN fournit à l'utilisateur une justification de la réponse en lui renvoyant avec la réponse la description des traitements accomplis par le système.

### 1.4.5 Le système PowerAnswer

PowerAnswer est un système de question-réponse avec une architecture fondée sur le raisonnement logique. Proposé par (Moldovan et al., 2002) du LCC (Language Computer

Corporation), il repose sur la représentation sous forme de formules logiques de la question, de la réponse ainsi que des sources de données servant à extraire la réponse. La question est analysée afin de déterminer son type, le type de la réponse attendue et les mots-clés qui la composent. Cette analyse utilise les données sémantiques de WordNet ainsi qu'un module de reconnaissance d'entités nommées pour identifier les entités nommées présentes dans la question.

Pour l'extraction de la réponse, le système utilise le même module de reconnaissance d'entités nommées pour repérer et extraire l'entité nommée correspondant au type de la réponse attendue dans les passages sélectionnés. Dans le cas où la réponse recherchée n'a pas été trouvée, PowerAnswer utilise un programme de démonstration automatique de réponse en s'appuyant sur une base d'axiomes induite essentiellement à partir de WordNet pour construire un raisonnement permettant d'unir la représentation logique de la question et celle des réponses possibles. Cette base d'axiomes est dynamiquement enrichie par une liste d'axiomes produite automatiquement à partir des liens trouvés dans WordNet entre les mots-clés de la question et ceux des réponses. Cette justification des réponses permet au système de retourner à l'utilisateur non seulement les passages contenant la réponse mais aussi la chaîne de raisonnement liant la question et la réponse.

### 1.4.6 Le système WEBCOOP

Le système WEBCOOP (COOPérativité pour le WEB) (Benamara, 2004) est un système de génération de réponses coopératives. L'idée est de proposer à l'utilisateur des informations additionnelles (explications, justifications, etc.). Ce système de question-réponse permet de retourner une réponse même quand la question posée comporte des fausses présuppositions ou des malentendus. WEBCOOP se fonde sur l'intégration de procédures de raisonnement couplées à des modes de représentation de connaissances. Il a été développé pour répondre à des questions portant sur le domaine du tourisme en s'appuyant essentiellement sur une ontologie du domaine et des bases de connaissances regroupant les aspects hébergement et transport. L'analyse de la question repose sur une classification, construite manuellement, fondée sur la forme de la question ainsi que sur son focus. Les questions sont classées selon deux catégories où chaque catégorie se compose de plusieurs classes sémantiques permettant de mieux préciser le type de la réponse recherchée. Une première catégorie concerne les questions qui attendent une entité nommée comme réponse, tandis qu'une deuxième catégorie

porte sur les questions dont la réponse est une entité textuelle (définition, description, procédure, etc.). À l'issue de cette étape, la question est représentée sémantiquement sous forme d'un triplet : la catégorie de la question, le type de la réponse attendue et une représentation de la question en formules logiques du premier ordre.

L'extraction de la réponse repose sur l'utilisation d'un moteur d'inférences qui permet de comparer la question avec les documents de la base documentaire grâce à des procédures de raisonnement, en construisant des formules logiques associées aux réponses potentielles. La réponse est décomposée en deux parties selon le type de la question. La première partie consiste en une réponse directe à la question tandis que la deuxième représente une réponse coopérative et peut prendre différentes formes de coopérativités : justifications, avertissements, explications ou bien commentaires.

### 1.4.7 Le système d'InsightSoft

L'équipe d'InsightSoft-M a développé un système de question-réponse fondé sur l'application massive de patrons d'extraction sous forme d'expressions régulières et l'exploitation d'une base de connaissances factuelles (Soubbotin et al., 2002). Les questions sont analysées pour en déterminer le type et sélectionner les patrons à appliquer. Chaque patron ainsi sélectionné est appliqué à l'ensemble des passages candidats. L'appariement est réalisé sur la base du patron ainsi que des éléments de la base de connaissances (pays, monnaies, etc.). Ce système s'est classé premier lors de l'évaluation TREC-10, avec 77% de réponses correctes.

## 1.5 Problématique des systèmes de question-réponse en domaine restreint – Cas particulier du domaine médical

Le domaine de recherche des systèmes de question-réponse a considérablement évolué depuis les premiers résultats obtenus par ces systèmes lors de la première campagne d'évaluation des systèmes de question-réponse TREC-QA (Voorhees, 1999). Ainsi, il est apparu que les systèmes combinant des outils de traitement linguistique et des techniques de recherche d'information obtenaient les meilleurs résultats. Depuis, l'intérêt pour ce type de systèmes

s'est beaucoup développé puisqu'ils se sont imposés comme des systèmes capables d'extraire une information précise en réponse à une requête utilisateur.

Les systèmes de question-réponse en domaine ouvert se présentent comme des systèmes d'extraction d'information capables de traiter une grande masse documentaire. Néanmoins, la grande difficulté des systèmes travaillant en domaine ouvert réside dans la construction et l'exploitation de bases de connaissances génériques à tous les domaines. Cette problématique a un impact direct sur l'efficacité du système. En effet, pour pouvoir répondre correctement aux questions, ce type de systèmes requiert l'utilisation de connaissances sémantiques étendues nécessaires à la compréhension du texte en langage naturel. Une approche plus simple afin d'améliorer les performances des systèmes de question-réponse consiste à restreindre leur domaine d'application, c'est-à-dire à réduire la fonctionnalité des systèmes pour ne répondre qu'à des questions portant sur un domaine particulier. Une des particularités de ces systèmes est le fait qu'ils recherchent des réponses dans une collection fermée de textes. En raison de ce nombre réduit de documents, les systèmes ne peuvent pas exploiter la redondance informationnelle pour extraire des réponses comme c'est le cas des systèmes qui utilisent un ensemble plus important de documents ou le Web comme source de données (Berthelin et al., 2003). En contrepartie, le fait de travailler sur des domaines restreints leur permet souvent d'exploiter des connaissances plus élaborées et plus complètes.

Les premiers systèmes de question-réponse en domaine restreint, comme les systèmes LUNAR et BASEBALL, sont apparus pour interroger des bases de données regroupant des informations relatives à un domaine précis. Depuis, les travaux sur les systèmes de question-réponse se sont intéressés de plus en plus aux domaines de spécialité, dont la caractéristique est de rechercher des réponses dans des collections de documents techniques. Pour ce faire, la plupart des systèmes de question-réponse en domaine restreint se fondent sur une approche sémantique, c'est-à-dire exploitant les bases de connaissances du domaine étudié. Cette approche doit leur permettre de mieux maîtriser la terminologie appropriée et bénéficier des connaissances spécifiques afin d'interpréter les termes employés dans les textes et dans les questions. En effet, chaque domaine se distingue par son propre vocabulaire et des connaissances plus au moins spécifiques. Toutefois, la principale difficulté des systèmes de question-réponse en domaine restreint réside dans le fait que l'on ne dispose pas forcément de

toutes les ressources sémantiques nécessaires facilitant la recherche d'une réponse, même si la situation est ici plus favorable qu'en domaine ouvert.

Dans notre étude, nous nous sommes intéressé au domaine médical, un domaine qui ne cesse d'évoluer. L'apparition de nouveaux cas cliniques et le nombre de recherches menées dans ce domaine sur des traitements médicaux plus efficaces en sont la preuve. En outre, grâce à l'apport de nouvelles techniques de communication, une grande quantité de connaissances médicales est devenue disponible. De ce fait, les spécialistes de la discipline éprouvent la nécessité d'organiser l'information médicale afin de centraliser ou plutôt de normaliser les données médicales en raison de la qualité et l'hétérogénéité de l'information existante sur le Web. Cette possibilité d'accéder aux différentes ressources médicales attire de plus en plus d'utilisateurs, qu'ils soient du domaine médical ou pas, curieux d'enquêter sur la nature d'une maladie, son étiologie[11] et éventuellement son traitement ou bien tout simplement pour approfondir leurs connaissances.

Le domaine médical, comme tout autre domaine de spécialité, est caractérisé par la complexité de son vocabulaire et la spécificité de sa terminologie très technique (Thieulle, 1993 ; Zweigenbaum, 2001). Par conséquent, l'accès à la connaissance médicale requiert un traitement particulier notamment à cause de la structure des différentes ressources existantes sur le Web. Cette problématique, dans un domaine sensible comme la médecine, contraint à développer des bases de connaissances spécifiques au domaine contenant des informations médicales plus structurées et surtout, d'un point de vue médical, des données approuvées. Pour accéder à ces données pertinentes, l'utilisation de systèmes de recherche d'information manipulant les connaissances du domaine s'impose. Aussi, pour atteindre une information médicale précise, l'utilisateur doit maîtriser au préalable la terminologie appropriée lui permettant ainsi d'exprimer son besoin en information avec des requêtes précises. En pratique, un professionnel de la santé est parfois contraint de fouiller dans des bases de connaissances à la recherche d'une réponse médicale afin de satisfaire son besoin en information, une tâche qui peut être coûteuse en temps. Le temps constitue d'ailleurs un critère primordial dans le domaine de la médecine. En effet, lors d'une consultation, dans le but d'apporter plus d'assurance à sa décision, un médecin se doit d'obtenir une réponse

[11] Étude des causes directes des maladies. Ce terme désigne aussi les causes elles-mêmes.

précise et rapide à sa requête, une fonction difficilement réalisable en utilisant un système de recherche d'information classique. L'intérêt des systèmes de question-réponse dans ce domaine est donc grand.

L'objectif d'un système de question-réponse dans le domaine médical est donc d'apporter en un minimum de temps une réponse valide à une question formulée en langage naturelle et ce, en étant sensible aux connaissances médicales (Zweigenbaum, 2003). Cela signifie par exemple, être capable de répondre à des questions telles que : « Quels sont les effets secondaires du médicament Y ? ». La recherche d'information dans ce domaine a fait l'objet de plusieurs travaux. À titre d'exemple, on peut citer les travaux de (Rinaldi et al., 2004) qui ont adapté le système de question-réponse ExtrAns pour répondre à des questions portant sur le domaine de la génomique ou encore (Pustejovsky et al., 2002a) qui ont développé un système d'extraction d'information fondé sur la variation terminologique dans le domaine biomédical en exploitant les articles disponibles dans la base de données bibliographiques MEDLINE (Medical Literature Analysis and Retrieval System Online). Il ressort de ces différents travaux l'importance et la nécessité de disposer d'une base de connaissances sémantiques structurée du domaine afin d'obtenir une bonne performance d'un système de question-réponse.

Les ressources sémantiques dans le domaine médical sont nombreuses et accessibles sur Internet. C'est le cas du thésaurus MeSH, qui est particulièrement utilisé pour indexer les documents et les sites Web médicaux, de la CIM, de la SNOMED et de bien d'autres. L'intérêt d'utiliser ces différentes ressources existantes pour un système de question-réponse est d'abord de se constituer une base de connaissances sous la forme de larges listes de concepts médicaux telles que les listes de noms de maladies, des noms de traitements, des noms de médicaments, etc. Cet ensemble de listes permet aux systèmes d'identifier les différentes entités médicales dans les questions et surtout dans les documents. Cependant, la grande majorité de ces ressources, bien que riches en terminologie, sont en revanche beaucoup plus pauvres en relations que l'on peut qualifier de syntagmatiques (Embarek et al., 2006), c'est-à-dire des relations qui peuvent déterminer qu'une maladie M peut être traitée par le traitement T ou qu'un médicament D est prescrit pour guérir la maladie M. En effet, les relations entre les concepts ou plutôt les entités médicales présentes dans les ressources existantes sont principalement des relations de type hiérarchique comme les relations de

synonymie ou d'hyperonymie. Cette absence de relations syntagmatiques constitue une difficulté majeure pour un système de question-réponse. En effet, les questions les plus communément posées par les médecins (Ely et al., 1999) portent beaucoup sur ce réseau de relations comme l'illustre une question comme « Quel est le médicament à prescrire dans le cas X ? » et il est donc très avantageux pour un système de bénéficier de ce type de ressources sémantiques. Pour faire face à cette problématique, les systèmes de question-réponse évolués s'appuient sur le réseau sémantique de l'UMLS qui regroupe un ensemble de 134 types sémantiques médicaux (comme les maladies, les symptômes, etc.) reliés entre eux par des liens sémantiques non hiérarchiques. Cependant, sa partie francophone reste peu exploitable en raison du faible pourcentage, de l'ordre de 2% (Delbecque et al., 2005), des concepts médicaux couverts. Pour pallier à ce type de constat, le développement d'une terminologie médicale francophone s'est imposé. Par exemple, le projet UMLF (Unified Medical Lexicon for French) (Zweigenbaum et al., 2003) a permis de regrouper et d'unifier plusieurs ressources lexicales du domaine.

## 1.6 Limites actuelles des systèmes de question-réponse

Même si la performance des systèmes de question-réponse dépend de leur capacité à trouver des réponses dans les documents, elle dépend aussi fortement des résultats retournés par les moteurs de recherche. Elle est donc fortement liée à la formulation de la requête adressée au moteur de recherche. Le but de cette formulation, qui s'appuie sur l'analyse de la question, est de composer la requête d'interrogation à passer au moteur de recherche pour récupérer les documents pertinents par rapport à la question posée et susceptibles de contenir la réponse souhaitée. Cette requête est constituée essentiellement des mots-clés de la question posée. Dans le cas où les mots sélectionnés se révèlent ambigus, le moteur de recherche retourne un grand nombre de documents sans rapport avec la question engendrant ainsi un bruit important. Dans d'autres cas, les documents renvoyés portent sur le sujet de la question sans pour autant contenir la réponse recherchée. De plus, les moteurs de recherche sont moins performants lorsque les requêtes formulées sont composées de tous les mots de la question (Kwok et al., 2001).

La sélection des mots-clés s'avère d'autant plus primordiale pour un système de question-réponse qu'elle reste délicate à réaliser. À la différence d'une recherche documentaire classique, les questions sont formulées en langage naturel, ce qui requiert un traitement particulier de celles-ci. Dans le but d'extraire les termes importants de la question, les systèmes se fondent généralement sur des techniques du traitement automatique des langues. Ces techniques permettent d'expliciter le contenu informatif de la question en extrayant différentes données utiles à la formulation de la requête de recherche, à savoir les entités nommées, le type de la question et le type de la réponse à retourner. Afin d'apporter plus de précision à la requête d'interrogation, les systèmes de question-réponse utilisent des bases de connaissances existantes comme WordNet pour le domaine général ou encore des ressources sémantiques spécifiques pour les domaines techniques. L'intérêt de faire appel à ces sources de connaissances lexicales est de faire le lien entre les mots de la question et ceux apparaissant dans les documents, lesquels peuvent apparaître sous une forme différente de celle exprimée dans la question initiale (par exemple des synonymes).

Au-delà du simple problème de la génération de requête, la problématique des connaissances concerne tout le domaine du question-réponse comme nous l'avons illustré précédemment. Pour trouver des réponses correctes et précises à des questions, les systèmes ont besoin de s'appuyer sur des bases de connaissances structurées et valides. À cet égard, le Web constitue un enjeu de première importance. De par le grand nombre de ressources semi-structurées (Lin J., 2002) qu'il abrite, à l'image de Wikipédia par exemple, il autorise en effet la mise en œuvre de tout un ensemble de méthodes d'acquisition de connaissances pouvant être utiles aux systèmes de question-réponse.

Un autre enjeu pour ces systèmes est le traitement des questions complexes et en particulier la capacité de répondre à des questions nécessitant de composer plusieurs éléments de réponse provenant de différents documents en s'appuyant sur des inférences mettant en jeu des sources de connaissances. C'est l'ambition des systèmes de question-réponse à venir. Car actuellement, les systèmes de question-réponse classiques trouvent des réponses à des questions à condition qu'elles soient explicitement présentes dans un seul document. Considérons la question suivante : « Quel coureur espagnol a remporté cinq fois le tour de France ? », pour répondre à cette question, dont la réponse attendue est un nom de personne (***Miguel Indurain***), le système de question-réponse doit fusionner plusieurs réponses

candidates découlant de différentes ressources documentaires : un premier document comprenant le premier élément de réponse « … le coureur espagnol Miguel Indurain … » et un deuxième élément de réponse issu d'un document différent « … Miguel Indurain a gagné cinq tour de France … ». La fusion de ces deux réponses va permettre de justifier que « ***Miguel Indurain*** » est bien un coureur de nationalité espagnole et qu'il a effectivement gagné cinq fois le tour de France.

C'est dans cette perspective qu'a vu le jour le projet CONIQUE (CONtexte et Inférences en QUEstion-réponse), projet débuté en 2006 dont l'objectif premier est d'étudier l'idée d'intégrer dans les systèmes de question-réponse des mécanismes de compréhension de textes en s'appuyant sur des inférences. Le but est de permettre aux systèmes de réaliser des raisonnements sur des fragments de textes issus de différents documents afin de construire et surtout de justifier une réponse à une question. Contrairement à la plupart des études allant dans ce sens, le projet a pour ambition non pas d'exploiter des sources de connaissances statiques mais plutôt de modéliser l'extraction des connaissances à partir des documents en fonction du contexte exprimé par la question.

## 1.7 Conclusion

Dans ce chapitre, nous avons illustré la notion de question-réponse ainsi que l'intérêt de la recherche d'information précise, une information devenue de plus en plus disponible grâce à l'avènement du Web. Cette information précise peut être extraite au moyen de systèmes de recherche d'information automatisés, plus précisément des systèmes capables de satisfaire des requêtes d'interrogation formulées par les utilisateurs en renvoyant uniquement une réponse précise et en un minimum de temps. Ces systèmes sont appelés « systèmes de question-réponse ».

Un bref panorama dressé à partir de quelques systèmes de question-réponse existants montre que l'architecture classique d'un tel système repose sur trois modules. Le premier porte sur l'analyse de la question, le deuxième sur la recherche et la sélection de documents pertinents tandis que le dernier module se concentre sur l'extraction de la réponse recherchée. L'ambition commune de ces systèmes est d'exploiter en premier lieu la question afin d'en extraire tous les traits syntaxiques et sémantiques qu'elle contient. C'est une étape cruciale

qui joue un rôle prépondérant sur la performance du système de question-réponse. L'analyse de la question renferme différentes tâches dont l'objectif est de déterminer une stratégie de recherche appropriée en s'appuyant sur le type de la question, la reconnaissance des entités nommées, l'extraction des mots-clés et le type de la réponse attendue. Pour ce faire, les systèmes s'appuient sur des techniques différentes mais faisant le plus souvent appel à des outils de traitement automatique des langues. L'extraction de réponses à partir des documents candidats retournés par le moteur de recherche est également réalisée par une grande diversité de méthodes dont les plus utilisées sont fondées sur des outils statistiques ou encore sur des patrons d'extraction linguistiques.

Dans le but d'améliorer leurs performances concernant l'extraction de réponses, les systèmes de question-réponse évolués se fondent sur des bases de connaissances sémantiques. En effet, les résultats des campagnes d'évaluation des systèmes de question-réponse, telle la campagne TREC, ont démontré que les systèmes exploitant des ressources sémantiques obtenaient les meilleurs scores (Voorhees, 2002). Ces sources de connaissances permettent aux systèmes d'une part, de mieux appréhender la question posée par l'utilisateur et d'autre part, de bénéficier des ressources structurées et valides indispensables pour renvoyer une réponse correcte. La plupart de ces ressources comportent généralement des relations sémantiques de nature paradigmatique telles que des liens de synonymie ou d'hyperonymie. En revanche, elles contiennent rarement des relations sémantiques plus spécialisées telles que celles intervenant entre deux concepts ne comportant aucune relation d'équivalence. Ces relations, dites « relations syntagmatiques », que l'on retrouve généralement dans des bases de connaissances spécifiques à un domaine particulier, peuvent indéniablement constituer la réponse souhaitée (Vargas-Vera et al., 2004 ; Nyberg et al., 2002). Cependant, pour répondre convenablement aux questions, un système de question-réponse se doit d'utiliser un grand nombre de ressources sémantiques concernant divers domaines. Ce principe s'avère déterminant mais reste néanmoins difficile à exploiter. Une autre approche pour améliorer la performance d'un tel système consiste à restreindre son domaine d'application, c'est-à-dire un système capable de ne répondre qu'à des questions portant sur un domaine particulier.

Pour notre étude, nous nous sommes intéressé à ce type de système et plus particulièrement à une stratégie de recherche dans le cadre d'un système de question-réponse dédié au domaine médical La langue médicale, qui se distingue par la complexité et la richesse de sa

terminologie, intéresse depuis longtemps la communauté de la recherche d'information. Cet engouement s'est accentué, ces dernières années, grâce au développement des nouvelles technologies de l'information qui a rendu disponible et accessible une quantité considérable de sources de données médicales. Toutefois, les systèmes de recherche d'information actuels s'avèrent peu appropriés à la pratique médicale et affichent quelques limites dues essentiellement à la structure et à la validité de l'information disponible sur le Web. Afin de surmonter les difficultés auxquelles il est confronté quotidiennement lorsqu'il se trouve en consultation avec un patient, un médecin a besoin d'obtenir des réponses valides, en un minimum de temps, aux questions qu'il se pose. D'où l'intérêt d'un accès facile et rapide à l'information médicale, c'est l'objectif d'un système de question-réponse médical.

Dans le chapitre qui suit, nous nous intéresserons à la construction de ressources sémantiques pour le domaine médical. Comme nous l'avons illustré, ces bases de connaissances s'avèrent en effet d'une utilité prépondérante pour un système de question-réponse médical pour trouver des réponses précises aux questions propres à ce domaine.

# Deuxième chapitre
# Ressources linguistiques et terminologiques du domaine médical

# 2. Ressources linguistiques et terminologiques du domaine médical

*Dans le premier chapitre, nous avons introduit la notion de question/réponse et présenté notre domaine d'étude ainsi que les différentes approches utilisées pour développer un système de question-réponse. Après avoir démontré l'importance des ressources sémantiques pour un tel système, nous présentons dans ce chapitre quelques ressources terminologiques existantes dans le domaine médical. Puis, nous exposons une ontologie du domaine médical proposée dans le cadre de notre travail qui nous a permis de dégager les concepts médicaux choisis pour notre étude.*

## 2.1 Introduction

Nous avons illustré dans le premier chapitre l'importance des bases de connaissances dans le fonctionnement et la performance des systèmes de question-réponse (Ferret & Zweigenbaum, 2007). Le principal intérêt d'exploiter ces différentes connaissances est d'allouer à un système de question-réponse la compétence nécessaire pour identifier et désambiguïser les termes apparaissant dans les questions et dans les documents susceptibles de contenir des réponses. Le domaine médical constitue l'un des domaines de spécialité les plus importants et les plus traités depuis l'essor de l'informatique. Il se caractérise par une terminologie riche et complexe qui ne cesse en outre de croître du fait des évolutions rapides des recherches qui y sont menées. Cette terminologie se révèle d'une utilité prépondérante dans le traitement de l'information médicale, contribuant comme source de connaissances pour de nombreux travaux consacrés principalement au traitement automatique de la langue médicale (Zweigenbaum et al., 1996).

La richesse et la complexité du vocabulaire médical ont conduit depuis de nombreuses années au développement d'un ensemble important de ressources terminologiques et lexicales telles que le MeSH ou l'UMLS par exemple. Ces ressources ont été constituées dans le but d'une part, de normaliser la terminologie médicale et d'autre part, de faciliter l'accès à l'information médicale. L'effort qui sous-tend leur réalisation a permis à la fois une modélisation de la connaissance médicale et une meilleure structuration des données. L'utilisation de ces

ressources permet d'identifier plus facilement dans les textes les termes médicaux ainsi que leurs différentes formes (synonymes, hyperonymes, etc.) ou variantes terminologiques (McCray et al., 1994), capacité qui est très utile pour de nombreuses applications comme l'indexation des documents médicaux, la recherche d'information ou même les systèmes de question-réponse adaptés à la médecine (Alper et al., 2001 ; Zweigenbaum, 2003).

## 2.2 Ressources terminologiques et sémantiques dans le domaine médical

L'outil informatique et les nouvelles technologies de l'information et de la communication ont largement favorisé ces dernières années le développement d'un nombre important de sources de données électroniques. Ces ressources rendent ainsi disponible et accessible une masse impressionnante de données, ce qui permet actuellement à n'importe quel utilisateur d'atteindre plus aisément une information désirée. Ce fait est particulièrement sensible dans plusieurs domaines de spécialité, notamment le domaine médical. La science médicale est un champ d'étude très vaste qui se caractérise par un vocabulaire spécialisé, très complexe sémantiquement, qui ne cesse d'évoluer. Cette dynamique contribue largement à la fréquence d'accès à l'information médicale et à la nécessité de la mise à jour de cette dernière.

La mise à disposition de ces ressources médicales n'est néanmoins pas garante de la qualité de l'information trouvée lors d'une recherche, un point qui constitue un souci primordial en ce qui concerne l'information médicale. Ce problème provient essentiellement de l'hétérogénéité des informations de santé publiées sur le Web. Les sources d'informations sont en effet plus au moins organisées et homogènes de façon intrinsèque et contiennent de plus des différences dans la description des données. Par exemple, deux termes appartenant à deux sources de données différentes peuvent avoir la même appellation alors que leurs définitions sont incompatibles et vice versa.

Pour pallier ces différentes contraintes qui représentent un réel handicap pour les systèmes de recherche d'information, comme les moteurs de recherche ou les systèmes de question-réponse, le recours à des bases de connaissances médicales certifiées s'impose. Le domaine des ontologies constitue la solution à laquelle s'intéressent de nombreux travaux actuels de recherche d'information afin de résoudre le problème de l'hétérogénéité sémantique des données (Hakimpour et al., 2001). Ce domaine est devenu un champ de recherche intéressant

pour toute une gamme d'applications faisant appel à des connaissances d'un domaine contribuant ainsi au développement d'une nouvelle génération du Web, soit le « Web sémantique »[12] (Golbreich et al., 2002).

Le terme ontologie est utilisé depuis le début des années 90 dans les domaines de l'ingénierie des connaissances et de l'intelligence artificielle. Il s'agit d'un mot dérivé du mot grec « onto » qui signifie « l'existence », ce qui définit l'ontologie comme une science d'un « existant ». Ce terme, emprunté à la philosophie, s'intéressait à la science de l'Être, c'est-à-dire l'étude des propriétés générales de ce qui existe (source Wikipédia). (Uschold et al., 1996) caractérise une ontologie comme une branche de la philosophie qui a comme objet de représenter ce qui existe sous la forme d'une description abstraite, en insistant sur des catégories, principes et traits généraux. En informatique, le terme « ontologie » signifie un ensemble structuré de concepts où les concepts sont organisés dans un graphe dont les relations expriment des relations sémantiques entre les différents concepts. En tant que domaine, l'ontologie consiste en l'étude des catégories d'entités abstraites et concrètes qui existent ou peuvent exister (Sowa, 1999). Cependant, il est très difficile d'attribuer au terme « ontologie » une définition précise du fait qu'il est employé dans des contextes très différents. Néanmoins, la littérature en propose plusieurs définitions. Commençons tout d'abord par celle éditée par le dictionnaire « le Petit Robert », qui définit l'ontologie comme suit : la partie de la métaphysique qui s'applique à l'être en tant qu'être, indépendamment de ses déterminations. (Welty, 1998) propose la définition suivante : une ontologie est la définition de concepts, relations entre concepts, contraintes et règles d'inférences qui seront utilisés par un système de représentation des connaissances. Pour (Chandrasekaran et al., 1999), une ontologie est une théorie du contenu sur les sortes d'objets, les propriétés de ces objets et leurs relations possibles dans un domaine spécifié de connaissances. Toutefois, la théorie donnée par (Gruber, 1993), qui présente l'ontologie comme une spécification qui exprime une conceptualisation des agents existants dans un domaine avec leurs propriétés et leurs relations, est peut-être l'une de celles qui caractérisent le mieux la compétence d'une ontologie pour le Web et les raisons de construire des ontologies.

[12] Un ensemble de programmes de recherche visant à rendre le contenu des ressources du Web accessible et utilisable par d'autres applications.

Le rôle d'une ontologie est typiquement de représenter les connaissances d'un domaine spécifique au moyen de concepts et de relations intervenant entre ces différents concepts. Cette représentation doit garantir d'une part, le contrôle de la cohérence des données et d'autre part, l'évolution de sa structure. L'élaboration d'ontologies à partir de ressources documentaires d'un domaine donné constitue un réel intérêt, en particulier pour les systèmes de recherche d'information, leur permettant de gérer et d'exploiter les connaissances formulées dans les documents (Staab et al., 2003). De plus, il existe de nombreuses méthodologies possibles pour construire une ontologie. Pour de plus amples détails sur ces différentes techniques, le lecteur pourra se référer à la synthèse effectuée par Gomez-Pérez et ses collègues (Gomez-Pérez et al., 2004). On distingue différents types d'ontologies suivant le domaine modélisé et selon le degré de formalisation de leur structure et les modalités de définition de leurs concepts, ici nous les avons classés suivant les principaux travaux réalisés dans ce domaine :

- les ontologies génériques (globales) : ce sont des ontologies formelles qui couvrent plusieurs domaines, telle que WordNet par exemple (Fellbaum, 1998 ; Miller, 1990) ;
- les ontologies de domaine : ce sont des ontologies spécifiques à un domaine particulier. Elles se limitent à représenter les concepts d'un domaine précis (comme la, la géométrie, l'enseignement, etc.). Par exemple, l'ontologie « OntoPneumo » qui couvre le domaine de la pneumologie (Baneyx, 2007) ;
- les ontologies d'application : ce sont des ontologies très caractéristiques. Elles contiennent les connaissances spécifiques à une application. C'est le cas de l'ontologie « Toronto Virtual Enterprise » (Fox & Gruninger, 1994) qui décrit l'enchaînement des tâches d'une application, leurs coûts, etc.

Nous présentons dans ce qui suit quelques ressources terminologiques et ontologiques existantes explicitement conçues pour le domaine médical. Ces ressources ont été construites pour répondre à des besoins précis et divers : le thésaurus MeSH (cf. Section 2.2.1) est utilisé pour indexer des documents médicaux dans des bases documentaires, l'UMLS (cf. Section 2.2.5) a comme objectif de faciliter le développement de systèmes informatisés afin d'améliorer l'accès à l'information médicale, la CIM (cf. Section 2.2.3) permet le codage des dossiers patients à des fins statistiques, l'ORPHANET (cf. Section 2.2.4) répertorie tous les

noms de maladies rares et enfin la SNOMED (cf. Section 2.2.2) est une nomenclature[13] utilisée pour le codage des dossiers électroniques des patients. Nous décrivons également deux exemples de projets dont l'objectif était de construire une ressource ontologique pour le domaine médical : GALEN (cf. Section 2.2.6) est une ontologie médicale généraliste et MENELAS (cf. Section 2.2.7) est une ontologie couvrant les maladies coronaires.

### 2.2.1 MeSH

Le MeSH (Medical Subject Heading)[14] est un thésaurus numérisé. Il a été développé par la National Library of Medicine (NLM), principalement pour indexer la base bibliographique MEDLINE. Il est traduit en français par l'INSERM[15]. De nos jours, ce thésaurus est également utilisé pour l'indexation de nombreuses sources de données médicales. Le MeSH est une liste structurée de termes médicaux organisés en une arborescence. Au fur et à mesure que l'on descend dans la hiérarchie, les termes sont de plus en plus spécifiques. Ces termes sont appelés « descripteurs » car ils expriment de manière précise et spécifique le contenu d'un document. Les descripteurs, au nombre de 23 000 (en 2005), sont regroupés en 15 branches majeures. Par exemple la branche « A » correspond à l'anatomie (*Anatomy*), la branche « B » aux organismes (*Organisms*), la branche « C » aux noms de maladies (*Diseases*), etc. Chacune de ces branches contient plusieurs sous branches qui constituent les différents niveaux de la hiérarchie. Par exemple « C01 » pour la catégorie « Infections bactériennes et mycoses » (*Bacterial Infections and Mycoses*), « C02 » pour « Maladies virales » (*Virus Diseases*) ou encore « C03 » pour « Maladies parasitaires » (*Parasitic Diseases*).

Par ailleurs, chaque terme du thésaurus MeSH est associé à sa définition, ses synonymes et sa position dans l'arborescence (identifiant hiérarchique). Cependant, certains descripteurs peuvent apparaître dans plusieurs branches de l'arborescence, c'est-à-dire qu'un même terme peut appartenir à plusieurs catégories du MeSH et par conséquent, il peut donc avoir plusieurs identifiants. Un identifiant est composé d'un numéro alphanumérique : une lettre qui précise la catégorie (comme C = Maladies) et une série de nombres qui indiquent la position du terme

[13] Termes exhaustifs d'un domaine classés méthodologiquement.
[14] http://www.nlm.nih.gov/mesh/
[15] http://dicdoc.kb.inserm.fr:2010/basismesh/mesh.html

dans la hiérarchie. Par exemple, l'identifiant attribué au descripteur « Hépatite C » est « C02.440.440 », ce qui signifie : « C » pour Maladie, « C02 » pour la catégorie « Maladies virales », « C02.440 » pour « Hépatites virales humaines » (*Hepatitis, Viral, Human*) et ainsi de suite.

Le MeSH est utilisé par de nombreux systèmes de recherche bibliographique notamment pour indexer des sites et documents médicaux. C'est le cas par exemple de MEDLINE et de CISMEF :

- MEDLINE (Medical Literature Analysis and Retrieval System Online) est une base de données bibliographiques couvrant tous les domaines des sciences de la vie. Cette base est maintenue et mise à jour par la NLM depuis 1966. Elle est devenue la base de données la plus utilisée pour la recherche bibliographique dans le domaine biomédical. MEDLINE contient plus de 15 millions de références bibliographiques provenant d'environ 70 pays totalisant ainsi plus de 5000 sources biomédicales distinctes, indexées principalement par le thésaurus MeSH. Toutefois, les résumés, les titres et les descripteurs sont toujours en anglais. D'ailleurs, les articles en anglais sont majoritaires dans la base puisqu'ils représentent presque 85 % des références.

  L'interrogation de la base de données MEDLINE peut être effectuée via l'interface de plusieurs sites spécialisés, notamment le site « PUBMED » (http://www.pubmed.org), qui est le principal moteur de recherche de données bibliographiques du domaine biomédical.

- CISMeF (Catalogue et Index des Sites Médicaux Francophones) (http://www.chu-rouen.fr/cismef) est un annuaire électronique proposé par le Centre Hospitalier Universitaire (CHU) de Rouen. Développé en 1995, dès la création du site Web du CHU de Rouen, ce portail s'adresse en priorité aux professionnels de la santé. Il contient également des informations destinées aux patients et à leurs familles. CISMeF permet de trouver rapidement et plus facilement des sites et des documents médicaux francophones disponibles sur le Web. À ce jour, il recense et indexe environ 24 000 ressources francophones de qualité du domaine de la santé, soit plus de 24 000 documents et publications médicales indexés. Ces derniers sont organisés selon un

classement thématique en incluant les principales spécialités médicales. Comparativement aux bases de connaissances médicales en langue anglaise, les ressources francophones sont plus restreintes mais CISMeF ne les couvre encore que partiellement.

Le CISMeF s'appuie sur deux outils standards pour structurer l'information : le format de métadonnées du Dublin Core (http://www.dublincore.org) pour la description des ressources médicales et les mots-clés du thésaurus MeSH, pour l'indexation de ces ressources. De plus, le catalogue privilégie la qualité et la pertinence de l'information. Ainsi, il ne répertorie que les sites médicaux francophones répondant à des critères de qualité de l'information médicale sur Internet « Netscoring ». Ces critères sont regroupés en huit catégories principales qui sont : la crédibilité, le contenu, les liens, le design, l'interactivité, les aspects quantitatifs, les aspects déontologiques et l'accessibilité.

### 2.2.2 SNOMED

La SNOMED (Systematized Nomenclature of Medicine) (http://www.snomed.org/) est une nomenclature de type classification multiaxiale. La version SNOMED 3.5 (1998) comprend plus de 200 000 termes médicaux couvrant plusieurs domaines de la médecine. SNOMED a été élaborée en complémentarité avec la CIM et est actuellement traduite en 11 langues.

La SNOMED occupe une position intermédiaire entre un thésaurus et un système formel de concepts (ontologie) (Gangemi et al., 1992). Elle renferme des concepts de base qui peuvent être associés pour décrire des diagnostics ou des actes professionnels, ce qui autorise la constitution de bases de données médicales à partir de l'ensemble des informations constituant le dossier du patient ou son compte-rendu de sortie. Son vocabulaire est organisé selon onze axes de classification définis par une lettre (par exemple, T pour topographie, M pour morphologie, etc.). Les éléments à l'intérieur de chaque axe sont organisés suivant une structure hiérarchique. La classification d'un terme repose sur une décomposition de celui-ci en combinaison de termes appartenant à différents axes. Par exemple, la juxtaposition : M4405 (granulome éosinophile), F0300 (fièvre), E2001 (tuberculose) et T2800 (poumon) correspond à la phrase « tuberculose pulmonaire ». Cette possibilité de combiner des termes

appartenant à des classes différentes avec des qualificatifs et des termes relationnels permettant ainsi de composer des expressions fait de la SNOMED une terminologie très importante dans le domaine médical, notamment pour l'indexation des dossiers médicaux.

### 2.2.3 CIM-10

La Classification Internationale des Maladies (CIM-10)[16] (en anglais ICD pour International Classification of Diseases) publiée par l'Organisation Mondiale de la Santé (OMS), est apparue en 1993. Elle a pour but de répertorier les maladies, les traumatismes et l'ensemble des motifs de recours aux services de santé. Elle est notamment utilisée pour recenser les informations sanitaires utiles concernant les causes de mortalité et de morbidité dans différents pays. La CIM bénéficie d'une remise à niveau régulière, le chiffre 10 correspond à la dernière version exploitable de la classification (1993). Une nouvelle révision de la CIM est en cours de lancement dans le cadre du projet (CIM-11) administré par l'OMS.

La classification dans CIM-10 est monoaxiale comprenant 21 chapitres principaux dont 17 concernent des maladies et 4 concernent les signes, les causes et les facteurs de recours aux soins. Les maladies sont classées selon plusieurs catégories telles que : les maladies endocriniennes (E), les maladies du système nerveux (G), les maladies de l'appareil circulatoire (I), etc. Elles sont répertoriées suivant leur degré de gravité. Par exemple, le chapitre des maladies infectieuses recense le plus grand nombre d'entrées car ces maladies sont la première cause de morbidité et de mortalité dans le monde. Chaque entrée est identifiée dans la CIM par un code. Ce dernier est composé de quatre caractères : une lettre correspondant au chapitre suivie de trois chiffres pour spécifier les maladies définies à un niveau général. Par exemple, le code A15.9 indique une tuberculose de l'appareil respiratoire ou encore le code C91.1 désigne une leucémie lymphoïde chronique.

### 2.2.4 ORPHANET

ORPHANET (http://www.orpha.net) est une base de données sur les maladies rares et les médicaments orphelins en libre accès pour tous publics. Elle a été créée conjointement par la

[16] http://www.who.ch/hst/icd-10/icd-10.htm

Direction Générale de la Santé et l'Institut National de la Santé et de la Recherche Médicale (INSERM). Disponible sur Internet depuis 1997, le portail d'ORPHANET a pour objectif principal de faciliter pour les professionnels de la santé, les chercheurs, les malades et tous les autres types de publics l'accès aux informations validées et actualisées dont ils ont besoin sur les maladies rares et les médicaments orphelins. Ce portail reçoit en moyenne plus de 20 000 visiteurs par jour. Il est constitué d'une encyclopédie réunissant un vocabulaire d'environ 3800 maladies et couvrant une information détaillée sur plus de 1500 maladies rares, rédigée par des experts internationaux. Il propose également un répertoire de services spécialisés, à destination des professionnels et des malades, donnant de l'information sur les consultations spécialisées, les centres de références, les laboratoires de diagnostics, les projets de recherche en cours, les essais cliniques et les associations de malades.

De plus, la base ORPHANET est multilingue. Elle offre un choix de six langues : Français, Allemand, Anglais, Italien, Espagnol et Portugais. Enfin, afin de répondre à l'évolution continuelle des connaissances dans le domaine des maladies rares, une nouvelle version d'ORPHANET a été développée et rendue disponible fin 2006. Cette nouvelle version, plus exhaustive, doit apporter davantage d'information sur l'épidémiologie des maladies et leur prise en charge en situation d'urgence.

### 2.2.5 UMLS

L'UMLS[17] (Unified Medical Language System) (pour Système d'unification de la langue médicale) est actuellement la ressource terminologique de référence pour le domaine biomédical. Cette ressource, développée et maintenue par la NLM depuis 1986, est le résultat de la compilation d'une centaine de thésaurus de langues et structures différentes dont le MeSH et la SNOMED pour les plus connus d'entre eux, ce qui lui confère le statut de métathésaurus multilingue. Ce métathésaurus comporte donc la terminologie résultant de l'union des vocabulaires de ces différentes sources médicales tout en préservant les relations intervenant entre les termes.

[17] http://nlm.nih.gov/research/umls/

L'UMLS est constitué de plus d'un million de concepts (version 2006) et indique les relations existant entre les concepts. Ces derniers, au nombre graduellement croissant, sont reliés entre eux par des liens sémantiques hérités des ressources initiales. Les relations sémantiques présentes dans l'UMLS sont principalement des relations de nature paradigmatique telles que les relations de synonymie[18] ou d'hyperonymie ainsi que d'autres relations plus spécifiques comme la relation « affecte ». Par ailleurs, l'UMLS dispose d'un vaste réseau sémantique (Delbecque et al., 2005 ; Zweigenbaum, 2004) comportant 134 types hiérarchisés par le lien « is-a ». Ce réseau fait de l'UMLS la ressource terminologique du domaine médical la plus largement exploitée. Elle s'avère très appropriée pour le traitement de l'information biomédicale et par conséquent, elle constitue un outil précieux pour les systèmes de recherche documentaire, notamment pour repérer dans les documents médicaux les concepts spécifiques au domaine biomédical comme les gènes, les maladies ou encore les médicaments.

Cependant, l'utilisation de l'UMLS et de son réseau sémantique se révèle difficile pour la langue française puisque la majorité des termes intégrés dans le métathésaurus UMLS sont en langue anglaise. En fait, selon (Delbecque et al., 2005), la terminologie en français ne couvre que 2% des concepts présents dans l'UMLS. Ce constat est à l'origine de deux projets : le premier projet s'intitule l'UMLF[19] (Unified Medical Lexicon for French) (Zweigenbaum et al., 2003) et a pour objectif d'effectuer la collecte, la synthèse et la validation de ressources lexicales pour le traitement informatique du français médical. Il vise à générer un lexique contenant les variantes flexionnelles et dérivationnelles des termes médicaux ; le second, VUMeF[20] (Vocabulaire Unifié Médical Français) (Darmoni et al., 2003) a la tâche d'enrichir le vocabulaire en français dans l'UMLS afin d'augmenter les ressources terminologiques francophones du domaine médical.

Pour mieux appréhender le processus d'une recherche dans une base d'articles aidée par le métathésaurus UMLS, prenons l'exemple suivant (Lindberg et al., 1990) : il s'agit d'un médecin désirant prendre connaissance d'éventuelles recherches dans le domaine de l'efficacité de l'AZT[21] dans la prévention de l'apparition du SIDA[22] chez des personnes

[18] La synonymie est représentée de façon implicite par le fait que deux termes étiquettent le même concept.
[19] http://www-test.biomath.jussieu.fr/umlf/
[20] http://www.vidal.fr/vumef/
[21] Zidovudine : azidothymidine
[22] « AIDS » en anglais

pouvant être exposées au virus et qui ne sont pas HIV-positives[23]. Le médecin soumet une requête d'interrogation au système de recherche sous la forme de mots-clés : « AIDS and AZT ». L'interface UMLS assigne aux deux termes leur équivalent en vocabulaire MeSH, à savoir, « Acquired Immunodeficiency Syndrome » et « Zidovudine » et détermine au passage dans quelles sources d'information les deux termes apparaissent le plus fréquemment ensemble. Si les articles pour cette recherche sont trop nombreux, le système devra s'appuyer sur le réseau sémantique de l'UMLS pour mieux affiner la recherche en proposant au médecin de valider le type de relation le mieux approprié à son besoin. Une fois le choix effectué, le système peut alors générer une requête typique en termes MeSH.

### 2.2.6 GALEN

Le projet GALEN (Rector et al., 1996), développé à l'université de Manchester, visait à mettre en place une ontologie pour le domaine biomédical. L'objectif principal était donc de construire une représentation des concepts du domaine médical. La version initiale de GALEN (en 1995) comptait une hiérarchie de plus de 4000 concepts. Actuellement, plus de 52 000 concepts sont recensés. GALEN utilise un langage de représentation de la connaissance médicale « GRAIL » (Galen Representation And Integration Language) (Rector et al., 1997), une variété de logique de description.

Les concepts de l'ontologie de GALEN sont organisés en une arborescence formant un réseau sémantique où les différents concepts sont reliés entre eux essentiellement par la relation sémantique « sorte-de ». Cette propriété donne aux concepts une structure de graphe dirigé acyclique (Zweigenbaum et al., 1996). Chaque concept de cette ontologie est complété d'une déclaration des relations qui doivent ou peuvent le lier à d'autres concepts. Ces mêmes concepts et relations peuvent être combinés à volonté pour créer de nouveaux concepts structurés.

[23] Virus de l'Immunodéficience Humaine

### 2.2.7 MENELAS

MENELAS (Zweigenbaum et al., 1994) est un projet européen dont l'objectif était de construire une représentation formelle de concepts médicaux par une analyse des comptes-rendus hospitaliers, écrits en texte libre, en utilisant les techniques du traitement automatique des langues. Ce projet, qui a duré trois ans (1992-1995), a débuté en parallèle avec le projet GALEN. Le domaine d'étude concernait principalement les pathologies coronariennes.

À la fin du projet MENELAS, une ontologie spécifique a été élaborée. Cette ontologie comptait un ensemble d'environ 1800 types de concepts et une hiérarchie de plus de 300 types de relations. La construction de cette ontologie a été réalisée en exploitant plusieurs sources de données médicales et en adoptant une méthodologie proche de celle employée dans le projet GALEN, explicitée dans (Bouaud et al., 1995). Enfin, l'ontologie du projet MENELAS comportait également des lexiques sémantiques et morpho-syntaxiques de mots simples et composés. Les données syntaxiques incluaient la construction syntaxique des verbes alors que les données sémantiques étaient typiquement constituées par la définition des mots en termes de descriptions conceptuelles (1000 entrées pour le français) (Zweigenbaum et al., 1996).

### 2.2.8 Synthèse

Dans cette section, nous avons présenté quelques ressources terminologiques du domaine médical accessibles sur le Web. Un premier constat s'impose : contrairement aux terminologies médicales en langue anglaise, les terminologies disponibles en langue française sont plus limitées. De plus, les ressources francophones existantes s'avèrent peu adaptées à notre problématique. Elles souffrent parfois d'un manque d'informations ou dans certains cas, sont sur-définies, c'est-à-dire qu'elles contiennent plus de détails qu'il n'en faut. Nous n'avons donc pas trouvé une ontologie opérationnelle satisfaisant notre besoin. À l'image de réseaux lexicaux de même type mais plus généraux, comme WordNet, ces ressources sémantiques contiennent majoritairement des relations paradigmatiques de type synonymie ou hyperonymie et sont beaucoup moins riches en relations syntagmatiques comme celles spécifiant qu'une maladie « M » peut être soignée par le traitement « T » ou encore que l'examen « E » permet de diagnostiquer la maladie « M ». Ces différentes contraintes nous ont donc amené à proposer une ontologie du domaine médical exprimant des concepts

médicaux et des relations sémantiques spécifiques utilisés par un médecin généraliste dans le cadre d'une consultation de médecine générale.

## 2.3 Proposition d'une ontologie du domaine médical

L'intérêt des ontologies pour les systèmes de recherche et d'extraction d'information nous a donc amené à nous intéresser plus particulièrement à l'utilisation des ontologies pour les systèmes de question-réponse du domaine médical. Une ontologie médicale doit permettre de représenter et de structurer les connaissances du domaine de la santé. En représentation des connaissances, les ontologies existent sous la forme de concepts et de relations. Ce travail entrant dans le cadre du projet du Guide des Bonnes Pratiques Médicales (GPBM), nous nous sommes plus spécifiquement intéressé aux concepts et aux relations permettant de répondre aux questions auxquelles sont confrontés quotidiennement les professionnels de la santé dans l'accomplissement de leur pratique. En l'absence d'ontologie couvrant spécifiquement ce champ de connaissances, nous proposons une représentation du domaine de la médecine générale sous la forme d'une ontologie (voir Figure 2.1), adaptée à notre besoin. Les concepts sont organisés en une arborescence où ils sont reliés entre eux par des relations sémantiques spécifiques, c'est-à-dire non hiérarchiques.

Pour ce faire, il fallait donc tout d'abord déterminer et organiser les concepts et relations sémantiques, dans le cadre de ce domaine particulier, pour nous permettre de collecter le vocabulaire indispensable à la description des concepts sélectionnés.

L'ontologie proposée a été définie à la fois en sollicitant directement des médecins et par l'analyse des questions typiquement posées par des médecins généralistes (Ely et al., 1999 ; Ely et al., 2000)[24]. Dans (Ely et al., 1999), l'étude concerne une collection de 1101 questions médicales provenant de 103 médecins généralistes tandis que dans (Ely et al., 2000), il s'agit d'une collection de 4653 questions collectées auprès de plus de 100 médecins généralistes. Selon les auteurs, presque la moitié des questions recueillies porte sur des prescriptions médicamenteuses, les maladies infectieuses et l'étiologie des pathologies, avec des questions telles que « What is the dose of drug X ? », « What is the drug for condition X ? » ou « What is the cause of symptom X ? ».

[24] http://www.bmj.com/cgi/content/full/321/7258/429

Pour compléter notre étude concernant la sélection des concepts médicaux, nous nous sommes appuyé également sur les questions de la tâche médicale posées dans le cadre de la campagne d'évaluation des systèmes de question-réponse EQueR (cf. Section 6.1.1) qui proposait une collection de 200 questions. Enfin, une fois élaborée, l'ontologie a été supervisée et validée par des médecins participant au projet GPM.

Dans l'ontologie, les concepts médicaux sont représentés par des rectangles tels que le concept « Maladie », qui centralise toutes les expressions désignant des noms de maladies ou le concept « Symptôme », qui regroupe toutes les manifestations cliniques révélant la présence potentielle d'une maladie. Les relations, qui expriment le type d'interaction entre deux concepts, sont quant à elles représentées par des flèches permettant ainsi de déterminer le sens de lecture d'une relation telle que la relation « Traite » entre les deux concepts « Traitement » et « Maladie ». De plus, entre deux mêmes concepts, plusieurs relations peuvent intervenir, comme c'est le cas des deux relations « Contre-indication » et « Soigne » entre « Médicament » et « Maladie ». Enfin, les concepts et relations sélectionnés pour notre étude ont été différenciés par une couleur rouge.

Il est à noter que le concept « Phénomènes » représente ici toutes les manifestations pathologiques. Elles peuvent être de nature psychologique (stress, colère, troubles psychiques, etc.) ou physiologique (formation de vaisseaux collatéraux, eczéma, mycoses etc.). Nous avons également introduit quelques propriétés liées aux médicaments (classe, posologie, forme, etc.). Ces informations sont parfois très utiles pour un médecin lors des prescriptions des ordonnances. Enfin, pour distinguer les différents types de traitements, trois concepts ont été définis : traitements physiques (cure, massages, exercices physiques, …), traitements médicamenteux et traitements annexes (conseils).

Contrairement aux différentes ressources sémantiques disponibles pour le domaine médical qui contiennent majoritairement des relations de type hiérarchique (hyperonymie), l'ontologie proposée se concentre sur des relations spécifiques du domaine telles que la relation « Traite », la relation « Détecte », etc. Il est à noter que le métathésaurus UMLS recense des relations similaires comme la relation « May-Treat » (115 055 instances), « Has-Manifestation » (39 738 instances) ou encore « Disease-Has-Finding » (14 146 instances). Cependant, l'utilisation de ce réseau s'avère difficile en raison du faible taux de la terminologie en français.

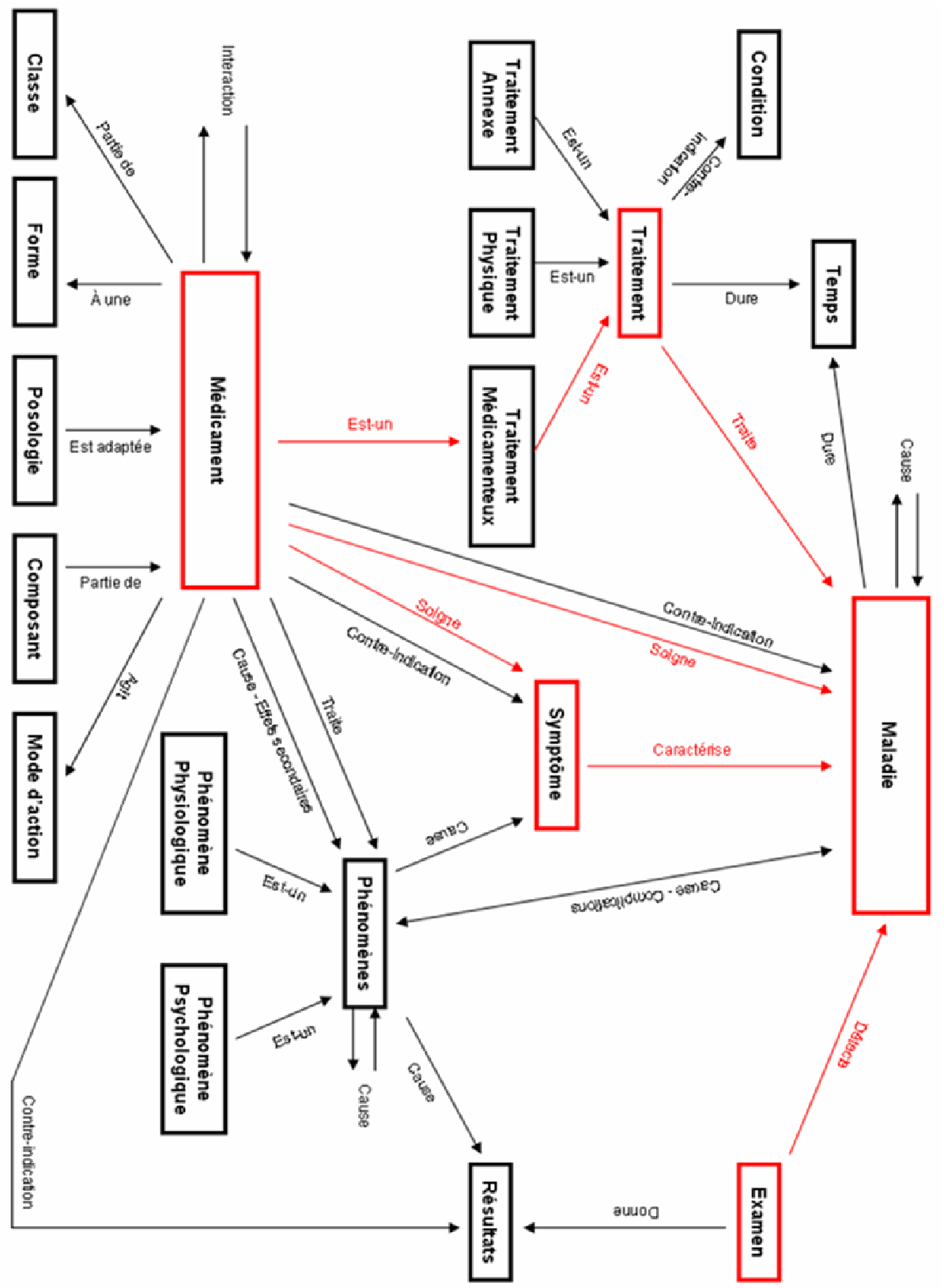


Figure 2.1 Ontologie du domaine médical

### 2.3.1 Concepts médicaux retenus

L'ontologie définie dans le cadre de notre travail (voir Figure 2.1) est une ontologie classique du domaine médical regroupant les concepts médicaux les plus courants. Ce choix est bien évidemment discutable mais il répond de façon adéquate à nos besoins. Ces concepts ont été retenus principalement sur la base des questions posées, au quotidien, par les médecins généralistes dans le cadre d'une consultation générale. Cette ontologie représente avant tout un simple échantillon d'étude pour un domaine aussi vaste que la médecine et peut éventuellement, dans l'avenir, être enrichie par de nouveaux concepts et relations.

Dans la suite de notre étude, nous nous sommes plus spécifiquement restreints à un sous-ensemble des concepts de notre ontologie, notre objectif étant de montrer l'intérêt et la fiabilité de notre approche et non de construire un système directement commercialisable. Sont ainsi concernés les cinq concepts médicaux suivants : Maladie, Traitement, Médicament, Symptôme et Examen. Elles sont définies comme suit :

- le concept « Maladie » regroupe tous les noms de maladies,
- le concept « Symptôme » concerne toute manifestation extérieure révélant la présence d'une pathologie comme la fièvre, l'insomnie, les douleurs abdominales, etc.,
- le concept « Traitement » comporte tous les traitements physiques (chimiothérapie, radiothérapie, cure, …) et annexes (conseils),
- le concept « Médicament » centralise tous les noms de médicaments pouvant être prescrits par un médecin (aspirine, upfen, voltarène, …),
- le concept « Examen » contient tous les noms des examens médicaux (IRM, bilan sanguin, coloscopie, etc.).

À noter que dans le cadre de notre travail, bien qu'il soit possible de considérer le concept « Médicament » comme un traitement, nous l'avons traité indépendamment du concept « Traitement » puisqu'il représente une classe sémantique importante dans une consultation de médecine générale.

### 2.3.2 Relations sémantiques retenues

La plupart des ontologies disponibles pour le domaine médical sont organisées sous la forme d'une classification hiérarchique de concepts référencés entre eux par des relations sémantiques paradigmatiques, c'est-à-dire des relations comme l'hyperonymie. Ces ressources sémantiques sont néanmoins beaucoup moins riches en relations syntagmatiques, sans doute parce que beaucoup plus nombreuses, plus diverses et moins bien caractérisées sur un plan formel. Ces relations sont pourtant extrêmement utiles pour les systèmes d'extraction d'information ou les systèmes de question-réponse car elles portent une grande part des connaissances recherchées du fait même de leur spécificité.

Nous avons donc constitué notre ontologie médicale en nous basant essentiellement sur les relations sémantiques les plus spécifiques pouvant intervenir entre les différents concepts médicaux définis dans l'ontologie (voir Figure 2.1) comme celles spécifiant qu'un symptôme *S* est un signe clinique de la maladie *M*, représentée dans l'ontologie par la relation « Signe », ou qu'un médicament *D* est contre-indiqué pour la maladie *M*, exprimée par la relation « Contre-indication ». Cette volonté est motivée par l'utilisation qui est faite de ces relations sémantiques pour aider un système de question-réponse médical à interpréter et à répondre aux questions habituellement posées par des professionnels de la santé, questions à l'image de « Quel médicament prescrire dans le cas d'une aniridie ? » ou « Quel examen médical permet de détecter un cancer du côlon ? ».

Pour notre travail, nous nous sommes focalisé sur l'étude de quatre relations sémantiques, à savoir les relations intervenant entre le concept « Maladie » et les quatre autres concepts médicaux retenus également pour cette étude (Médicament, Symptôme, Traitement et Examen) (voir Figure 2.2). Ainsi, parmi toutes les relations de la Figure 2.1, les quatre relations auxquelles nous nous sommes attaché sont :

- la relation « Traite » : Maladie – Traitement,
- la relation « Soigne » : Maladie – Médicament,
- la relation « Détecte » : Maladie – Examen,
- la relation « Signe » : Maladie – Symptôme.

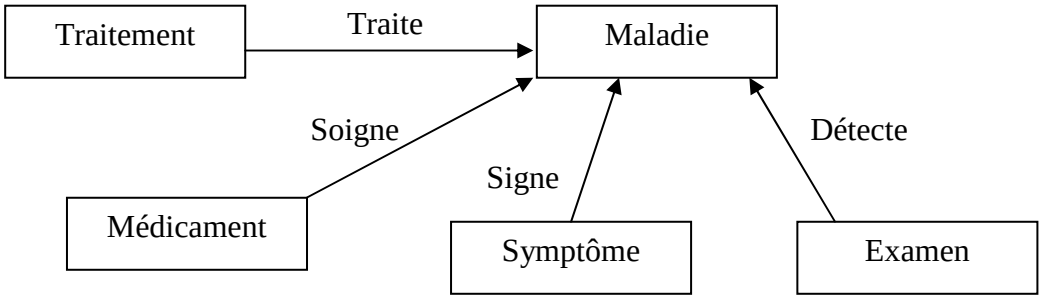


Figure 2.2 Sous-ensemble de l'ontologie du domaine médical de la Figure 2.1 retenu pour notre étude

## 2.4 Conclusion

Dans ce chapitre, nous avons introduit le domaine des ontologies et nous avons fait un tour d'horizon sur les ontologies médicales existantes. Ce domaine suscite un intérêt particulier de la part de la communauté de la recherche d'information et fait l'objet de larges attentes pour toute une gamme d'applications faisant appel aux connaissances d'un domaine. Les ressources sémantiques existantes se révèlent peu satisfaisantes pour notre problématique (manque d'informations, de liens, etc.) ou sur-définies. Nous n'avons donc pas trouvé d'ontologie opérationnelle satisfaisant notre besoin, en particulier concernant la présence de relations non paradigmatiques. Cette contrainte nous a amené à proposer une ontologie sur les bonnes pratiques médicales allant dans ce sens, plus clairement une ontologie fondée sur les concepts et relations utilisés par les médecins généralistes dans le cadre d'une consultation de médecine générale. Une telle ontologie peut éventuellement être utilisée par la suite comme support à des opérations de raisonnement pour de nombreuses applications à l'instar de systèmes de recherche et d'extraction d'information comme les systèmes de question-réponse. Pour notre étude, nous nous sommes limité à quelques concepts et relations sémantiques (voir Figure 2.2) mais avec la perspective de pouvoir étendre la couverture des relations de notre ontologie médicale.

Dans le chapitre qui suit, nous nous attacherons au problème du peuplement automatique de cette ontologie. Nous exposerons d'abord la manière dont nous identifions ses concepts dans des textes médicaux ainsi que les ressources médicales constituées pour supporter cette reconnaissance. Puis nous évoquerons l'extraction des relations sémantiques intervenant entre ces différents concepts médicaux.

# Troisième chapitre
# Enrichissement d’une ontologie du domaine médical

# 3. Enrichissement d'une ontologie du domaine médical

*Le chapitre 2 nous a permis d'introduire le domaine des ontologies et de mettre en évidence leur intérêt pour de nombreuses applications fondées sur les connaissances d'un domaine, en particulier pour les systèmes d'extraction d'information. Nous avons également présenté notre ontologie du domaine médical puis décrit les concepts médicaux et relations retenus pour notre étude. Dans ce troisième chapitre, nous nous intéressons à la construction des ressources concernant les concepts choisis ainsi qu'à leur identification dans les documents médicaux. Enfin, nous abordons la méthode employée pour extraire et apprendre de nouvelles relations sémantiques.*

## 3.1 Introduction

L'une des étapes incontournables et importantes dans la chaîne de traitement des systèmes de question-réponse est sans aucun doute l'identification des entités nommées dans les collections textuelles (Ferret et al., 2001a). L'ensemble des systèmes de question-réponse présentés dans le premier chapitre (cf. Section 1.4) montre l'intérêt d'une telle tâche dans l'analyse de la question et l'extraction de la réponse à partir des documents candidats. Ce traitement permet d'apporter à la fois plus de précision sur le type de la réponse attendue et sur les informations importantes d'un document. Il est plus généralement utilisé dans de nombreuses applications qui relèvent du domaine du traitement automatique de la langue naturelle comme les systèmes de résumé automatique et les systèmes d'extraction d'information.

Une entité nommée (Ehrmann, 2008) est définie classiquement comme un terme ou un ensemble de termes désignant un objet précis. Elle concerne d'une manière générale les noms propres (tels que les noms de personnes, les organisations, les lieux) mais également des entités autres que des noms propres comme les expressions numériques et temporelles. Ces types d'entités nommées sont communément appelés des entités de type MUC (Message Understanding Conferences) (Grishman et al., 1995), la tâche d'identification et d'extraction d'entités nommées ayant été initiée par les conférences MUC. Cette tâche consistait à repérer

ces différents types d'entités dans des textes journalistiques en anglais (Grishman et al., 1995 ; MUC-7, 1998). Dans le cas du domaine particulier qui nous intéresse, à savoir le domaine médical, la reconnaissance des entités consiste à identifier automatiquement dans des textes médicaux les instances des classes sémantiques du domaine (comme les noms de maladies, des traitements, des examens médicaux, etc.) à partir des techniques traditionnellement utilisées pour extraire des entités nommées.

Cependant, la reconnaissance des entités nommées doit s'appuyer sur de nombreuses ressources terminologiques concernant les instances de concepts du domaine à étudier. Elle se révèle une tâche complexe du fait du caractère ouvert et donc nécessairement incomplet du vocabulaire exploité. Il est à la base difficile de recenser la terminologie spécifique à un domaine de façon exhaustive, notamment pour le domaine médical, qui est caractérisé par la grande richesse de son vocabulaire. Cet exercice est ensuite compliqué lorsque le domaine concerné, ce qui est le cas du domaine médical, est en évolution constante et rapide. Cette évolution rend en effet la maintenance et l'enrichissement de ces ressources très lourds. Face à ces difficultés, on peut considérer les informations stockées sur le Web comme étant une base de connaissances intéressante permettant la construction de ressources terminologiques appropriées. Dans ce contexte, plusieurs travaux ont été réalisés concernant l'intérêt d'exploiter le Web comme source de connaissances pour le traitement automatique des langues (Grefenstette, 1999) et plus particulièrement pour la reconnaissance des entités nommées (Jacquemin et al., 2000a ; Fourour et al., 2003).

Ce chapitre s'intéresse au peuplement de l'ontologie médicale définie (voir Figure 2.1). Nous nous intéressons dans un premier temps à l'identification des attributs des concepts médicaux retenus pour notre étude dans les collections textuelles. Dans un second temps, nous abordons l'extraction de relations sémantiques entre les concepts étudiés.

## 3.2 Identification des concepts

Les entités nommées occupent une place prépondérante dans les systèmes de question-réponse, à la fois parce qu'elles sont souvent l'objet de la question (en particulier pour les questions factuelles) et sont donc les informations recherchées mais aussi parce qu'elles

constituent des éléments très discriminants. Leur identification à partir de textes écrits constitue la tâche actuelle d'extraction d'information la plus efficace. Elle obtient en moyenne des taux combinés de précision et de rappel de l'ordre de 0,90 sur des dépêches journalistiques (MUC-7, 1998).

Il existe différentes techniques pour repérer les entités nommées dans les textes. Cependant, on distingue généralement deux grandes approches utilisées pour l'identification des entités nommées : une approche fondée sur des règles de reconnaissances écrites manuellement et composées d'un ensemble d'automates à nombre fini d'états ou d'expressions régulières et de dictionnaires ; une approche à base d'apprentissage s'appuyant sur des corpus étiquetés. La première approche repose sur la description syntaxique et lexicale des syntagmes recherchés. Ces derniers sont identifiés et typés par des règles de grammaire exploitant des marqueurs lexicaux, des dictionnaires de noms propres et des dictionnaires de la langue générale. Ces techniques permettent de repérer différentes expressions comme les noms propres, les expressions temporelles ou numériques, les sigles ou acronymes, les mots inconnus ou encore de gérer les amorces, par exemple M. (pour Monsieur) ou Dr. (pour docteur). La seconde approche quant à elle, utilise des techniques d'apprentissage pour construire un modèle, soit descriptif, soit discriminant, à partir de larges corpus de textes étiquetés manuellement et utilise ensuite le modèle obtenu pour étiqueter de nouveaux textes. Le résultat de la phase d'apprentissage peut se présenter comme un ensemble de règles d'induction logique (Sasaki et al., 2000), un arbre de décision (Béchet et al., 2000) ou encore un modèle numérique (Kubala et al., 1999). Cette approche est actuellement largement reprise par de nombreux travaux de recherche concernant le domaine de l'extraction d'information (Bikel et al., 1997). Elle se révèle particulièrement robuste pour les systèmes traitant des textes bruités comme les systèmes dédiés à l'oral (Kubala et al., 1999) alors que les techniques relevant de la première approche s'emploient plutôt sur des corpus homogènes de textes respectant des critères rédactionnels stricts facilitant ainsi la tâche d'identification (Poibeau et al., 2003).

Enfin, d'autres systèmes ont été développés proposant une approche mixte selon le type du document analysé. Dans ces systèmes, les règles sont apprises automatiquement puis révisées par un expert (Aberdeen et al., 1995). L'approche inverse consiste à élaborer manuellement des règles qui seront par la suite étendues automatiquement pour améliorer la couverture du

corpus (Cucchiarelli et al., 2001). Toutefois, les différentes approches utilisées pour identifier les entités nommées obtiennent des résultats globalement comparables.

L'identification dans les documents des entités médicales que nous avons retenues a été réalisée indépendamment de leur nature intrinsèque (entité nommée au sens littéral ou terme) en adoptant une approche à base de règles mêlant patrons morpho-syntaxiques et listes d'entités ou d'éléments caractéristiques de ces entités. Ces règles ont été définies manuellement à partir d'un travail sur corpus (voir des exemples de règles dans l'Annexe 3). Nous avons repris en l'occurrence une des approches classiquement utilisées pour identifier des entités nommées de nature plus générale comme les noms de personnes, les organisations ou les lieux. À la différence de ces dernières, les patrons morpho-syntaxiques ont ici une importance moindre ce qui, à l'inverse, donne un rôle plus central aux listes d'entités ou de parties d'entités et justifie le soin tout particulier apporté à la constitution de ces listes pour chaque type d'entités médicales en exploitant pour ce faire à la fois des ressources existantes sur le Web (cf. Section 3.2.1) et les dictionnaires de l'Académie de Médecine sous forme électronique.

Dans la suite de cette section, nous présentons dans un premier temps les ressources du domaine de la santé utilisées pour construire notre base de connaissances médicales. Dans un second temps, nous exposons l'approche adoptée pour identifier les différents concepts médicaux dans les textes spécialisés.

### 3.2.1 Construction des ressources

Le degré de spécialisation, la structure ou l'incomplétude des bases de connaissances existantes (comme WordNet dans le domaine général ou le MeSH et l'UMLS dans le domaine médical) rendent leur exploitation difficile pour beaucoup d'applications qui doivent s'adapter à un corpus de textes (Charlet et al., 1996). De ce fait, la sélection des ressources sémantiques à exploiter se révèle comme une phase déterminante pour des systèmes ayant comme objectif de désambiguïser et interpréter le contenu des sources de données textuelles, à l'instar des systèmes de question-réponse.

Chaque ressource terminologique est construite pour une tâche spécifiée, ce qui rend une ressource sémantique conforme au type d'application que l'on veut développer difficile à trouver. D'autre part, construire une nouvelle ressource pour chaque nouvelle application est très coûteux. Il est en revanche possible, dans certaines situations, de spécialiser une base de connaissances et de l'adapter au domaine et au traitement envisagés (Basili et al., 1998).

Dans le cadre de ce travail, nous avons réalisé la construction d'une base de connaissances concernant la terminologie médicale recouvrant pour l'essentiel les concepts médicaux retenus à partir du modèle de l'ontologie médicale proposée (voir Figure 2.1). Pour ce faire, nous avons exploité dans un premier temps les ressources sémantiques francophones du domaine médical disponibles sur le Web et dans un second temps, les dictionnaires médicaux édités par l'Académie de Médecine. Plus précisément, nous avons collecté une liste de termes pour chaque entité médicale (Maladie, Médicament, Symptôme, Traitement et Examen). Par exemple pour les noms des maladies, nous nous sommes appuyé à la fois sur la terminologie de Orphanet et sur le vocabulaire de la base medisite[25], également exploitée pour récupérer les noms des médicaments et des traitements médicaux. Pour les noms de médicaments, nous avons en outre utilisé les données du site médical Doctissimo. En ce qui concerne l'entité symptôme, nous avons exploité les sources médicales citées ci-dessus ainsi que d'autres sites annexes comme la liste des symptômes en médecine humaine[26], Passeportsanté[27] ou encore l'encyclopédie Vulgaris-médical[28]. Quant aux noms des examens médicaux, nous avons utilisé le contenu des dictionnaires de l'Académie de Médecine. Ces mêmes dictionnaires contiennent aussi des définitions sur certaines maladies ainsi que leurs symptômes et traitements. Enfin, pour mieux caractériser les symptômes et les maladies (par exemple *cancer du poumon gauche*), cette collection de termes a été complétée par une terminologie concernant l'anatomie récupérée du thésaurus MeSH (correspondant à la branche « A » de l'arborescence).

À noter que l'ensemble des listes d'entités médicales constituées a été examiné et validé par les médecins participant au projet GPM. Ces derniers ont aussi contribué à l'enrichissement de la base de connaissances par l'ajout de quelques expressions.

[25] http://www.medisite.fr/medisite/
[26] http://fr.wikipédia.org/wiki/Liste_des_symptomes_en_medecine_humaine
[27] http://www.passeportsante.net/
[28] http://www.vulgaris-medical.com

Nous présentons ci-dessous les sources les plus importantes utilisées pour constituer notre terminologie médicale, à savoir le site ORPHANET (voir Section 2.2.4), le site Doctissimo et les dictionnaires de l'Académie de Médecine :

**- Le site Doctissimo** (http://www.doctisssimo.fr) est un portail francophone, géré par des médecins, dédié exclusivement à la santé. Il est accessible et consultable gratuitement par tout type de publics. Lancé en 2000, Doctissimo est une filiale de MEDCOST, société de services spécialisée dans le secteur de la santé. Depuis, le site est devenu le premier site d'information médicale en français. Il est en effet le site le plus consulté par le grand public dans le domaine de la santé (plus de 6 millions de visiteurs mensuels). Doctissimo propose plusieurs services consacrés à la santé et au bien être tels que des forums, des articles, des reportages, etc. Son forum est d'ailleurs le plus actif des forums médicaux francophones. La particularité de Doctissimo est certainement son encyclopédie médicale qui recense une terminologie abondante sur les principales maladies ainsi que son dictionnaire sur les médicaments commercialisés en France. Ce dictionnaire de médicaments, classés par ordre alphabétique, est la deuxième base de données sur les médicaments la plus interrogée sur le Web francophone après celle de Vidal[29]. À chaque médicament est associée une fiche descriptive synthétisant les résumés des caractéristiques du produit comme la dénomination officielle, la molécule active, la classe thérapeutique, le laboratoire fabricant, les indications, le mode d'action, les effets secondaires, les contre-indications, les interactions médicamenteuses et le surdosage.

**- Les dictionnaires de l'Académie de Médecine** sont des dictionnaires spécialisés dans le domaine de la médecine et édités par l'Académie de Médecine. Leur but est de rassembler l'ensemble du vocabulaire médical en usage afin de constituer une terminologie de ce domaine. Chaque dictionnaire relève d'une spécialité du domaine de la santé. Actuellement, il existe plusieurs de ces dictionnaires comme ceux concernant la biologie, la dermatologie, la cardiologie, la neurologie, etc. Chaque volume répertorie un index de termes, classés par ordre alphabétique, spécifique au domaine de spécialité. Cependant, en raison de la taille du vocabulaire exprimé dans une discipline, il se peut qu'un volume soit composé de plusieurs tomes. Dans le cadre de notre travail, nous avons disposé de quatre dictionnaires sous forme

[29] Site réservé aux professionnels de santé qui répertorie l'ensemble des médicaments commercialisés en France accessible en ligne sur le site : http://www.vidal.fr

électronique : le dictionnaire de biologie, de l'imagerie médicale et des rayonnements, de l'appareil digestif et enfin un dernier concernant l'anesthésie et la réanimation. La plupart des termes contenus dans ces dictionnaires sont accompagnés de leur catégorie grammaticale, leur traduction en anglais, leur désignation, leur structure anatomique, leurs synonymes, leurs antonymes et dans certains cas, une explication complémentaire leur est associée.

Nous avons réalisé un premier travail de formatage des dictionnaires. Ce travail consistait à transformer ces ressources sous forme électronique (Word) en un format XML (eXtensible Markup Language ) de description des données permettant ainsi l'exploitation du contenu des dictionnaires. Ci-dessous, deux exemples du terme « adénocarcinome » présent dans le dictionnaire de biologie : le premier est au format initial, c'est-à-dire avant la transformation, et le second dans un format XML après adaptation.

**adénocarcinome** n.m.
adenocarcinoma
Tumeur maligne épithéliale dont l'aspect morphologique reproduit, de façon plus ou moins fidèle et différenciée, la structure d'un tissu glandulaire.
*Syn*. carcinome glandulaire, carcinome cylindrique
*Étym.* gr. *adên* : glande ; *karkinos :* crabe

```
<MOT>
   <INTITULE>adénocarcinome </INTITULE>
   <CATEGORIE>n.m.</CATEGORIE>
   <DEFINITIONANG>adenocarcinoma</DEFINITIONANG>
   <DEFINITION>Tumeur maligne épithéliale dont l'aspect morphologique reproduit, de façon plus
    ou moins fidèle et différenciée, la structure d'un tissu glandulaire.</DEFINITION>
   <SYNONYME> carcinome glandulaire, carcinome cylindrique</SYNONYME>
   <ETYMOLOGIE> gr. adên : glande ; karkinos : crabe</ETYMOLOGIE>
</MOT>
```

### 3.2.2 Reconnaissance des entités médicales

La reconnaissance des entités nommées dans le domaine médical et surtout biomédical a déjà fait l'objet de nombreux travaux, surtout centrés sur la détection des noms de gènes (Proux et al., 1998 ; Collier et al., 2000), des noms de protéines (Fukuda, 1998) ou encore les termes MeSH, UMLS (Delbeque et al., 2005), etc.

Pour identifier les entités médicales dans les documents, nous avons utilisé une approche à base de règles de reconnaissance. Ces règles ont été écrites manuellement associant patrons morpho-syntaxiques et listes d'entités médicales constituées au préalable à partir de bases de connaissances existantes pour la spécialité. Chaque règle de reconnaissance d'une entité est composée d'un déclencheur, d'un contexte précédent, d'un contexte suivant, du type d'entité identifié et éventuellement d'une forme normalisée de l'entité, laquelle n'est pas utilisée dans le cas présent. Le déclencheur est constitué d'un mot ou d'une liste de mots permettant de se positionner dans le texte sur l'élément identifié et de déclencher l'application de la règle. Les contextes précédent et suivant représentent le contexte dans lequel doit se trouver le déclencheur pour que la règle s'applique. Ils sont définis par des expressions régulières et dans certains cas, ils ne contiennent aucun élément. Les règles sont compilées sous la forme d'automates. Leur application s'effectuant à la suite de l'étiquetage morpho-syntaxique réalisé par l'analyseur LIMA (LIC2M Multilingual Analyzer) (cf. Section 4.3), le module de reconnaissance des entités médicales dispose pour chaque mot à la fois de sa forme fléchie, de sa catégorie morpho-syntaxique et de sa forme normalisée. Le déclencheur, le contexte précédent et le contexte suivant d'une règle prennent donc la forme d'expressions régulières pouvant porter sur la forme fléchie, la forme normalisée ou la catégorie d'un ou de plusieurs mots. Plus formellement, les règles prennent la forme suivante :

**déclencheur : contexte_précédent : contexte_suivant : type_d'expression**

Par exemple, la règle

@AnnonceurMaladie **:** $L_DET ? **:** ($L_DET) ($L_NC|$L_NP) **:** MALADIE[30]

où :

déclencheur : @AnnonceurMaladie

contexte_précédent : $L_DET ?

contexte_suivant : ($L_DET) ($L_NC|$L_NP)

type_d'expression : MALADIE

[30] ? marque classiquement un élément optionnel tandis que ( | ) note une alternative. $L_DET, $L_NC et $L_NP sont des catégories morpho-syntaxiques, correspondant respectivement à déterminant, nom commun et nom propre.

permet d'identifier « *maladie de Lyme* » comme une maladie dans la phrase « La maladie de Lyme est une… », tandis que la règle

[@AnnounceurSymptome] **: :** [,] [comme] [$L_DET] $L_NC **:** SYMPTOME[31]

reconnaît « *fièvre* » comme un symptôme dans « … symptôme, comme la fièvre … ». On peut noter à cette occasion la présence de références à des listes permettant de regrouper des éléments linguistiques ayant un même rôle, comme les éléments marquant la présence d'une maladie (@AnnonceurMaladie = {maladie, syndrome, pathologie …}) ou ceux marquant la présence d'un symptôme (@AnnonceurSymptome = {signe, symptôme …}).

Enfin, puisqu'il est impossible de disposer d'une ressource complète consacrée au domaine de la santé, nous utilisons cette même procédure pour augmenter la couverture des entités médicales grâce notamment à l'identification de certains termes médicaux composés. C'est le cas par exemple pour les noms de maladies dont un nombre considérable peut inclure des termes, fréquemment employés, tels que : bénin, malin, etc. Par exemple, la règle :

@maladies **: :** {0-1} (aigu| aiguë) **:** MALADIE[32]

où :

@maladies contient tous les mots simples faisant référence à des noms de maladies.

permet de repérer dans les textes des expressions spécifiant des noms de maladies telles que : « pleurésie aiguë » ou encore « otite moyenne aiguë ».

Ainsi, nous avons constitué un ensemble de 153 règles de reconnaissance (cf. Tableau 3.2, page 111). À titre de comparaison, le système de Xerox utilise un ensemble de plus de 250 règles manuelles pour identifier des entités biologiques.

## 3.3 Extraction de relations sémantiques

Plus encore que l'identification de concepts, l'extraction de relations sémantiques à partir de textes se situe au carrefour de nombreux champs de recherche : extraction d'information,

[31] [ ] permet de spécifier la non appartenance d'un élément à l'entité reconnue.
[32] { } permet de limiter le nombre de mots minimum et maximum entre deux expressions.

sémantique lexicale, construction d'ontologies, terminologie. Ces différents champs ont en commun l'objectif de formaliser et d'exploiter le contenu des documents d'un domaine en construisant des modèles fondés sur les connaissances qu'ils contiennent. Le but ici est donc d'identifier les termes propres au domaine et leurs sens à travers les relations sémantiques intervenant entre ces termes.

Une relation sémantique se définit comme une liaison entre deux ou plusieurs types sémantiques, généralement de classes différentes. Elle permet de structurer un lexique et caractérise le lien sémantique existant entre différents termes (Skuce et al., 1991). Ainsi, les relations sémantiques permettent de déterminer le sens d'une unité lexicale au travers de l'ensemble des relations qui l'associent à d'autres unités (Cruse, 1986). Les relations ayant à la fois une même structure et une même signification sont regroupées dans des types de relations. La plupart des relations sont dites « binaires », c'est-à-dire ne faisant intervenir que deux concepts. L'ordre de ces concepts dans la relation est très important et significatif.

L'extraction de relations sémantiques, que ce soit en domaine ouvert ou restreint, a fait l'objet de nombreux travaux de recherche du fait de son intérêt majeur pour construire et structurer des bases de connaissances lexicales. Elle a comme préalable l'identification dans les textes des entités qui sont liés par la relation à extraire (Giuliano et al., 2007). La plupart des travaux réalisés se limitent à l'extraction de relations pouvant intervenir entre deux concepts sémantiques de classes différentes. Dans ce cas, la relation peut-être déterminée implicitement en identifiant les co-occurrences des concepts. D'autres travaux vont plus loin en abordant les cas d'ambiguïté, c'est-à-dire en explorant les différentes relations possibles entre deux concepts : par exemple les relations « contre-indiqué » et « traite » entre l'entité *Maladie* et l'entité *Traitement* (Rosario et al., 2004) ou encore des relations associant deux mêmes concepts comme illustrés dans (Bunescu et al., 2005 ; Ramani et al., 2005) sur les relations existantes entre deux protéines.

Il existe plusieurs types de relations sémantiques, regroupés en deux grandes familles : les relations paradigmatiques et les relations syntagmatiques (Cruse, 1986) :

**- Les relations paradigmatiques** sont des relations opérant principalement sur des concepts de même catégorie. Elles sont considérées comme des relations « non-prédicatives »

puisqu'elles n'apparaissent pas sous forme de lien syntaxique standard au sein des textes. Ainsi, ce type de relation est généralement représenté par des relations hiérarchiques (Condamines et al., 1993), appelées liens verticaux, qui permettent d'organiser les concepts en arborescence que l'on retrouve dans les thésaurus par exemple. Parmi ce type de relation, on peut citer les relations d'antonymie[33], de synonymie et d'hyperonymie (relation sorte-de).

**- Les relations syntagmatiques** sont des liens sémantiques intervenant entre deux unités linguistiques présentes dans une expression. À l'opposé des relations paradigmatiques, ces relations sont identifiables grâce à l'étude des formes syntaxiques dans les textes. Elles sont déterminées dans le texte par un prédicat. Celui-ci prend souvent une forme verbale, auquel cas les arguments de la relation s'identifient avec les arguments du verbe. Ce prédicat n'est cependant pas toujours explicite. Par exemple, on peut citer des relations spécifiques telles que : « X effet de Y » ou « X pour détecter Y ».

Nous présentons dans cette section les différentes familles de travaux portant sur l'acquisition des relations sémantiques à partir de corpus. Ensuite, nous exposons la méthodologie que nous avons utilisée pour induire des patrons linguistiques (schémas lexico-syntaxiques) propres à chaque relation traitée (cf. Section 2.3.2). Enfin, nous décrivons comment ces patrons linguistiques appris sont utilisés par la suite pour identifier de nouvelles relations sémantiques.

### 3.3.1 Travaux existants sur l'extraction de relations sémantiques

Il existe différentes méthodes d'acquisition de relations sémantiques entre termes à partir de textes. La majorité d'entre elles se fonde sur les occurrences des entités et sur les propriétés sémantiques qui leur sont associées. Cependant, on peut distinguer deux grandes approches : une approche à base de schémas lexico-syntaxiques, exploitant les formules linguistiques caractérisant une relation sémantique ; une approche distributionnelle, fondée sur l'analyse des propriétés contextuelles de chaque mot du texte.

[33] Mot dont le sens est opposé à celui d'un autre.

L'extraction de relations sémantiques à partir de corpus spécialisés implique principalement des entités sémantiques appropriées à la spécialité. Cette particularité engendre des relations sémantiques plus spécifiques (dites syntagmatiques) entre les concepts. Dans le domaine de la santé, les relations sémantiques concernent les liens sémantiques intervenant entre des entités caractéristiques du domaine médical, telles que les maladies, les médicaments ou les examens cliniques. Différents travaux ont déjà été menés concernant l'extraction de relations sémantiques dans le domaine médical ou biomédical, travaux parmi lesquels on peut citer (Craven, 1999), (Pustejovsky et al., 2002b), (Rosario & Hearst, 2004) ou encore (Mukherjea et al., 2006). La spécificité de chacun de ces travaux est illustrée à la Section 3.4.2. Les recherches menées en extraction d'information dans ce même contexte, bien qu'ayant *a priori* une finalité plus large, se ramènent dans bon nombre de cas à l'extraction de ce même type de relations, à l'instar de la détection des interactions entre gènes ou entre gènes et protéines. On se reportera à (Nédellec, 2004) pour un panorama de ces travaux, souvent fondés sur des règles d'extraction définies manuellement.

Dans la suite de cette partie, nous détaillons les deux grandes approches portant sur l'acquisition de relations sémantiques évoquées ci-dessus, à savoir l'approche à base de patrons lexico-syntaxiques et l'approche à base d'analyse distributionnelle.

#### 3.3.1.1 Approche à base de patrons lexico-syntaxiques

L'approche à base de patrons lexico-syntaxiques est l'une des méthodes les plus utilisées pour l'extraction de relations sémantiques. L'idée principale de cette approche est dans un premier temps de synthétiser, à partir d'un texte, les marqueurs caractéristiques d'une relation sémantique (ex. hyponymie) sous la forme de patrons linguistiques afin de les projeter dans un second temps pour extraire de nouvelles relations, c'est-à-dire identifier de nouveaux couples de termes correspondant à la relation spécifiée. Cette méthodologie a été initiée par M. Hearst qui, dans (Hearst, 1992), propose un processus itératif visant à apprendre des schémas lexico-syntaxiques à partir de textes. Ce processus se compose de cinq étapes :

1- Sélectionner une relation cible « R » pour laquelle on désire apprendre des patrons lexico-syntaxiques;

2- Fournir un ensemble d'exemples constitué de couple de termes respectant la relation sémantique spécifiée à l'étape précédente. Cette liste peut être définie manuellement ou extraite à partir d'un thésaurus ou d'une base de connaissances ;
3- Extraire des textes toutes les phrases contenant les couples de termes puis enregistrer leur contexte lexical et syntaxique ;
4- Trouver un environnement commun entre ces contextes. Cet environnement forme un schéma lexico-syntaxique ;
5- Utiliser les schémas identifiés pour extraire de nouveaux couples de termes et revenir à l'étape 3 du processus.

Cette technique est à la base de nombreux travaux sur l'acquisition de relations lexico-syntaxiques. Elle permet d'identifier des motifs d'extraction caractéristiques d'une relation choisie. Les résultats produits par la méthode montrent une certaine pertinence pour la relation d'hyponymie. Pour initier la démarche, Hearst s'est appuyée sur le réseau sémantique WordNet pour composer des couples de termes en relation d'hyponymie. Selon l'auteur, les résultats obtenus en appliquant cette approche à d'autres types de relations comme la relation de méronymie (Girju et al., 2006) sont moins encourageants du fait de la généralité des patrons linguistiques appris.

La méthodologie proposée par Hearst a fait l'objet de multiples travaux de recherche en traitement automatique des langues. Ces travaux tendent majoritairement à automatiser certaines étapes du processus en adoptant une approche partiellement ou complètement automatisée (approche non supervisée) selon les cas. La plupart des travaux réalisés proposent ainsi des techniques différentes dans le but d'automatiser la phase 4 du processus, entièrement manuelle dans (Hearst, 1992), c'est-à-dire l'observation des séquences en corpus correspondant à une relation spécifiée puis leur généralisation en schémas lexico-syntaxiques. Les patrons linguistiques extraits sont ensuite validés automatiquement. Dans cet esprit, Morin propose, avec son système Prométhée, une méthode d'acquisition automatique de relations sémantiques entre termes fondée sur l'étude des cooccurrences (Morin, 1999). Le système Prométhée extrait dans un premier temps, à partir de corpus de textes techniques, les contextes d'occurrences des termes avant de les analyser pour repérer des schémas lexico-syntaxiques similaires. Pour ce faire, il repose sur un calcul de similarité entre chaque paire de contextes lexico-syntaxiques. Cette technique permet au système de regrouper dans des

classes des expressions lexico-syntaxiques partageant des similarités. Ainsi, pour chaque classe, un patron linguistique candidat est sélectionné pour représenter une relation, patron qui est ensuite appliqué pour extraire de nouvelles relations. Cette généralisation automatique présente toutefois quelques limites, dues principalement d'une part, à la liste importante de couples de termes reliés par une relation dont doit disposer le système Prométhée pour être performant et d'autres part, à la fréquence d'apparition du schéma qui doit ressortir plusieurs fois pour être choisi.

De nombreux travaux de recherche s'intéressent également à l'acquisition de schémas lexico-syntaxiques caractérisant des relations sémantiques plus spécifiques entre les termes. Parmi ces recherches, on peut noter les travaux de (Rebeyrolle, 2000) et (Pearson, 1998) qui construisent des patrons lexico-syntaxiques sous la forme d'expressions portant sur les définitions des termes dans le but de repérer les énoncés définitoires dans un corpus de textes. (Malaisé et al., 2004) s'appuie sur le même principe des énoncés définitoires pour construire une ontologie. Ce type de relation (définition) est notamment utilisé par de nombreuses applications telles que les systèmes de question-réponse (Cui et al., 2005 ; Besançon et al., 2006), offrant aux systèmes la compétence nécessaire pour répondre à des questions définitoires du type « Qui est X ? » ou « Qu'est-ce que X ? ». Dans le même registre, (Ravichandran et al., 2002) propose une approche automatique d'acquisition de patrons lexico-syntaxiques simple et performante en vue d'extraire des réponses candidates dans un système de question-réponse. Ces patrons sont spécifiques des types de réponse attendus par les questions (comme les dates de naissance par exemple). L'approche consiste à fournir, dans un premier temps, des exemples (couples de termes) correspondant au type de la question pour lequel on désire acquérir des motifs d'extraction. Ensuite, une interrogation du Web est réalisée pour récupérer un ensemble important de sous-phrases contenant les couples de termes spécifiés. Dans un second temps, des séquences sont généralisées à partir des phrases sélectionnées. On substitue ensuite dans les séquences généralisées (schémas lexico-syntaxiques) les termes par leur type, c'est-à-dire que l'on remplace l'objet de la question par <*NAME*> et la réponse par <*ANSWER*>. Enfin, la dernière étape de cette méthode repose sur le calcul d'un score de précision pour chaque patron extrait.

Dans le domaine biomédical, plusieurs travaux s'inscrivent dans la perspective de l'acquisition de schémas spécifiques exprimant des relations entre des concepts sémantiques

du domaine. Un nombre important de ces études se concentrent sur l'étude des relations relevant du domaine de la génomique, et plus particulièrement sur l'identification des interactions entre gènes et protéines. Cependant, la plupart de ces travaux se fondent essentiellement sur des patrons lexico-syntaxiques construits manuellement. Ainsi, (Ng et al., 1999) utilise un ensemble de règles d'extraction constituées à la main sous la forme de patrons linguistiques pour spécifier des relations d'interactions entre protéines. Le système proposé par (Blaschke et al., 1999) permet également d'extraire des relations portant sur les interactions entre protéines. Pour ce faire, il se fonde sur un ensemble prédéfini de protéines et sur une liste composée de 14 motifs d'extraction. Dans (Khoo et al., 2000), l'idée est d'utiliser des patrons pour identifier et extraire des relations de causalité (relation Cause-Effect) à partir de résumés de la base médicale MEDLINE, ce qui se traduit par un repérage dans les textes des expressions exprimant une relation de causalité entre deux unités lexicales, par exemple les passages du type « *A à cause de B* » ou « *A est un effet de B* ». Plus récemment, (Rosario, 2005) s'est intéressée à l'extraction des différents types de relations intervenant entre les classes sémantiques maladie et traitement (l'étude s'est focalisée sur huit types de relations ).

#### 3.3.1.2 Approche à base d'analyse distributionnelle

L'approche à base d'analyse distributionnelle est l'autre grande technique exploitée pour extraire des relations sémantiques. Elle est classiquement utilisée par de nombreuses applications pour la structuration des termes d'un corpus afin de construire des bases de connaissances terminologiques ou ontologiques (Habert et al., 1996 ; Bourigault, 2002). Cette approche, qui se fonde essentiellement sur le principe présenté dans (Harris, 1968), s'appuie sur l'analyse des propriétés contextuelles des mots d'un corpus qui permet de regrouper tous les mots partageant les mêmes propriétés dans des classes de concepts afin de proposer des relations sémantiques intervenant entre ces concepts, plus précisément regrouper les concepts appartenant à une même classe. Le regroupement de ces termes s'avère très efficace pour la construction de modèles de connaissances à partir de textes spécialisés.

La majorité des travaux utilisant l'analyse distributionnelle repose sur un processus composé de trois phases, comme illustré dans (Grefenstette, 1994) : rechercher les caractéristiques contextuelles de chaque mot présent dans le texte ; collecter les mots partageant les mêmes contextes syntaxiques ; enfin, construire les classes à partir des mots sélectionnés à l'étape

précédente. Par exemple, à partir d'un corpus médical, un outil d'analyse distributionnelle rapprochera les termes échographie, radiographie et mammographie, car chacun fonctionne comme sujet des verbes montrer, détecter et confirmer et complément d'objet des verbes effectuer, prescrire et réaliser. Partant de cette méthodologie, plusieurs travaux se sont intéressés à l'étude des propriétés contextuelles des mots dans les corpus en vue de déterminer les dépendances syntaxiques entre mots avant de proposer les relations sémantiques, généralement de type paradigmatique, pouvant les associer. Certains d'entre eux se fondent sur les fréquences de cooccurrences des mots. Dans le prolongement de ces travaux, on peut citer pour l'anglais le système SEXTANT développé par Grefenstette (Grefenstette, 1992). SEXTANT généralise des classes de mots caractérisés par des dépendances identiques. Pour ce faire, il exploite la distribution des contextes syntaxiques de type Nom-Nom, Nom-Verbe ou encore Nom-Adjectif.

Pour le français, l'analyse distributionnelle a été mise en œuvre par des systèmes tels que LEXICLASS (Assadi, 1998) ou encore ZELLIG (Habert et al., 1996), qui varient selon les contextes syntaxiques étudiés. Assadi (Assadi, 1998) a développé le système LEXICLASS, un outil de classification des syntagmes nominaux extraits par le logiciel LEXTER (Bourigault, 1994), à partir d'un document technique selon leur contexte terminologique aidant ainsi le cogniticien dans la phase d'analyse conceptuelle. Cette classification se fonde sur un regroupement des têtes syntaxiques partageant les mêmes expansions. L'ensemble des syntagmes nominaux est centralisé suivant la distribution de leurs contextes adjectivaux. Enfin, de manière similaire, (Habert et al., 1996) présente un outil d'analyse de textes, ZELLIG, qui exploite les relations intervenant entre les composants au sein des syntagmes nominaux, plus précisément les relations de dépendance entre les têtes et leurs expansions dans les syntagmes. Pour collecter les différents syntagmes nominaux d'un corpus, ZELLIG utilise un extracteur tel que LEXTER. Cette approche forme des classes de noms selon leur distribution syntaxique dans les groupes nominaux. Pour chaque mot, deux classes de contextes sont constituées concernant respectivement son contexte précédent et son contexte suivant. Cependant, selon les auteurs, la proximité conceptuelle entre deux mots repose sur le nombre de contextes partagés par ces mots.

Dans un domaine technique comme le domaine biomédical, on recense également de nombreux travaux, portant majoritairement sur le domaine de la génomique et se fondant sur

les propriétés contextuelles des termes dans les textes (Nazarenko et al., 1997). L'objectif est généralement d'extraire les informations sur les interactions génétiques, c'est-à-dire les gènes impliqués dans un phénomène particulier. Ainsi, le système Bibliometrics (Stapley et al., 2000) s'appuie sur la fréquence d'apparition des gènes dans un même document. Si la fréquence entre deux gènes est signifiante, les gènes sont nécessairement en relation. Le système peut aussi déterminer la nature des relations existantes entre les gènes. Les auteurs soulignent cependant que le type de la relation intervenant entre un couple de gènes est implicitement représenté graphiquement. Dans (Stephens et al., 2001), le système proposé se fonde sur des statistiques de cooccurrence pour repérer les relations intervenant entre les gènes. Chaque couple de gènes est regroupé selon une liste de descripteurs prédéfinie correspondant à des relations. Un graphe est ensuite construit automatiquement où les nœuds représentent des gènes et les branches les relations de cooccurrence. La longueur d'une branche est déterminée en fonction de la probabilité de la présence du couple de gènes dans les mêmes documents.

#### 3.3.1.3 Synthèse

Les différentes techniques que nous avons présentées en acquisition de relations lexicales et sémantiques montrent des approches très diverses et des résultats relativement satisfaisants qui permettent de couvrir les connaissances d'un domaine. Les méthodes utilisées peuvent être classées selon deux grandes familles : les approches exploitant l'aspect structurel des données textuelles et celles exploitant leur aspect numérique. Les approches numériques se révèlent pertinentes pour inférer des classes sémantiques et sont généralement faciles à mettre en œuvre étant donné qu'elles ne requièrent pas de connaissances préalables sur le domaine étudié ; elles ne reposent sur aucune donnée autre que le corpus. Toutefois, il est parfois très difficile de déduire la relation sémantique existant entre les termes au-delà d'une notion de proximité sémantique issue de la classification de ces termes. L'interprétation des classes sémantiques et des relations extraites nécessite donc un investissement humain afin de valider les informations identifiées. Contrairement à ces techniques, les connaissances extraites par les méthodes structurelles s'avèrent plus facilement interprétables. Les patrons d'extraction utilisés au sein de ces méthodes permettent de déterminer la nature exacte des connaissances extraites puisqu'ils sont supposés caractéristiques de ces connaissances. Néanmoins, les approches à base de patrons lexico-syntaxiques exigent des connaissances préalables pour l'apprentissage des schémas d'extraction et, dans le cas où ces schémas sont appris

automatiquement, un ensemble d'exemples d'apprentissage par rapport aux relations sémantiques désirées. Ces exemples sont principalement fournis par un expert du domaine. L'utilisation des patrons linguistiques se révèle une méthode robuste et très utile pour la construction de bases de connaissances à partir de corpus techniques. Les travaux exploitant des motifs d'extraction pour repérer les relations sémantiques dans les textes montrent que les patrons induits peuvent être utilisés de différentes manières suivant les besoins attendus de l'application. De ce fait, leur niveau de généralisation peut varier selon le degré de la précision ou du rappel souhaité. Certaines applications privilégient le rappel au détriment de la précision pour acquérir un nombre plus important de relations sémantiques.

La méthode que nous proposons dans le cadre de ce travail s'inscrit dans la même perspective que les approches structurelles. Elle repose pour sa part sur l'identification puis l'application de patrons linguistiques caractérisant les relations visées (cf. Section 2.3.2), dans le prolongement direct de (Pantel et al., 2004). Cette application se déroule en deux étapes (Embarek et al., 2007). La première consiste à identifier dans les textes les entités du domaine médical intervenant dans les relations étudiées. Dans la phrase « …en novembre 2001, année d'un cancer de la prostate traité par radiothérapie et qu'il affirme aujourd'hui disparu, … », le premier objectif est ainsi de repérer que « *cancer de la prostate* » est une maladie et que « *radiothérapie* » est un traitement. Dans un second temps, l'application du patron « <*maladie*> traité par <*traitement*> » construit automatiquement à partir d'un corpus de référence permet de valider la présence d'une relation entre ces deux entités, relation stipulant dans le cas présent que la *radiothérapie* est un traitement possible du *cancer de la prostate*. L'utilisation des patrons doit contribuer au peuplement de notre ontologie médicale et ainsi garantir au système de question-réponse la compétence nécessaire pour trouver les réponses candidates aux questions.

La section suivante présente plus en détail la méthodologie que nous avons utilisée pour apprendre les patrons d'extraction de relations sémantiques à partir de corpus médicaux. Cette approche se fonde essentiellement sur l'algorithme d'extraction de patrons multi-niveaux explicité dans (Pantel et al., 2004). Par la suite, nous présentons également le processus d'application de ces patrons pour extraire de nouvelles relations.

### 3.3.2 Apprentissage de patrons lexico-syntaxiques

Nous présentons dans cette section l'approche utilisée pour apprendre des patrons lexico-syntaxiques.

#### 3.3.2.1 Principe

Le terme de patron linguistique désigne dans le cas présent un schéma lexico-syntaxique spécifique d'une relation intervenant entre deux entités médicales. Ces patrons sont dits multi-niveaux, c'est-à-dire qu'ils s'appuient sur des informations provenant de plusieurs niveaux de traitement des textes. À l'instar des règles de reconnaissance des entités médicales, ils peuvent ainsi faire intervenir la forme fléchie des mots, leur forme normalisée ou bien encore leur catégorie morpho-syntaxique. Le processus (présenté aussi à la Figure 3.1) que nous avons élaboré pour extraire à partir d'un corpus les patrons linguistiques caractérisant une relation est le suivant :

1- appliquer sur le corpus considéré les règles de reconnaissance des entités médicales impliquées dans la relation cible. Nous prendrons à titre d'exemple la relation « *Traite* » entre une *Maladie* et un *Traitement* ;
2- extraire du corpus toutes les phrases contenant les deux entités de la relation cible, à savoir ici les phrases contenant à la fois une maladie et un traitement ;
3- sélectionner manuellement les phrases dans lesquelles la relation entre les deux entités correspond effectivement à la relation cible. Cela implique en particulier d'écarter les phrases telles que « la <maladie> n'est pas traitée par le <traitement> » ;
4- réaliser l'analyse linguistique de chaque phrase sélectionnée pour faire apparaître les différents niveaux d'information. Cette analyse est réalisée comme pour la reconnaissance des entités par l'analyseur LIMA ;
5- remplacer dans chaque phrase les entités par leur type ;
6- appliquer l'algorithme d'extraction de patrons multi-niveaux (voir Figure 3.2) entre chaque couple de phrases parmi celles sélectionnées précédemment.

Pour extraire les patrons linguistiques propres à chaque relation sémantique traitée (cf. Section 2.3.2), nous faisons appel à l'algorithme proposé par (Pantel et al., 2004) (voir Figure

3.2) pour apprendre des patrons multi-niveaux. Cet algorithme est composé de deux parties. La première consiste à calculer la distance d'édition minimale entre deux phrases, ce qui permet de déterminer le nombre minimum d'opérations (insertion, suppression et remplacement) à appliquer pour passer d'une phrase à l'autre. La deuxième étape extrait le patron multi-niveau le plus spécifique permettant de généraliser les deux phrases. Enfin, pour compléter certains alignements, deux opérateurs génériques classiques sont introduits : (*s*), qui représente 0 ou 1 instance de n'importe quel mot (présence facultative) et (*g*), qui représente exactement une instance de n'importe quel mot.

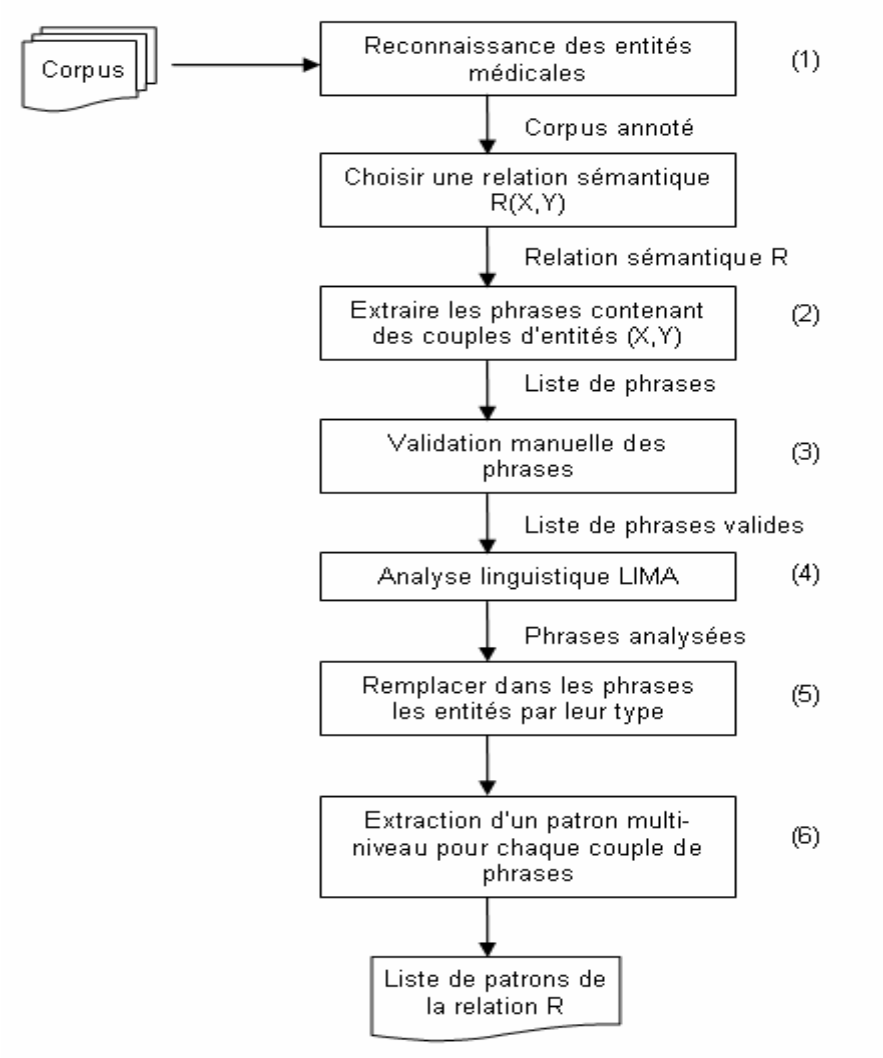


Figure 3.1 Processus d'extraction de patrons multi-niveaux

Dans le cadre de notre travail, nous avons décidé d'éliminer tous les patrons contenant plus de deux opérateurs d'alignement (*s*) et (*g*), c'est-à-dire éviter des patrons tels que « X (*s*) (*g*) (*s*) (*g*) (*s*) Y ». Cette décision a pour but d'améliorer la pertinence et

l'expressivité des patrons appris. Les patrons linguistiques de chaque relation sont ensuite classés selon leur fréquence d'apparition pour ne retenir que les N premiers patrons. Pour notre étude, nous avons fixé le seuil N à 50 afin de limiter le degré de spécificité des patrons construits (voir Tableau 3.1). Il s'agit dans ce cas d'éliminer les patrons les plus spécifiques issus d'une généralisation parfaite de deux phrases avec une fréquence d'apparition égale à un.

**Algorithme pour le calcul de la distance d'édition minimale**

```
Soit A et B sont deux phrases composées de n et m mots,
D[0,0] = 0
Pour i = 1 jusqu'à n faire D[i,0] = D[i-1,0] + coût (insertion)
Pour j = 1 jusqu'à m faire D[0,j] = D[0,j-1] + coût (suppression)
Pour i = 1 jusqu'à n faire
   Pour j = 1 jusqu'à m faire
   D[i,j] = min (D[i-1,j-1] + coût (remplacement),
          D[i-1,j] + coût (insertion),
          D[i,j-1] + coût (suppression))
Afficher (D[n,m])

Où : D[n,m] est la distance d'édition entre ces deux phrases
```

**Algorithme pour l'extraction du patron multi-niveau optimal**

```
i = n, j = m
Tant que i ≠ 0 et j ≠ 0
   Si D[i,j] = D[i-1,j] + coût (insertion)
      Afficher (*s*), i = i-1
   Sinon Si D[i,j] = D[i,j-1] + coût (suppression)
      Afficher (*s*), j = j-1
   Sinon Si A1i = B1j
      Afficher (A1i), i = i-1, j= j-1
   Sinon Si A2i = B2j
      Afficher (A2i), i = i-1, j= j-1
   Sinon Si A3i = B3j
      Afficher (A3i), i = i-1, j= j-1
   Sinon
      Afficher (*g*), i = i-1, j= j-1
Où :
- A1, A2, A3 sont le niveau 1 (forme lexical), niveau 2 (forme lemmatisée) et niveau 3 (catégories grammaticales) de la phrase A
composée de n mots.
- B1, B2, B3 sont le niveau 1 (forme lexical), niveau 2 (forme lemmatisée) et niveau 3 (catégories grammaticales) de la phrase B
composée de m mots.
```

Figure 3.2 Algorithme d'extraction de patrons multi-niveaux (Pantel et al., 2004)

L'algorithme présenté ci-dessus est un algorithme permettant de déterminer l'alignement optimal entre deux séquences en calculant la distance d'édition minimale entre elles. Le calcul de cette distance d'édition (Levenshtein, 1966) s'effectue au moyen de la programmation dynamique. L'algorithme est composé de deux parties. La première partie consiste à calculer le nombre minimum d'opérations d'édition pour passer d'une séquence à l'autre alors que la seconde partie de l'algorithme produit l'alignement optimal.

À noter que les patrons construits par ce processus sont assez spécifiques puisqu'ils sont issus de la généralisation de deux phrases. En effet, dans notre expérience, on ne procède pas à la généralisation des patrons induits. Ce choix est motivé par l'importance que nous accordons à la précision au détriment du rappel. Le but ici est d'utiliser les patrons lexico-syntaxiques pour maximiser l'extraction de nouvelles relations valides. C'est d'ailleurs pour cette raison que les patrons sont appris uniquement à partir de phrases dans lesquelles la présence de la relation visée a été validée manuellement (étape 3 du processus). Cette tâche peut s'avérer fastidieuse, surtout si la taille du corpus étudié est importante ou le nombre de phrases extraites contenant le couple d'entités médicales impliquées dans la relation cible est élevé. Néanmoins, il est possible d'utiliser des expressions pour éliminer automatiquement les phrases qui ne correspondent pas à la relation cible telles que : « intraitable », « déconseillé », etc. Enfin, l'aspect multi-niveau des patrons est exploité pour repérer dans les textes les relations exprimées par des termes différents mais partageant la même forme normalisée ou la même catégorie morpho-syntaxique. Par exemple, le patron « X pour *VERBE_PRINC_INFINIT* la maladie Y » peut être utilisé pour identifier des expressions comme « X pour *guérir* la maladie Y » ou « X pour *traiter* la maladie Y ».

#### 3.3.2.2 Résultats

Nous avons appliqué le processus (cf. Figure 3.1) décrit plus haut sur la totalité du corpus médical (16 millions de mots) de la campagne d'évaluation des systèmes de question-réponse EQueR (cf. Section 6.1.1). Nous avons extrait ainsi des patrons multi-niveaux pour les quatre relations considérées dans cette étude. Nous donnons, à titre illustratif, quelques exemples de patrons extraits pour chaque relation :

**Maladie – Examen (Relation Détecte)**
<examen> en suspicion de <maladie>
<examen> pour le NC_GEN[34] (*g*) <maladie>
<examen> (*g*) le diagnostic (*g*) <maladie>
<maladie>, (*s*) <examen>

[34] NC_GEN : Nom commun général.

**Maladie – Traitement (Relation Traite)**

<traitement> dans le traitement des <maladie>

<traitement> être (*g*) PREP_GENERAL[35] le traitement de le (*s*) <maladie>

<traitement> est recommandé pour le traitement des <maladie>

<traitement> contre le <maladie>

**Maladie – Symptôme (Relation Signe)**

<maladie>, se manifeste par une <symptome>

<symptome> VERBE_PRINC_INFINIT[36] la NC_GEN de le (*s*) <maladie>

<maladie> (*g*) avec <symptome>

<symptome> ( <maladie>

**Maladie – Médicament (Relation Soigne)**

<medicament> est indiqué dans le traitement de la <maladie>

<medicament>, utilisée (*s*) (*s*) dans le traitement de <maladie>

<medicament> est un médicament utilisé pour traiter <maladie>

<maladie> chez les NC_GEN traité par <medicament>

Les exemples donnés ci-dessus montrent que les patrons multi-niveaux construits peuvent être classés selon trois catégories. La première catégorie concerne les patrons linguistiques contenant uniquement la forme fléchie des mots. Cela s'explique par le fait que ces patrons sont généralisés à partir de phrases exprimées par des termes similaires pour illustrer une relation sémantique intervenant entre des instances différentes des deux concepts concernés par la relation. La deuxième catégorie porte sur les patrons regroupant des opérateurs ((*s*), (*g*)) et des informations provenant de plusieurs niveaux de traitement des textes. Dans ce cas, la généralisation est déterminée entre des couples de phrases composées de termes partageant les mêmes catégories morpho-syntaxiques ou les mêmes formes normalisées. Enfin, la dernière catégorie regroupe les phrases comprenant simplement les opérateurs d'alignement. Cette catégorie est le résultat d'un alignement entre des couples de phrases qui ne partagent aucun niveau de traitement des textes.

[35] PREP_GENERAL : préposition générale.
[36] VERBE_PRINC_INFINITIF : infinitif du verbe principal.

Les patrons multi-niveaux sont construits pour chaque type de relation à partir d'un ensemble de phrases exprimant une relation valide entre les concepts concernés. Pour ce faire, on procède à une sélection manuelle des phrases pour éliminer les phrases n'abritant pas la relation recherchée. Les phrases non retenues sont considérées comme non valides, c'est-à-dire qu'elles n'expriment pas la relation cible. Sont ainsi écartées les phrases suivantes : « la chimiothérapie pour le traitement adjuvant des cancers de l'œsophage n'est pas indiquée dans l'état actuel des connaissances », « le tamoxifène augmente le risque du cancer de l'endomètre », etc. Le Tableau 3.1 récapitule l'évaluation réalisée sur le pourcentage de phrases retenues pour chaque type de relation pour l'acquisition de patrons linguistiques multi-niveaux. Les chiffres montrent qu'en moyenne 79% des phrases extraites du corpus sont retenues.

| **Patrons de relations** | **Nombre de phrases extraites** | **% de phrases retenues** | **Nombre de patrons appris** |
|---|---|---|---|
| Maladie – Examen | 315 | 81% | 131 |
| Maladie – Médicament | 182 | 75% | 68 |
| Maladie – Traitement | 255 | 84% | 142 |
| Maladie – Symptôme | 201 | 76% | 102 |
| **Moyenne** | **238** | **79%** | **111** |

Tableau 3.1 Statistiques sur la sélection automatique / manuelle des phrases exemples

Il est intéressant de noter qu'en procédant à une évaluation sans tenir compte de l'étape concernant la sélection des phrases contenant la relation cible (étape 3 du processus), nous avons obtenu pratiquement 80% des patrons initiaux et plus de 87% en moyenne des patrons retenus (sur les 50 patrons sélectionnés). Ces chiffres montrent que les phrases non retenues pour la construction des patrons lexico-syntaxiques n'influent pas trop sur le résultat final, à savoir sur la liste des patrons pouvant être appris.

### 3.3.3 Application des patrons appris à l'identification de relations

Pour acquérir de nouvelles relations sémantiques à partir d'un corpus, c'est-à-dire de nouveaux couples d'entités liées par une relation identifiée, nous appliquons une démarche en deux temps. Comme dans le cas de l'extraction de patrons de relations, nous commençons par

sélectionner des relations candidates en repérant les phrases contenant un couple d'entités intervenant dans une des relations cibles. Dans un second temps, nous confrontons la phrase contenant la relation candidate avec les patrons linguistiques spécifiques de cette relation. Si l'un au moins de ces patrons peut s'appliquer à la phrase considérée, la relation est considérée comme valide. Dans le cas contraire, elle est écartée. Plus formellement, le processus mis en œuvre pour un type de relation est le suivant :

1- appliquer sur le corpus considéré les règles de reconnaissance des entités médicales impliquées dans la relation cible ;
2- extraire du corpus toutes les phrases contenant simultanément les deux entités de la relation cible ;
3- réaliser l'analyse linguistique de chaque phrase sélectionnée en utilisant l'analyseur LIMA ;
4- remplacer dans chaque phrase les entités par leur type ;
5- pour chaque phrase, calculer sa distance d'édition (cf. Section 3.3.2) avec tous les patrons multi-niveaux de la relation. Si la distance d'édition est égale à 0, c'est-à-dire si la relation entre les deux types sémantiques de la phrase respecte strictement le schéma du patron, alors la relation est validée. La distance d'édition a été fixée dans un premier temps à 0 pour déterminer une identité parfaite entre la phrase réponse et un des patrons multi-niveaux de la relation afin de privilégier la précision. Cependant, l'utilisation de cette distance va permettre dans un second temps d'étudier les contextes précédent et suivant de la phrase réponse.

Il est à noter que nous avons utilisé ici un critère strict d'appariement entre les phrases sélectionnées et les patrons de relation mais l'utilisation de la distance d'édition pour réaliser cet appariement autoriserait sans changement un appariement plus flou.

Nous avons appliqué notre algorithme d'extraction et de validation de relations sémantiques pour les quatre relations retenues dans notre étude sur un corpus de textes médicaux recueillis dans le cadre du projet Technolangue Atonant[37], corpus différent de celui utilisé pour induire

[37] Projet Technolangue portant sur l'enrichissement semi-automatique d'ontologies.

les patrons de caractérisation de ces relations. Voici quelques exemples de relations détectées par notre méthode (le patron utilisé est introduit par =>) :

**Maladie – Examen (Relation Détecte)**

…*tomodensitométrie* dans le diagnostic des *tumeurs du médiastin*

=> <examen> dans le diagnostic (*s*) <maladie>

…*radiographie pulmonaire* pour le diagnostic de *tuberculose*

=> <examen> (*g*) le diagnostic (*g*) <maladie>

**Maladie – Symptôme (Relation Signe)**

…L'*intoxication* peut provoquer des *vomissements*

=><maladie> peut VERBE_PRINC_INFINIT DET_ART (*s*) <symptome>[38]

…*Botulisme*, se manifeste par une *sécheresse de la bouche*

=> <maladie>, se manifeste (*s*) par une <symptome>

**Maladie – Médicament (Relation Soigne)**

…*insuffisance rénale chronique* traitée par *Eprex*

=> <maladie> traitée par <medicament>

…*vaccin* utilisé pour prévenir la *fièvre aphteuse*

=> <medicament> utilisé pour VERBE_PRINC_INFINIT (*g*) <maladie>

**Maladie – Traitement (Relation Traite)**

…*chimioprophylaxie* contre la *malaria*

=> <traitement> contre la <maladie>

…*radiothérapie* dans le traitement de la *resténose*

=> <traitement> dans le traitement de la <maladie>

Les exemples donnés ici caractérisent l'intérêt d'utiliser des patrons multi-niveaux. Ces patrons donnent la possibilité d'identifier dans les textes des relations respectant le schéma du patron mais formulées avec des termes différents, comme illustré par la relation « *vaccin* utilisé pour prévenir la *fièvre aphteuse* » extraite par le patron « <medicament> utilisé pour VERBE_PRINC_INFINIT (*g*) <maladie> ». De plus, les patrons construits à partir d'une généralisation stricte entre deux phrases contenant une relation valide entre deux concepts

[38] VERBE_PRINC_INFINITIF : infinitif du verbe principal ; DET_ART : article

médicaux permettent automatiquement d'extraire des relations candidates valides. Par exemple, le patron « <traitement> dans le traitement de la <maladie> » permet de repérer la relation *« radiothérapie* dans le traitement de la *resténose »*.

## 3.4 Évaluation

Dans cette section, nous présentons successivement les résultats des évaluations menées sur deux corpus en français, constitués chacun d'articles scientifiques et de recommandations de bonne pratique médicale téléchargées à partir du site du CISMeF. La première (cf. Tableau 3.3) concerne l'identification des entités médicales dans les textes en appliquant les règles de reconnaissance présentées à la Section 3.2.2. La seconde (cf. Tableau 3.4) porte sur l'extraction et la validation de relations sémantiques grâce à la méthode présentée à la Section 3.3.3.

### 3.4.1 Évaluation de l'identification de concepts

Le Tableau 3.2 donne le nombre de règles développées pour identifier chaque type d'entités médicales. On notera que globalement, il n'y a pas de grosses disparités dans le nombre de règles nécessaires pour reconnaître ces différents types d'entités. Néanmoins, la complexité des noms de maladie se traduit par un nombre de règles plus important pour ce type sémantique contrairement aux noms de médicament qui peuvent être facilement répertoriés. L'ensemble des règles de reconnaissance est complété par des listes d'entités regroupant des éléments caractéristiques qui permettent d'identifier la présence d'un type sémantique comme les noms de médicaments, les noms des examens cliniques, etc. Ainsi, nous avons construit les listes suivantes : maladie (1516 instances), symptôme (438 instances), traitement (600 instances), examen (836 instances) et enfin médicament (2429 instances).

| Type d'entités | Nombre de règles |
|---|---|
| Maladie | 38 |
| Symptôme | 32 |
| Examen | 27 |
| Traitement | 30 |
| Médicament | 26 |
| **Total** | **153** |

Tableau 3.2 Nombre de règles de reconnaissance développées

Le Tableau 3.3 résume les résultats obtenus en appliquant nos règles de reconnaissance d'entités médicales sur un sous-ensemble sélectionné aléatoirement, d'une taille de 1,5 Mo (soit environ 130 000 mots), du corpus médical de la campagne d'évaluation EQueR. Les mesures utilisées sont classiquement la précision et le rappel, qui se définissent ici de la façon suivante :

$$\text{précision} = \frac{\text{nombre d'entités correctes extraites par notre système}}{\text{nombre total des entités extraites par notre système}}$$

$$\text{rappel} = \frac{\text{nombre d'entités correctes extraites par notre système}}{\text{nombre total des entités présentes dans le corpus}}$$

| Entités médicales | Nombre d'entités[39] | Précision | Rappel | F1-mesure |
|---|---|---|---|---|
| Maladie | 1826 | 0,95 | 0,80 | 0,86 |
| Symptôme | 444 | 0,84 | 0,76 | 0,79 |
| Examen | 226 | 0,94 | 0,93 | 0,93 |
| Traitement | 581 | 0,86 | 0,81 | 0,83 |
| Médicament | 191 | 0,93 | 0,88 | 0,90 |
| **Moyenne** | **654** | **0,90** | **0,84** | **0,86** |

Tableau 3.3 Résultats de la reconnaissance des entités médicales

La F1-mesure, moyenne harmonique entre la précision et le rappel, est utilisée comme mesure synthétique. Ces mesures sont calculées par comparaison avec une annotation manuelle que j'ai faite du corpus d'évaluation. Les résultats de notre méthode, donnés par le Tableau 3.3, montrent une précision et un rappel supérieurs ou égaux à 83% en moyenne, ce qui constitue un bon niveau pour ce type de tâche. Globalement, ils sont comparables aux résultats des meilleurs systèmes de reconnaissance d'entités nommées concernant des concepts très généraux tels que les noms de personnes ou les lieux (conférence CoNLL) : la F1-mesure du meilleur système est de l'ordre de 88% pour l'anglais, 72% pour l'allemand et 81% pour l'allemand. On peut noter en particulier le niveau élevé de la précision qui caractérise un niveau de fiabilité très significatif. Cette propriété est d'autant plus importante dans le cas

[39] Nombre d'entités présentes dans le corpus d'évaluation.

présent que la détection des entités sert ensuite de point de départ à l'extraction des relations. Le rappel pourrait quant à lui être amélioré en étant plus exhaustif dans les listes d'entités constituées. L'analyse des erreurs résultant de l'application des règles de reconnaissance montre que la majorité des erreurs porte sur la reconnaissance partielle des entités médicales. Par exemple, dans le passage : « l'*encéphalopathie de Gayet Wernicke*… », le terme « Wernicke » n'est pas identifié en tant que complément de l'entité maladie « encéphalopathie de Gayet » ou encore dans l'expression suivante : « … *pleurectomie partielle* …», seulement le terme « pleurectomie » est reconnu comme un traitement.

### 3.4.2 Évaluation de l'extraction des relations

Concernant l'extraction et la validation des relations sémantiques, nous avons appliqué la méthode présentée à la Section 3.3.3 sur 65 Mo du corpus utilisé dans le cadre du projet Technolangue Atonant, soit environ 10 millions de mots. Les patrons d'extraction appliqués avaient été préalablement appris à partir de la totalité du corpus médical EQueR, soit environ 16 millions de mots. Contrairement au cas des entités, l'annotation manuelle de référence n'a pas été réalisée en parcourant tout le corpus mais en jugeant de la présence effective d'une des quatre relations cibles parmi les phrases abritant des relations candidates, c'est-à-dire les phrases contenant au moins deux entités compatibles avec des relations cibles. Par conséquent, seule la validation des relations candidates est évaluée ici. Pour les mesures d'évaluation, nous avons à nouveau fait appel à la précision et au rappel, définis comme suit :

$$\text{précision} = \frac{\text{nombre de relations validées correctes}}{\text{nombre total des relations validées par notre système}}$$

$$\text{rappel} = \frac{\text{nombre de relations validées correctes par notre système}}{\text{nombre total de relations annotées dans le corpus}}$$

Comme dans le cas de la reconnaissance des entités, l'extraction et la validation des relations se caractérisent par une forte précision et un rappel un peu moins élevé (cf. Tableau 3.4). Cependant, la différence entre précision et rappel est plus accentuée dans ce cas. On peut donc avancer que les relations produites par la méthode que nous avons proposée sont globalement d'une bonne fiabilité mais que les patrons linguistiques appris sur le corpus médical EQueR

ne couvrent pas toutes les formes par lesquelles les relations cibles se manifestent dans le corpus Atonant.

La bonne précision obtenue par l'application des patrons lexico-syntaxiques construits est incontestablement liée à la spécificité de ces derniers qui sont appris à partir de la généralisation de deux phrases candidates contenant la relation cible. Toutefois, nous avons relevé deux grandes causes d'erreurs. Une première cause concerne le degré de généralité de certains patrons. Il s'agit en effet de l'utilisation des patrons moins spécifiques tels que les patrons regroupant plusieurs opérateurs d'alignement ((*s*) et (*g*)). Ces patrons permettent ainsi de détecter automatiquement des relations sémantiques non valides, *i.e.* des couples d'entités médicales non liées par la relation identifiée. Par exemple, la relation *Soigne* (cf. Figure 2.2) entre le médicament « insuline » et la maladie « acidocétose diabétique » est validée à partir de la phrase suivante « L'**insuline** provoquera un phénomène d'**acidocétose diabétique**. » par le patron « **<medicament>** (*g*) DET_ART (*g*) (*s*) **<maladie>** ». La deuxième cause d'erreurs porte sur l'imperfection de certaines règles de reconnaissance des entités médicales. Plus précisément, il s'agit des relations validées entre des entités médicales incorrectes. Par exemple, dans la phrase « L'antibiothérapie pour traiter les personnes atteintes d'une maladie gastroentérique. », la relation *Traite* (cf. Figure 2.2) entre « antibiothérapie » et « personnes atteintes » a été validée car l'expression « personnes atteintes » a été identifiée comme une maladie en appliquant les règles de reconnaissance. Cette difficulté peut néanmoins être surmontée en améliorant l'écriture des règles de reconnaissance. Le rappel quant à lui est le résultat de l'absence de reconnaissance des entités médicales dans le corpus déclenchant le processus d'extraction. En effet, reconnaître les deux entités impliquées dans la relation cible est nécessaire pour identifier les relations candidates dans les textes et par conséquent, augmente le nombre de relations sémantiques à valider.

| Relations | Précision | Rappel | F1-mesure |
|---|---|---|---|
| Maladie – Examen | 0,92 | 0,63 | 0,74 |
| Maladie – Médicament | 0,91 | 0,59 | 0,71 |
| Maladie – Traitement | 0,92 | 0,69 | 0,78 |
| Maladie – Symptôme | 0,90 | 0,65 | 0,75 |
| **Moyenne** | **0,91** | **0,64** | **0,75** |

Tableau 3.4 Résultats de la validation des relations sémantiques

La comparaison avec d'autres travaux est quant à elle difficile du fait de la diversité des types de relations considérés, des corpus et des approches adoptées. Néanmoins, il est possible de donner quelques éléments de situation. En utilisant des patrons linguistiques élaborés manuellement pour caractériser des relations d'inhibition dans des phrases extraites de la base Medline, (Pustejovsky et al., 2002b) obtient ainsi une précision de 94% et un rappel de 58,9%. Le Tableau 3.4 montre que nous obtenons des résultats globalement comparables en construisant ces patrons linguistiques de manière automatique. Le processus de validation des relations extraites peut également être envisagé sous l'angle de la classification : une relation candidate est alors classée comme pertinente ou non pertinente. C'est l'approche retenue par (Craven, 1999) ou par (Rosario et al., 2004). En utilisant un classifieur bayésien naïf[40] sur des relations candidates de type « *subcellular-location* » (Identité de la protéine – localisation cellulaire de la protéine) extraites de Medline, (Craven, 1999) fait état d'une précision de 78% et d'un rappel de 32%. Dans le cas de (Rosario et al., 2004), le classifieur n'est plus seulement binaire. Il s'agit en effet de discriminer les relations intervenant entre un traitement et une maladie : 8 relations sont ainsi distinguées qui recouvrent la relation *Traite* à laquelle nous nous sommes attaché mais également des relations exprimant qu'un traitement peut prévenir une maladie ou qu'une maladie est un effet secondaire d'un traitement. (Rosario et al., 2004) rapporte les évaluations menées avec plusieurs types de classifieurs et obtient les meilleurs résultats avec un réseau de neurones, la précision étant alors de 96,9%. Il est à noter que ce travail s'appuie sur des ressources plus étendues que les nôtres puisqu'il fait appel à un analyseur syntaxique de surface pour produire les groupes syntaxiques correspondant à la structure des phrases et qu'il exploite également la ressource sémantique que constitue le MeSH. Enfin, (Mukherjea et al., 1999) exploite les sources de données du Web pour apprendre automatiquement des relations sémantiques intervenant entre des entités médicales comme la relation de causalité entre une maladie et une entité biologique. Pour ce faire, ils interrogent des moteurs de recherche avec comme requête des patrons lexico-syntaxiques écrits manuellement en utilisant les termes issus des concepts médicaux de l'UMLS tels que les gènes, les protéines, les vitamines, etc. Ainsi, pour la relation de causalité, (Mukherjea et al., 1999) obtient une précision de 82% et un rappel de 85%.

[40] Classifieur fondé sur le théorème de Bayes permettant de calculer les probabilités conditionnelles.

## 3.5 Discussion

La méthodologie proposée pour l'extraction de relations sémantiques dans le domaine médical repose sur l'identification des entités du domaine puis la validation des relations candidates extraites sur la base de la cooccurrence de ces entités en utilisant des patrons linguistiques. L'utilisation de schémas lexico-syntaxiques pour l'extraction de relations sémantiques a déjà fait l'objet de nombreux travaux. Comme nous l'avons vu, Hearst (Hearst, 1992) est l'une des premières à avoir proposé une approche fondée sur des patrons pour extraire des relations d'hyperonymie. Cependant, sa méthode, qui consiste à extraire un environnement commun à un ensemble de phrases, était essentiellement manuelle. Cette approche a été reprise et complétée par d'autres travaux, toujours dans le domaine de l'extraction de relations sémantiques, dans le but notamment d'automatiser l'apprentissage des patrons. La méthode développée par (Pantel et al., 2004) dont nous nous sommes inspiré se situe précisément dans cette perspective. Cette démarche s'est également avérée particulièrement productive dans des domaines de spécialité comme en attestent par exemple les travaux rapportés dans (Finkelstein-Landau et al., 1999) ou (Séguéla et al., 1999), qui se sont focalisés sur des textes techniques.

La méthode que nous avons exposée ici se différencie de tous ces travaux par le mode d'application des patrons linguistiques induits. Au lieu de les appliquer à la manière d'expressions régulières, nous calculons une distance d'édition entre le patron et la phrase abritant une relation candidate. Cette façon de faire autorise une plus grande souplesse dans l'application des patrons et permet également d'avoir le même mode de fonctionnement lorsque les relations sont caractérisées par des patrons, comme c'est le cas ici, et lorsqu'elles sont caractérisées par des exemples, comme dans une approche de type Memory-Based Learning (Daelemans et al., 2005). On peut même envisager ainsi de mêler les deux approches.

Une autre différence notable avec les travaux tels que (Pantel et al., 2004) est que les résultats de la Section 3.4.2 ont été obtenus sans utilisation d'un filtrage *a posteriori* des relations extraites. En dépit de cette absence, la précision se situe à un haut niveau sans que le rappel ne soit trop faible. Plusieurs explications complémentaires peuvent être avancées. Tout d'abord, cette extraction intervient dans un domaine spécialisé et se focalise sur des relations

intervenant entre des entités spécifiques à ce domaine. Ensuite, les relations sont de type syntagmatique et non paradigmatique comme dans (Pantel et al., 2004). Enfin, les patrons linguistiques appris restent assez spécialisés puisqu'ils ne sont issus que de la généralisation de couples d'exemples.

## 3.6 Conclusion

Dans ce chapitre, nous avons présenté notre méthode développée pour extraire des relations sémantiques intervenant entre des entités du domaine médical. Le but de cette extraction est d'identifier les instances de concepts médicaux dans les documents afin d'enrichir notre ontologie médicale. La méthode que nous avons définie et appliquée s'appuie sur des patrons linguistiques multi-niveaux pour valider des relations candidates extraites des textes. Ces patrons sont appris automatiquement à partir de textes annotés en s'appuyant sur une notion de distance d'édition étendue.

La méthode proposée ici a montré des résultats encourageants en regard des travaux comparables existants. La principale amélioration doit porter sur le rappel. Les évaluations relatives à la validation des relations ont montré que les patrons linguistiques appris ne couvrent pas toutes les manifestations des relations cibles. En outre, ces évaluations étant réalisées seulement à partir des phrases extraites et non de toutes les phrases du corpus d'évaluation du fait de la taille de ce dernier, elles masquent le déficit de rappel résultant de l'absence de reconnaissance des entités médicales déclenchant le processus d'extraction. Même si le niveau de reconnaissance de ces entités peut être considéré comme bon, la nécessité de reconnaître les deux entités d'une relation amplifie l'impact de leur éventuelle mauvaise reconnaissance.

Pour améliorer à la fois la couverture des patrons linguistiques et la reconnaissance des entités médicales, nous envisageons d'adopter une démarche itérative classiquement utilisée dans un tel cas : au lieu de limiter l'usage des patrons linguistiques à la seule validation des relations extraites, il est aussi possible de les utiliser pour extraire de nouvelles entités en ne fixant qu'une seule des entités d'une relation. Ces nouvelles entités viennent à leur tour enrichir la reconnaissance des entités médicales et peuvent ainsi servir à acquérir de nouveaux patrons linguistiques. Une autre voie d'amélioration du rappel est l'utilisation des ressources

sémantiques existant dans le domaine médical, comme le thésaurus MeSH ou le métathésaurus UMLS. Il serait ainsi possible d’inclure la vérification de relations sémantiques telles que l’hyperonymie au niveau de la distance d’édition étendue permettant à la fois de construire les patrons linguistiques et de les appliquer. Enfin, parmi les extensions envisagées de ce travail figure également une extension de la couverture des relations de notre ontologie médicale, dont la Figure 2.1 ne montre qu’une partie. Nous nous sommes limité pour le moment à quatre relations mais les principes testés peuvent tout à fait être appliqués aux autres relations de cette ontologie.

# Quatrième chapitre
# Le système Œdipe

# 4. Le système Œdipe

*Ce chapitre est essentiellement consacré à la présentation de notre système de question-réponse Œdipe. Pour ce faire, nous présentons en premier lieu l'architecture du système puis nous développons tour à tour les différents modules intervenant dans sa chaîne de traitement, c'est-à-dire de l'analyse de la question jusqu'à l'extraction de la réponse candidate. Nous décrivons également dans ce chapitre l'analyseur linguistique LIMA sur lequel repose Œdipe pour l'analyse des questions et des documents sélectionnés par le moteur de recherche. Enfin, la dernière partie de ce chapitre concerne le développement du système Œdipe dans le traitement des questions définitoires.*

## 4.1 Présentation du système Œdipe

Le système Œdipe a été développé à l'occasion de la campagne d'évaluation des systèmes de question-réponse en français EQueR (Ayache, 2005). Il est le premier système de question-réponse développé au LIC2M (CEA/LIC2M)[41] et s'appuie essentiellement sur ses outils d'analyse linguistique. Doté d'une architecture classique pour un système de question-réponse, le système Œdipe a été conçu initialement pour répondre à des questions en retournant des passages de taille fixe à partir d'un ensemble de documents sélectionnés par le moteur de recherche du LIC2M. Cette première version a été étendue dans le cadre des campagnes d'évaluation CLEF-QA dont les réponses attendues sont des réponses courtes (Besançon et al., 2005b). Cependant, aucune modification n'a été effectuée sur la conception globale du système Œdipe.

Dans ce chapitre, nous présentons le fonctionnement du système de question-réponse Œdipe. Après avoir présenté l'architecture du système Œdipe ainsi que l'analyseur linguistique sur lequel repose ce système pour l'analyse des questions et les documents candidats, nous détaillerons plus spécifiquement les différents modules intervenant dans sa chaîne de traitement, c'est-à-dire de l'analyse de la question jusqu'à l'extraction des réponses

[41] Commissariat à l'Energie Atomique / Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue.

candidates. Enfin, nous présentons l'amélioration apportée au système Œdipe concernant l'analyse de questions et le traitement des questions définitoires.

## 4.2 Architecture d'Œdipe

L'architecture du système Œdipe, comme illustrée par la Figure 4.1, est tout à fait classique pour un système de question-réponse. Elle s'appuie à la base sur l'analyseur linguistique LIMA (LIC2M Multilingual Analyzer) (Besançon et al., 2004) qui permet d'une part, de normaliser les mots apparaissant dans les documents et dans les questions, et d'autre part, d'en extraire des entités nommées de type MUC (personnes, lieux, organisations, dates et unités de mesure ainsi que les produits). Il s'agit du même analyseur que celui utilisé pour l'identification des concepts médicaux, les types d'entités reconnus étant différents. La normalisation des mots est réalisée par une analyse morphologique et un étiquetage morpho-syntaxique. La reconnaissance des entités nommées est effectuée quant à elle par une série d'automates appliqués au résultat de l'analyse linguistique. Chaque question posée est analysée afin de déterminer si la réponse attendue est une entité nommée et le cas échéant, le type d'entité nommée concernée. Cette analyse repose sur un ensemble d'environ 150 automates du même type que ceux définis pour la reconnaissance des entités nommées (cf. Section 3.2.2). Les mots pleins de la question sont par ailleurs pondérés afin de caractériser leur importance *a priori*.

Une fois la question analysée, cette dernière est soumise à un moteur de recherche dans le but de récupérer des documents candidats susceptibles de contenir la réponse désirée. Les documents choisis font quant à eux l'objet d'un traitement en deux temps. Un premier traitement permet de localiser les extraits en relation directe avec la question. Ces extraits sont déterminés en fonction du nombre de mots qui composent la question. Chaque extrait se voit attribuer un poids en fonction des mots de la question qu'il contient et éventuellement, de la présence d'entités nommées correspondant au type de la réponse attendue. Les extraits sont ensuite ordonnés suivant leur score et les N premiers sont sélectionnés. Dans le cadre de la campagne d'évaluation EQueR, nous avons fixé le seuil N à 20 documents. Le second module est chargé de localiser la réponse à la question dans les extraits retenus. Si la réponse attendue est une entité nommée, Œdipe retient comme réponse possible la partie de l'extrait considéré centrée sur une entité nommée du type attendu et qui présente le score le plus élevé, score

comparable à celui calculé pour les extraits. Si la réponse attendue n'est pas une entité nommée, Œdipe applique une fenêtre glissante (égale à la taille souhaitée de la réponse) sur l'extrait en calculant pour chacune de ses positions un score comparable à celui de l'extrait. Il retient ensuite comme réponse possible la partie de l'extrait dans laquelle ce score est maximal. Un ensemble de réponses possibles dotées chacune d'un score est ainsi constitué. La liste finale des réponses est obtenue en ordonnant cet ensemble et en le tronquant en fonction du nombre de réponses désiré. Si le score de la meilleure réponse est trop faible, Œdipe suppose qu'aucune réponse n'existe dans les documents. Cette même heuristique est exploitée pour le traitement des questions polaires (oui/non) : une réponse négative est donnée lorsque le score de la meilleure réponse est trop faible. Œdipe prend également en compte les questions de type liste, sans autre particularité que de rechercher le nombre possible de réponses attendues lors du traitement de la question en prenant comme référence la première entité nommée numérique trouvée.

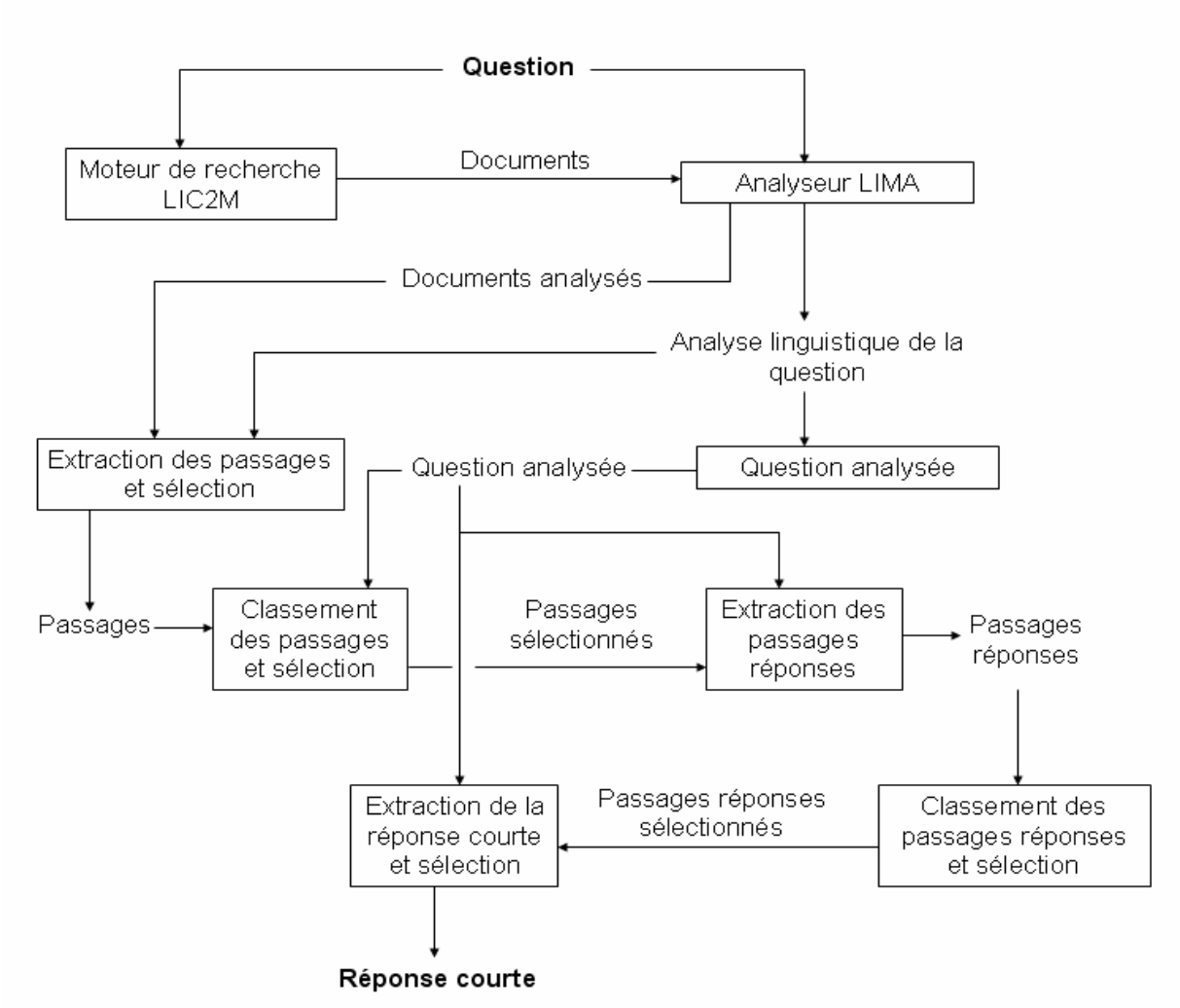


Figure 4.1 Architecture du système Œdipe

## 4.3 Présentation de l'analyseur LIMA

L'analyseur linguistique LIMA (pour LIc2m Multilingual Analyzer) (Besançon et al., 2004) a été développé par le LIC2M, dans la lignée des travaux de Christian Fluhr et ses collègues (Fluhr et al., 1997), en mettant l'accent à la fois sur le multilinguisme et sur une flexibilité permettant de mettre en œuvre différents niveaux d'analyse allant jusqu'à la réalisation d'une analyse linguistique profonde des textes pour en extraire des informations précises. LIMA prend ainsi en charge 7 langues différentes, à savoir le français, l'anglais, l'arabe, l'allemand, l'espagnol, l'italien et enfin le chinois mandarin. Cette analyse linguistique va de la normalisation des termes contenus dans un document jusqu'à l'analyse syntaxique et même plus récemment, la résolution de coréférence. Dans le cadre de notre travail, l'utilisation de LIMA permet d'effectuer une analyse des questions et des documents retournés par le moteur de recherche afin d'obtenir les éléments nécessaires à l'extraction des réponses candidates.

La Figure 4.2 résume l'organisation des modules qui composent la chaîne de traitements de l'analyseur LIMA. Cette architecture est constituée de plusieurs modules de traitement linguistique successifs : la tokenisation, l'analyse morphologique, l'identification des expressions idiomatiques, l'étiquetage morpho-syntaxique des mots, la détection des entités nommées, l'identification des groupes nominaux et verbaux et la mise en évidence des relations de dépendances syntaxiques.

### 4.3.1 Tokenisation et analyse morphologique

La première étape du processus d'analyse linguistique de LIMA est la tokenisation. Cette tâche consiste à découper le texte en « tokens » dans le but d'identifier des candidats mots et des coupures de phrases. Les tokens identifiés sont ensuite recherchés dans un dictionnaire de formes fléchies. Chaque token se voit ainsi associer un ensemble, éventuellement vide, d'entrées du dictionnaire, chaque entrée étant déterminée par une forme normalisée (lemme) et une catégorie morpho-syntaxique. Il est à noter que l'analyseur linguistique LIMA dispose d'un jeu d'étiquettes morpho-syntaxiques assez large qui le distingue des autres systèmes car

ses étiquettes comprennent une dimension positionnelle[42] (Besançon et al., 2004), autrement dit, les catégories concernées permettent de spécifier le positionnement d'un mot par rapport à d'autres mots qui l'entourent dans la phrase. Les propriétés positionnelles de ce jeu d'étiquettes (137 pour le français, 120 pour l'anglais, 100 pour l'espagnol et 38 pour l'allemand actuellement) ont pour principal objectif de faciliter la désambiguïsation morpho-syntaxique.

Dans le cas où le token n'est pas trouvé dans le dictionnaire, une catégorie par défaut lui est assignée à partir de ses propriétés typographiques. Par exemple, un mot inconnu commençant par une lettre majuscule sera considéré comme étant un nom propre. Dans le cas de langues telles que le français ou l'anglais, l'analyse morphologique se limite à un accès au dictionnaire de formes fléchies et au traitement des mots inconnus mais pour des langues dont la morphologie est beaucoup plus complexe, d'autres modules sont également sollicités pour découper des mots composés (en allemand par exemple), identifier des préfixes et des suffixes (cas de l'arabe) ou encore restaurer la voyellation des mots (cas de l'arabe encore une fois).

### 4.3.2 Identification des expressions idiomatiques

À la suite de l'analyse morphologique, les expressions idiomatiques sont repérées et remplacées par une seule unité. Les expressions idiomatiques sont des expressions ou des mots composés usuels dont le sens est non décomposable, à l'instar de « au fur et à mesure » ou de « prendre part » par exemple. La détection de ces expressions est réalisée en appliquant un ensemble de règles issues d'un dictionnaire spécifique, règles déclenchées par un mot particulier et validées par la satisfaction d'un certain nombre de contraintes par rapport au contexte gauche et au contexte droit du mot concerné. Ce principe est identique à celui adopté pour la reconnaissance des entités nommées. L'utilisation de ce type de règles permet de traiter des phénomènes d'insertion comme « prendre activement part ».

[42] Le jeu d'étiquettes permet de différencier entre les adjectifs épithètes antéposés et les adjectifs épithètes postposés.

### 4.3.3 Étiquetage morpho-syntaxique

Le module suivant du processus de traitement de l'analyseur LIMA concerne la désambiguïsation morpho-syntaxique. Cette phase s'appuie sur un étiqueteur morpho-syntaxique permettant de désambiguïser les catégories morpho-syntaxiques identifiées lors de l'analyse morphologique afin de réduire le nombre de mots possibles pour chaque token. Pour ce faire, le système repose sur un modèle statistique à base de matrices de bigrammes et trigrammes de catégories extraites au préalable à partir de corpus étiquetés manuellement.

Cette étape permet également de distinguer les mots pleins des textes, comme les noms, les verbes, les adjectifs et les noms propres, des mots grammaticaux, ce qui est en particulier utile pour toutes les tâches de recherche d'information.

### 4.3.4 Identification des entités nommées

Nous avons déjà évoqué la façon dont les entités nommées sont identifiées dans le cadre de LIMA au chapitre précédent lors de la présentation de l'identification des concepts médicaux (cf. Section 3.2.2) : cette identification s'appuie sur des règles définies manuellement et compilées sous la forme d'automates. Ces règles suivent d'ailleurs le même formalisme que celles définies pour la reconnaissance des expressions idiomatiques. Dans la version générale de LIMA, seuls les types d'entités considérés sont différents puisqu'il s'agit dans ce cas de noms de personnes, d'organisations, de lieux, de dates et de nombres, de produits et enfin d'évènements. Les tokens identifiés comme une entité nommées sont ensuite regroupés en une seule unité pour la suite de l'analyse. Pour les types d'entités généraux mentionnés, ce module obtient en moyenne une précision de 80% et un rappel de l'ordre de 60% pour le français (5000 textes), l'anglais (5000 textes) et l'espagnol (50 textes) (Besançon et al., 2004).

### 4.3.5 Analyse syntaxique

La dernière étape de l'analyse linguistique est l'analyse syntaxique. Un premier module de cette analyse, assimilable à un chunker, permet de délimiter les groupes nominaux et les

groupes verbaux en utilisant des données sur les successions possibles de catégories morpho-syntaxiques à l'intérieur de chacun de ces groupes. À sa suite, un second module détermine à la fois les relations de dépendances internes à chaque groupe et celles intervenant entre les groupes, comme par exemple les relations existant entre un verbe et son sujet ou son objet. L'établissement de ces relations de dépendances s'appuie sur des règles écrites manuellement utilisant le même moteur d'automates que l'identification des entités nommées ou des expressions idiomatiques.

Les relations de dépendances internes aux groupes nominaux sont par ailleurs exploitées pour construire la liste des mots composés possibles de chaque phrase. Ces mots composés, outre leur intérêt du point de vue terminologique, ont un rôle particulièrement important dans le fonctionnement du moteur de recherche du LIC2M qui est utilisé par Œdipe.

### 4.3.6 Exemple du résultat de l'analyse linguistique

Les résultats du traitement linguistique présenté précédemment contiennent des informations linguistiques de niveaux différents (morpho-syntaxiques et syntaxiques). Pour le moteur de recherche du LIC2M, le résultat de l'analyse réalisée par LIMA sert à la fois à indexer la base documentaire et à analyser les requêtes soumises. Ci-dessous un exemple du résultat d'une partie de l'analyse linguistique (jusqu'à l'étiqueteur morpho-syntaxique) réalisée par LIMA de la phrase suivante : « une glande endocrine est un organe qui sécrète des hormones. ». Œdipe se limite d'ailleurs à ce niveau de l'analyse et n'exploite pas l'analyse syntaxique.

```
1 | Une | un#L_DET_ARTICLE_INDEF
5 | glande | glande#L_NC_GEN
12 | endocrine | endocrine#L_ADJ_QUALIFICATIF_EPITHETE_POSTN
22 | est | être#L_VERBE_PRINCIPAL_INDICATIF
26 | un | un#L_DET_ARTICLE_INDEF
29 | organe | organe#L_NC_GEN
36 | interne | interne#L_ADJ_QUALIFICATIF_EPITHETE_POSTN
44 | qui | qui#L_PRON_REL_COI
48 | sécrète | sécréter#L_VERBE_PRINCIPAL_INDICATIF
56 | des | un#L_DET_ARTICLE_INDEF
60 | hormones | hormone#L_NC_GEN
68 | . | .#L_PONCTU_FORTE
```

La première colonne du fichier de sortie représente la position du token dans le document (en caractères) sans prendre en compte les balises XML qui figurent dans le document. La deuxième et la troisième colonne indiquent respectivement la forme fléchie du terme, c'est-à-dire le token tel qu'il figure dans le document, et sa forme normalisée (lemme). Enfin, la dernière colonne donne la catégorie grammaticale des mots : « L_VERBE_PRINCIPAL_INDICATIF » pour un verbe, « L_NC_GEN » pour un nom commun, « L_DET_ARTICLE_INDEF » pour un déterminant article indéfini, « L_ADJ_QUALIFICATIF_EPITHETE_POSTN » pour un adjectif qualificatif épithète et « L_PONCTU » pour les ponctuations.

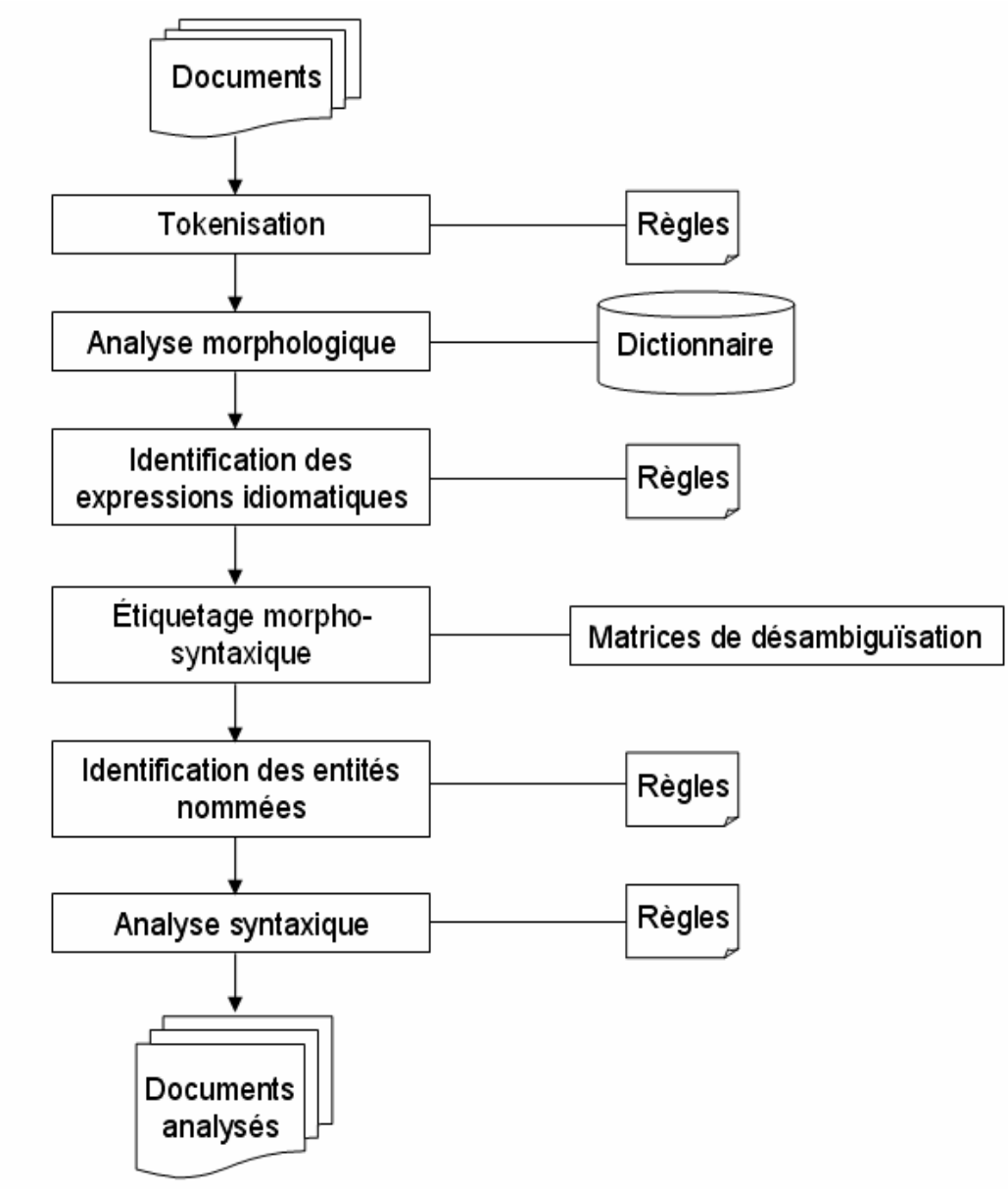


Figure 4.2 Chaîne de traitements de l'analyseur LIMA

## 4.4 Description des modules du système Œdipe

Dans cette section, nous détaillons plus particulièrement les différents modules du système de question-réponse Œdipe. Pour ce faire, nous avons décomposé la chaîne de traitement du système en deux phases : une première phase concerne l'extraction des passages candidats susceptibles de contenir la réponse recherchée à partir d'une sélection de documents renvoyés par le moteur de recherche ; la deuxième phase se focalise sur l'extraction des réponses candidates à partir des passages sélectionnés.

### 4.4.1 Sélection des passages candidats

La première phase du processus de traitement du système Œdipe consiste à extraire les passages réponses susceptibles de contenir une réponse candidate à une question. Pour ce faire, le système Œdipe repose sur un enchaînement de traitements s'appuyant sur l'analyseur LIMA pour le traitement linguistique, l'analyse de la question et enfin l'interrogation du moteur de recherche du LIC2M pour récupérer une sélection de documents en rapport avec la question posée. Nous présentons ci-dessous les étapes essentielles de la chaîne de traitement.

#### 4.4.1.1 Moteur de recherche

Pour sélectionner des documents à partir d'une collection initiale, le système de question-réponse Œdipe utilise le moteur de recherche développé par le LIC2M. Ce dernier a été également exploité lors de différentes campagnes d'évaluation comme CLEF 2003 (Besançon et al., 2004) et CLEF 2004 (Besançon et al., 2005a). Cette sélection de documents est une étape cruciale pour un système de question-réponse car si le moteur de recherche échoue dans la recherche de documents pertinents par rapport à la requête, le système de question-réponse n'aura donc aucune chance de trouver la bonne réponse. Le moteur de recherche est guidé par l'identification dans les documents des concepts significatifs de la requête, c'est-à-dire que les documents retournés par le moteur ne sont pas classés selon un score de précision mais plutôt par rapport au nombre de concepts (les termes simples et surtout complexes ainsi que les entités nommées) de la requête d'interrogation présents dans ces documents. Il privilégie, dans la sélection des documents, en premier lieu, ceux qui contiennent une occurrence de

chaque concept, tel que formulé dans la requête, puis ensuite ceux contenant le plus grand nombre de concepts, sous leur forme originelle ou sous la forme d'une variante[43] et sans tenir compte de leur nombre d'occurrences dans les textes. Par conséquent, les résultats du moteur de recherche du LIC2M se présentent comme une liste de classes où chaque classe se compose d'un ensemble de documents correspondant au même ensemble de concepts. Puisque tous les documents appartenant à une même classe sont considérés similaires, il serait logique de sélectionner tous les documents de la classe. Pour des raisons d'efficacité, nous fixons cependant un nombre minimal et un nombre maximal de documents à retenir pour chaque question traitée. À titre d'exemple, pour la campagne d'évaluation CLEF-QA 2005, le nombre minimum de documents à retenir était limité à 25 documents, et même à 20 pour la campagne CLEF-QA 2006, le nombre maximum à 50 documents pour chaque question. Ces contraintes sont mises en œuvre en appliquant l'algorithme suivant :

```
Documents sélectionnés ← {}
i ← 1
Tant que card (Documents sélectionnés) < 20  ∧  i ≤ card (classes) faire
   Classe actuelle ← classes[i]
   i ← i + 1
   Si card (Documents sélectionnés) + card (Classe actuelle) ≤ 50
      alors
        Documents sélectionnés ← Documents sélectionnés ∪ Classe actuelle
      sinon
        randNbDocsSel = 50 – card (Documents sélectionnés)
        Documents sélectionnés ← Documents sélectionnés ∪ random (Classe actuelle,
                                  randNbDocsSel)
   fin
```

Où random(S, N) est la fonction qui permet de sélectionner aléatoirement N éléments à partir d'un ensemble S. Le principe est de retenir le nombre de classes permettant de sélectionner au moins le nombre minimal requis de documents tout en ne dépassant pas le nombre maximal fixé et, dans le cas où la dernière classe est suffisamment large pour couvrir l'intervalle entre la borne inférieure et la borne supérieure, de choisir aléatoirement les documents dans cette classe pour atteindre le nombre maximum de documents fixé. Cette stratégie de choix aléatoire répond au principe d'équivalence des documents d'une classe. Pour la campagne d'évaluation CLEF-QA 2006, un nombre moyen de 33 documents par question ont été sélectionnés par cet algorithme.

[43] Dans le cas où le terme apparaît dans le document sous une forme différente de celle employée dans la requête comme un synonyme, acronyme ou un sous-terme.

#### 4.4.1.2 L'analyse linguistique

Le traitement linguistique des questions et des documents retournés par le moteur de recherche du LIC2M est assuré par l'analyseur linguistique LIMA que nous avons décrit à la Section 4.3. Cependant, dans le cadre du système de question-réponse Œdipe, celui-ci n'est exploité que pour une partie de ses compétences. Plus précisément, le système Œdipe s'appuie sur la normalisation morpho-syntaxique des mots, l'identification des mots pleins et enfin la reconnaissance et le typage des entités nommées. La normalisation des mots et l'identification des mots pleins sont réalisées par la combinaison de l'analyse morphologique et de l'étiquetage morpho-syntaxique. L'analyse syntaxique n'est exploitée que partiellement et indirectement au niveau du moteur de recherche. Celui-ci reposant sur une analyse des documents mettant en avant les termes complexes, il fait appel à l'extracteur de termes de LIMA qui exploite lui-même la partie de l'analyse syntaxique mettant en évidence les relations de dépendance à l'intérieur des groupes nominaux.

L'analyse linguistique des documents et des questions repose donc sur les modules suivants :

- tokenisation,
- analyse morphologique,
- identification des expressions idiomatiques,
- étiquetage morpho-syntaxique,
- identification des mots pleins,
- identification des entités nommées.

#### 4.4.1.3 Analyse de la question

Une des parties importantes d'un système de question-réponse est l'analyse des questions. Elle a pour principal but de caractériser le type de la réponse attendue tout en construisant la requête d'interrogation à soumettre au moteur de recherche. Plus spécifiquement, chaque question posée est analysée afin de déterminer si le type de la réponse attendue est une entité nommée et le cas échéant, le type d'entité nommée recherché. Dans le cas du système Œdipe, le module « analyse de la question » permet de réaliser deux tâches différentes :

- identification du type de la réponse attendue ;
- identification des mots pleins de la question qui sont par ailleurs pondérés afin de caractériser leur importance *a priori.*

La première tâche permet de définir la stratégie à adopter par le système Œdipe pour extraire les réponses : si le type de la réponse attendue correspond aux types d'entités nommées identifiés par l'analyseur LIMA, Œdipe recherche dans les passages candidats, extraits à partir des documents sélectionnés par le moteur de recherche du LIC2M, l'entité nommée du type concerné dont le contexte est le plus compatible avec la question. Autrement, il considère que la question est une question de type « définition » et applique une fenêtre glissante (égale à la taille souhaitée de la réponse) sur les passages candidats en calculant pour chacune de leurs positions un score pour ne retenir que l'extrait ayant le plus grand score. La deuxième tâche consiste principalement à identifier les mots pleins de la question et à rechercher leur information normalisée dans un corpus de référence pour évaluer leur degré de spécificité.

Ainsi, pour déterminer le type de la réponse attendu, l'analyse de la question repose principalement sur l'application d'un ensemble de patrons morpho-syntaxiques. La stratégie généralement adoptée pour les construire s'inspire de stratégies d'extraction de patrons couramment employées en extraction d'information (Riloff, 1994) et de travaux dans le domaine de l'apprentissage dit « par alignement » (Van Zaanen, 2001 ; Balvet et al., 2005). Concrètement, cette stratégie de co-analyse suit la procédure schématisée ci-dessous :

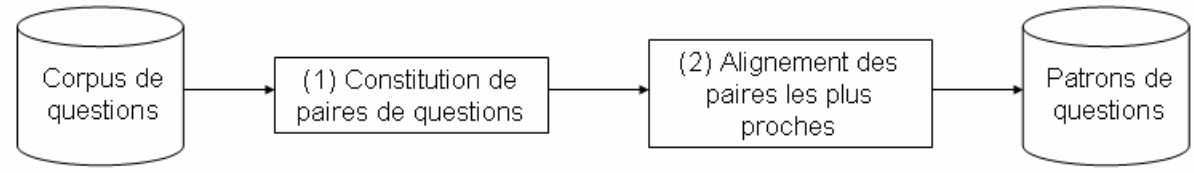


Figure 4.3 Étapes pour la constitution d'une base de données de patrons de questions

De même que pour l'extraction de patrons lexico-syntaxiques présentée à la Section 3.3.2, l'étape (1) se base sur la mesure de la distance d'édition entre deux chaînes de caractères (Levenshtein, 1966), calculée à partir d'opérations d'insertion, d'élision et de déplacement. Elle aboutit à une liste de paires de questions associées à un score de distance d'édition. L'étape (2) cherche, pour toutes les paires, la plus longue sous-chaîne commune de mots en

s'inspirant de l'algorithme Longest Common Substring (Hirschberg, 1977). Le résultat de ces deux étapes est une liste de paires de questions, chaque paire étant caractérisée par un score de distance d'édition, ainsi que des scores dérivés de celui-ci, et par la plus longue sous-chaîne de mots commune aux deux questions de la paire. Par exemple, la recherche de la plus longue sous-chaîne commune de mots pour la paire de questions ci-dessous donne le patron suivant, où les '_' marquent des positions possibles dans la séquence de mots analysée[44] :

*Quelle est la capitale de la Bosnie ?*
*Quelle est la capitale de Madagascar ?* ⟶ *Quelle est la capitale de _ _ ?*

Le patron extrait est ensuite traduit sous la forme d'une expression régulière typée utilisée lors de l'identification du type d'une question. Par ailleurs, les mots non alignés peuvent être considérés comme les membres d'un même paradigme (*i.e.* des noms de pays pour *Bosnie* et *Madagascar*). Il est à noter que le système Œdipe repose sur une collection de 149 types de questions auxquelles sont associées un ensemble de types de réponse attendus. Par exemple, pour le patron « Quelle est la capitale _ _ ? », la réponse attendue est de type « Lieu ».

L'approche adoptée ici, concrétisée par la plate-forme CoPT[45] (Corpus Processing Tools) développée par Antonio Balvet, est donc une approche de surface qui ne met en œuvre aucune connaissance linguistique explicite (*i.e.* morphologique, syntaxique ou sémantique) autre que des récurrences de chaînes de caractères et des coïncidences de position pour ces chaînes, ce qui lui confie un large champ d'application. Elle requiert simplement une certaine stabilité dans les patrons morpho-syntaxiques employés. Ainsi, l'application de l'algorithme d'extraction de patrons multi-niveaux (Pantel et al., 2004) est une extension naturelle à envisager pour améliorer et automatiser davantage l'extraction de patrons de typage de questions.

#### 4.4.1.4 Extraction, classement et sélection des passages

Après la sélection d'un ensemble restreint de documents par le moteur de recherche du LIC2M, le système de question-réponse Œdipe procède à la délimitation des passages

[44] La plus longue sous-chaîne commune est alignée sur la plus longue des séquences traitées.
[45] http://french.osstrans.net/software/copt.html

candidats susceptibles de contenir la réponse correcte à une question posée. Cette délimitation repose particulièrement sur la détection de certaines zones dans les documents contenant une forte densité des mots de la question soumise. Cette tâche est effectuée en attribuant à chaque position du document une valeur d'activation : quand une position contient un mot de la question, une valeur fixe est ajoutée à sa valeur d'activation et à la valeur d'activation des positions qui l'entourent. Enfin, les passages candidats correspondent aux positions contiguës du document dont la valeur d'activation est supérieure à un seuil fixé.

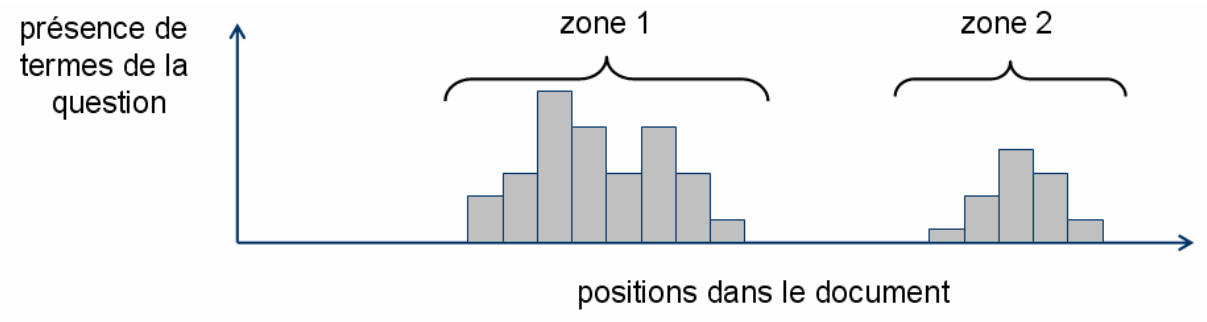


Figure 4.4 Extraction de passages dans le cadre du système Œdipe

Après cette étape, un score est calculé pour chaque passage extrait. Ce score est déterminé par trois facteurs :

- le nombre des mots de la question contenus dans le passage,
- la présence dans le passage d'une entité nommée correspondant au type de la réponse attendue si ce dernier est une entité nommée,
- la densité des mots de la question dans le passage.

Plus précisément, le score d'un passage « Pi » est calculé par la formule suivante :

$$Score(Pi) = \alpha \cdot wordScore(Pi) + \beta \cdot neScore(Pi) + \gamma \cdot densityScore(Pi)$$

Où $\alpha$, $\beta$ et $\gamma$ sont des modulateurs[46] et tous les autres scores sont compris entre 0 et 1.

Une fois leur score calculé, les passages candidats sont classés selon un ordre décroissant de leur score puis ne sont retenus, pour les étapes suivantes, que les N premiers passages[47].

[46] Pour la campagne CLEF-QA, $\alpha$, $\beta$ et $\gamma$ étaient égaux à 1.
[47] Pour l'évaluation CLEF-QA, nous avons fixé N égal à 20.

### 4.4.2 Extraction de la réponse candidate

Le système de question-réponse Œdipe a été développé initialement pour trouver des passages réponses susceptibles de contenir la réponse à la question posée plutôt que pour extraire des réponses courtes. Cette dernière fonctionnalité a été néanmoins ajoutée à Œdipe pour la campagne d'évaluation CLEF-QA qui n'évalue que des réponses courtes. Néanmoins, aucune modification n'a été effectuée sur la conception globale du système Œdipe. Le système commence donc par identifier des passages candidats avant d'extraire dans un second temps des réponses courtes.

#### 4.4.2.1 Extraction des passages réponses

Le système Œdipe détecte un extrait correspondant à la réponse candidate à partir de chaque passage sélectionné. Ce processus consiste à déplacer une fenêtre au-dessus du passage concerné et à calculer un score à chaque position de la fenêtre suivant son contenu. La taille de la fenêtre est égale à la taille de la réponse attendue (lors de la campagne d'évaluation EQueR, la taille était égale à 250 caractères). La réponse extraite est constituée par le contenu de la fenêtre pour la position ayant le score le plus élevé. Le déplacement de cette fenêtre dépend principalement du type de la réponse attendue. Plus précisément, si la réponse désirée n'est pas une entité nommée, la fenêtre se déplace alors sur chaque mot plein du passage. Autrement, elle se déplace en se positionnant directement sur les entités nommées compatibles avec le type de la réponse attendu. Dans les deux cas, le score attribué à chaque position de la fenêtre est la somme des deux scores suivants :

- un score concernant le nombre de mots de la question présents dans la fenêtre,
- un score égal à la proportion des entités nommées de la question présentes dans la fenêtre.

Pour les questions dont la réponse attendue n'est pas une entité nommée, il est fréquent d'obtenir plusieurs positions adjacentes avec des scores élevés identiques. Dans ce cas particulier, le passage réponse sélectionné est extrait du milieu de cette zone et non à partir du début car souvent, la réponse apparaît après les mots de la question. Enfin, comme pour la sélection des passages, tous les passages réponses sont classés selon un ordre décroissant de

leurs scores. Si le score du meilleur passage réponse est trop faible, Œdipe suppose qu'il n'y a aucune réponse possible à la question considérée.

#### 4.4.2.2 Extraction des réponses courtes

Si la réponse attendue est une entité nommée, l'extraction des réponses courtes est directe : le passage réponse avec le score le plus élevé est choisi puis l'entité nommée sur laquelle la fenêtre était centrée dans le passage réponse est retournée comme étant la réponse courte. Dans le cas où la réponse attendue n'est pas une entité nommée, la recherche d'une réponse courte s'appuie sur un ensemble restreint d'heuristiques faisant l'hypothèse que la réponse est un groupe nominal. Le système Œdipe commence donc par localiser tous les groupes nominaux contenus dans le passage réponse en appliquant le patron morpho-syntaxique suivant :

(DET|NP|NC) (NC|NP|ADJ|PREP) (ADJ|NC|NP)[48]

Ensuite, Œdipe calcule un score pour chaque groupe nominal repéré. Ce score tient compte à la fois de la taille de la réponse et de ses contextes :

- il est proportionnel à la taille de la réponse, avec néanmoins une limite fixe ;
- il est augmenté par une valeur fixe à chaque fois qu'un élément spécifique est trouvé dans son contexte restreint (2 mots). Cet élément peut être une des entités nommées présentes dans la question initiale ou d'une manière générale, un élément qui caractérise la présence d'une définition comme une virgule, une parenthèse ou encore le verbe « être ».

Le score final d'une réponse est donné par la somme du score relatif au passage réponse et du score de la réponse courte, le score calculé pour les réponses courtes étant destiné à faire un choix au sein d'un passage réponse et non à confronter directement les réponses courtes provenant de différents passages. La réponse ayant le plus grand score est alors retournée comme réponse pour la question considérée.

[48] DET : déterminant, NP : nom propre, NC : nom commun, ADJ : adjectif, PREP : préposition

## 4.5 Traitement des questions définitoires

Dans cette section, nous exposons une amélioration apportée au système de question-réponse Œdipe. Cette amélioration concernait dans un premier temps le typage des questions en révisant essentiellement les règles de typage. Ces règles ont également été complétées pour identifier l'objet sur lequel porte la question (focus). Dans un second temps, il s'agissait de fournir au système Œdipe la capacité de traiter les questions définitoires, ce qui constitue la principale amélioration du système Œdipe lors de la campagne d'évaluation des systèmes de question-réponse CLEF-QA 2006 (Besançon et al., 2006). L'objectif est donc le traitement des questions portant sur des définitions telles que : « Qu'est ce que X ? » ou « Qui est X ? ». Les questions définitoires s'avèrent très difficiles à traiter car, d'une part, elles sont généralement courtes, et d'autre part, la recherche d'une réponse ne peut pas se focaliser sur un type spécifique d'éléments comme dans le cas des entités nommées. De ce fait, ne pas adopter de stratégie adaptée dans ce contexte produit en général de mauvais résultats (Besançon et al., 2005b).

Pour extraire des réponses à ce type de questions à partir de phrases réponses sélectionnées, la plupart des systèmes de question-réponse utilisent un ensemble de patrons linguistiques construits manuellement, à l'instar de (Soubbotin et al., 2001). Un certain nombre de travaux ont été également réalisés pour apprendre automatiquement de tels patrons à partir d'exemples en s'appuyant sur des travaux du domaine de l'extraction d'information. Une des premières tentatives dans ce sens est le travail effectué par Ravichandran et Hovy (Ravichandran et al., 2002), qui propose un processus d'apprentissage de partons d'extraction pour un type de question donné (date de naissance, nom de personne, etc.). Les patrons appris sont spécifiques au type de la question représentant ainsi des formes de réponse possibles. Par exemple, pour une question portant sur une date de naissance, on pourrait avoir des réponses de la forme « Mozart was born in 1756 » identifiées par le patron « <*NAME*> was born in <*BIRTHDAY*> » où <*NAME*> et <*BIRTHDAY*> sont des balises remplaçant respectivement le focus de la question et la réponse candidate à la question. Pour construire ces patrons d'extraction, (Ravichandran et al., 2002) exploite à la fois le Web et une structure de donnée de type arbre de suffixes. L'utilisation du Web pour l'apprentissage des patrons linguistiques est aussi la solution adoptée par (Du et al., 2004). (Jousse et al., 2005) ont également testé plusieurs approches concernant l'apprentissage automatique des patrons d'extraction, tandis

que (Cui et al., 2005) ont proposé un nouvel algorithme pour induire des schémas lexico-syntaxiques probabilistes.

Des travaux tels que (Ravichandran et al., 2002 ; Jousse et al., 2005) ont prouvé que la construction d'un ensemble d'exemples de question-réponse se révèle une tâche tout à fait aisée, spécialement à partir du Web. C'est la raison pour laquelle nous avons adopté pour le système Œdipe une approche fondée sur des patrons lexico-syntaxiques construits à partir d'exemples de réponses à des questions définitoires. Elle consiste dans un premier temps à constituer des patrons linguistiques exprimant une relation de type « définition » puis à les appliquer dans un deuxième temps pour extraire des réponses à des questions définitoires. Par ailleurs, cette approche est plus souple et moins coûteuse lorsqu'un système de question-réponse doit être étendu à de nouveaux domaines.

### 4.5.1 Identification du focus

L'identification du focus des questions constitue une part de notre contribution dans le développement du système Œdipe. Lors de la campagne d'évaluation CLEF-QA 2006, nous avons appliqué cette identification uniquement aux questions définitoires mais d'un point de vue plus général, la détermination du focus de la question s'avère également utile pour améliorer le traitement des questions factuelles. Le focus de la question exprime formellement la partie de la question supposée être présente à proximité de la réponse dans les passages sélectionnés (Ferret et al., 2002a). Comme pour l'identification des entités nommées (cf. Section 3.2.2), cette tâche se fonde sur l'application d'un ensemble de règles prédéfinies. Prenons par exemple les trois règles suivantes :

[Qu'] **::** [est] [ce] [@Que] [$L_DET] *{1-20} [\ ?] **:** DEFINITION_FOCUS **:**
Qu'est ce que l'Atlantis ?

[Qui] **::** [être$L_V] *{1-20} [\ ?] **:** DEFINITION_FOCUS **:**
Qui est Hugo Chavez ?

[Comment] **::** [$L_DET] *{1-5} [peut-il] [être$L_V] [défini] [\ ?] **:** DEFINITION_FOCUS **:**
Comment l'IMC peut-il être défini ?

La première règle identifie le terme *Atlantis* comme focus de la question définitoire « Qu'est ce que l'Atlantis ? » alors que la deuxième permet d'extraire *Hugo Chavez* comme le focus de la question « Qui est Hugo Chavez ? »[49]. La troisième règle quant à elle extrait le terme *IMC* comme focus de la question « Comment l'IMC peut-il être défini ? ». Il est à noter que le focus est spécifié dans la règle par l'élément non délimité par des crochets, *i.e.* l'élément qui n'appartient pas à l'entité reconnue. Par exemple, dans les deux premières règles citées ci-dessus, le focus est déterminé par l'élément « *{1-20} » pour préciser que le focus est l'expression restante de la question et peut être composé de 1 à 20 mots. Le chiffre 20 est donné parce qu'il est peu probable d'avoir des questions de plus de 20 mots. Dans la troisième règle, le focus « *{1-5} » est l'élément qui précède l'expression (peut-il).

L'ensemble des règles élaborées pour déterminer le focus et les questions de définition (29 règles) a été intégré à la collection de règles utilisée pour le typage des questions. Au total, l'analyse des questions est réalisée par un ensemble de 238 règles implémentées sous la forme d'automates à états finis. Ces automates, intégrés dans l'analyseur LIMA, sont de même nature que ceux définis pour la reconnaissance des entités nommées et l'identification des expressions idiomatiques. Chaque règle est une sorte de patron lexico-syntaxique pouvant aussi inclure des classes sémantiques. Une fois déclenchée, la règle associe la question à un type de questions parmi les 158 types de questions distingués (contre 149 initialement). En effet, les règles de typage dans Œdipe ont fait l'objet d'une révision dans le but d'améliorer le typage existant. Cette révision a porté essentiellement sur la correction des règles existantes et l'intégration de nouvelles règles telles que les règles concernant l'analyse des questions définitoires. Par exemple, on peut citer les règles suivantes :

typage initial : [Qui] **::** (être$L_V) *{1-20} [\ ?] **:** F_QUI_PERSONNE **:**
nouveau typage : [Qui] **::** (être$L_V) (T_AS)[50] *{1-20} [\ ?] **:** F_QUI_DEFINITION **:**
*/Qui est Michel Platini ?/*

typage initial : [Où] **::** *{0-1} (être$L_V) *{1-20} [\ ?] **:** F_OU_LIEU **:**
nouveau typage : [Où] **::** (en) (être$L_V) *{1-20} [\ ?] **:** F_OU_NONE[51] **:**
/*Où en est-on avec le développement durable ?*/

[49] Ces deux questions sont issues de la campagne d'évaluation CLEF-QA 2006.
[50] Première lettre du mot est en majuscule.
[51] NONE permet d'indiquer que le type de la réponse attendu n'est pas une entité nommée.

Enfin, pour identifier le type de la réponse recherchée, nous procédons à un mapping entre le type de la question et le type de la réponse attendu. Plus explicitement, nous associons chaque type de question à un type de la réponse attendu, comme c'est le cas des types de questions suivants : F_QUI_PERSONNE (Personne), F_QUAND_DATE (Date), D_DEFINITION (Définition), etc.

### 4.5.2 Apprentissage des patrons de définition

L'algorithme utilisé pour apprendre les patrons linguistiques permettant de répondre aux questions définitoires est une extension de l'algorithme présenté par Ravichandran et Hovy (Ravichandran et al., 2002). Il s'appuie sur l'algorithme d'extraction de patrons multi-niveaux que nous avons déjà utilisé pour l'extraction de patrons de relations (Pantel et al., 2004). Au lieu d'apprendre des patrons ne faisant intervenir que la forme de surface des phrases comme dans (Ravichandran et al., 2002), nous pouvons ainsi apprendre des patrons intégrant différents niveaux d'information linguistique. Plus généralement, il faut noter que cet algorithme peut être employé pour induire des patrons à partir de textes afin d'extraire différentes sortes d'éléments : des relations sémantiques pour le peuplement des bases de connaissances (Embarek et al., 2006 ; Pantel et al., 2004), des réponses dans le cadre des systèmes de question-réponse ou encore des relations intervenant entre des entités dans le domaine de l'extraction d'information.

Dans notre cas, l'induction des patrons linguistiques est effectuée à partir d'un ensemble d'exemples concernant des réponses à des questions définitoires. L'élément de base d'un patron peut être la forme fléchie d'un mot, sa catégorie grammaticale ou sa forme normalisée. Ces trois niveaux d'information sont obtenus en appliquant l'analyseur linguistique LIMA sur les phrases contenant les exemples de réponses. Plus précisément, l'apprentissage des patrons spécifiques pour extraire des réponses à des questions de type définition est réalisé par le processus suivant :

1- Construire un corpus d'exemples constitué de phrases contenant les réponses à des questions définitoires. Contrairement à (Ravichandran et al., 2002) ou à (Du et al., 2004), notre corpus d'exemples n'est pas collecté à partir du Web mais récupéré à la

fois des résultats de l'évaluation EQueR et des évaluations de la campagne CLEF-QA. Pour chaque question « définition » de ces évaluations, toutes les phrases contenant une réponse correcte sont extraites (339 phrases pour 74 questions) ;

2- Appliquer l'analyseur linguistique LIMA à toutes les phrases réponses afin d'obtenir les différents niveaux d'informations linguistiques des mots ;
3- Remplacer dans chaque phrase réponse l'objet de la question (focus) par la balise <*focus*> et la réponse courte de la question par la balise <*answer*>. Cette notion du focus est ici plus générique que l'élément <name> utilisé dans (Ravichandran et al., 2002), qui le limite aux entités nommées ;
4- Appliquer l'algorithme d'apprentissage de patrons multi-niveaux (voir Figure 3.2) entre chaque couple de phrases ;
5- Classer les patrons selon leur fréquence d'apparition puis sélectionner les P premiers patrons.

De même que pour l'apprentissage des patrons d'extraction de relations sémantiques, nous avons éliminé tous les patrons comprenant plus de deux opérateurs d'alignement (*s*) et (*g*). Voici quelques exemples des patrons de définition générés par le processus proposé ci-dessus :

<answer> ( <focus>
<focus> être L_DET_ARTICLE_INDEF <answer>
<focus> L_VERBE_PRINCIPAL_INDICATIF L_DET_ARTICLE_INDEF <answer>
<focus> , (*g*) <answer>
<answer> ( par ex. L_DET_ARTICLE_INDEF (*s*) <focus>
<answer> (*s*) comme L_DET_ARTICLE_INDEF <focus>

### 4.5.3 Application des patrons de définition

Les patrons de définition appris par l'algorithme présenté à la section précédente ont pour but d'extraire des réponses courtes à des questions portant sur des définitions. Pour cette étude, nous avons retenu la totalité des patrons construits (42) en raison du faible nombre de ces derniers. Cette liste de patrons a été intégrée à la chaîne de traitement du système Œdipe (voir

Figure 4.5) et est appliquée à la suite de la phase « sélection des passages réponses ». Cette application consiste à instancier les patrons de définition avec le focus de la question posée, identifié lors de son analyse, puis à aligner les patrons instanciés avec le passage réponse. Elle permet d'une part, de déterminer si le patron s'apparie avec la phrase réponse et d'autre part, d'extraire la réponse à la question dans le cas où un appariement est détecté. Plus précisément, la procédure appliquée est la suivante :

1- Instanciation des patrons de définition. La tâche consiste à remplacer toutes les balises <focus> dans les patrons par le focus de la question. Le focus concerné est identifié lors de la phase d'analyse de la question par des règles spécifiques présentées à la Section 4.5.1 ;
2- Application de l'analyseur linguistique LIMA à chaque passage réponse sélectionné par le système Œdipe pour acquérir les différents niveaux d'information linguistique présents au niveau des patrons ;
3- Extraction des réponses courtes. Cette extraction est réalisée en alignant le patron candidat avec le passage réponse. L'alignement commence à partir du focus dans le passage puis une comparaison est effectuée mot par mot entre le passage et le patron considéré jusqu'à la balise <*answer*> du patron. Si le processus de vérification aboutit, le groupe nominal du passage réponse qui correspond à la balise <*answer*> dans le patron est extrait et considéré comme la réponse courte à la question posée ;
4- Les réponses courtes extraites par l'utilisation des patrons de définition sont classées selon leur score, c'est-à-dire le nombre de patrons employés pour les extraire. La réponse courte avec le score le plus élevé est retournée par le système Œdipe comme la réponse à la question.

Ce processus a été appliqué dans le cadre de la campagne d'évaluation CLEF-QA 2006 (cf. Section 6.2.2.2). L'application des patrons de définition a permis de donner au système Œdipe la possibilité de répondre aux questions définitoires. Par exemple, pour la question « Qu'est-ce que la RKA ? », Œdipe a retourné la réponse « Agence Spatiale Russe » en appliquant le patron « <answer> (<focus> » sur le passage réponse « …relève un porte-parole de l'Agence Spatiale Russe (RKA). ». Par ailleurs, il est intéressent de noter qu'une même réponse peut être extraite par plusieurs patrons. Ce phénomène se produit lorsque les patrons ayant servi à identifier la réponse sont de différents niveaux de généralité. Il permet néanmoins dans notre

cas d'accentuer le degré de pertinence de la réponse trouvée. Ainsi, les deux patrons « <focus> L_VERBE_PRINCIPAL_INDICATIF L_DET_ARTICLE_INDEF <answer> » et « <focus> être L_DET_ARTICLE_INDEF <answer> » vont nécessairement extraire la même réponse. Enfin, dans le cas où plusieurs réponses sont extraites par le même nombre de patrons, le système Œdipe retourne la réponse qui figure en première position dans la liste des réponses possibles.

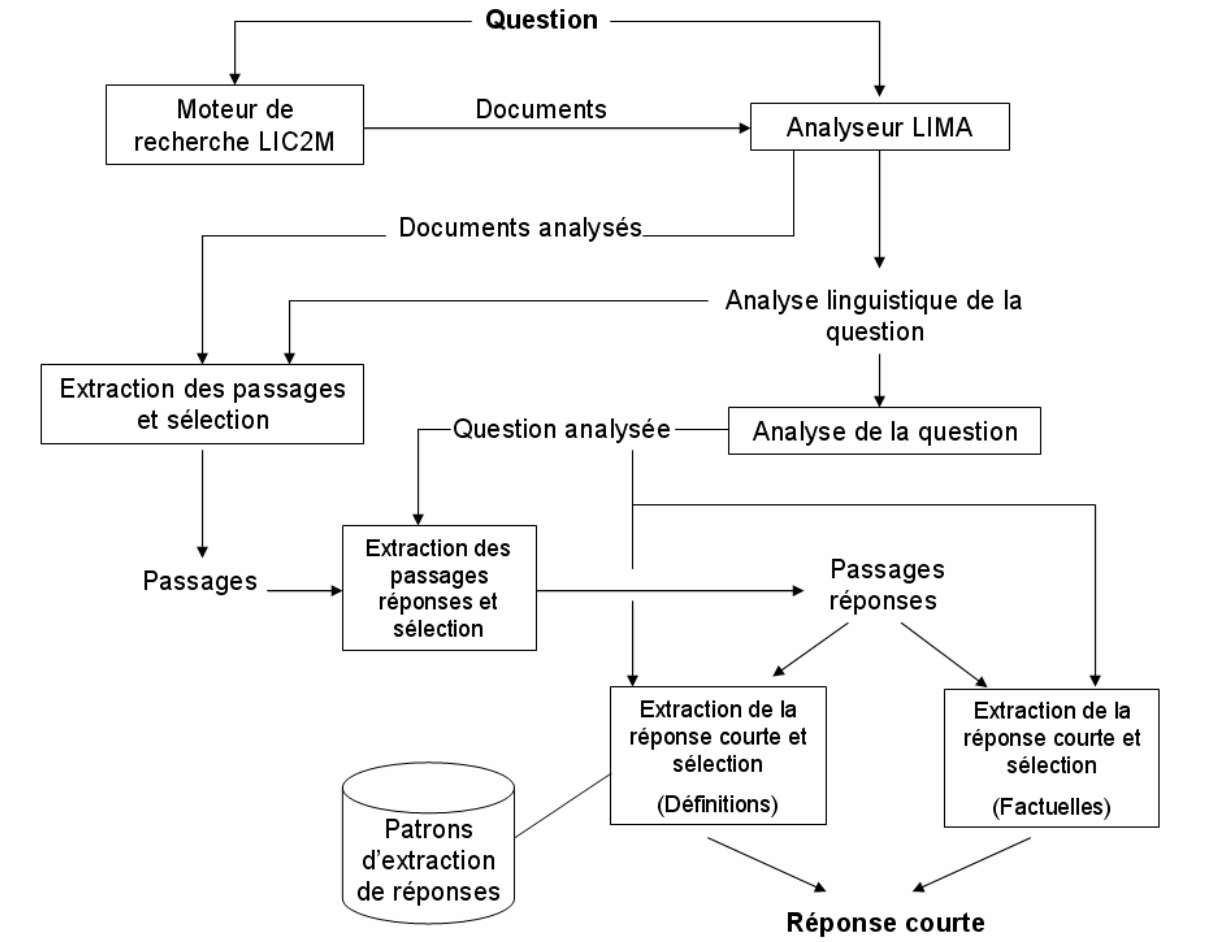


Figure 4.5 Intégration du traitement des questions de définition dans l'architecture d'Œdipe

## 4.6 Conclusion

Nous avons présenté dans ce chapitre le système de question-réponse Œdipe. Le système a été développé pour rechercher des réponses à des questions dans le domaine général. Il est doté d'une architecture classique et s'appuie essentiellement sur l'analyseur syntaxique LIMA.

Pour traiter les questions définitoires, le système utilise une approche fondée sur l'application de patrons linguistiques construits automatiquement à partir de couples exemples « question-réponse », issus des campagnes d'évaluation des systèmes de question-réponse. Cette approche a été évaluée sur les questions définitoires de la campagne d'évaluation CLEF-QA 2006 (cf. Section 6.2.2.2). Enfin, le système Œdipe a participé aux deux campagnes d'évaluation des systèmes de question-réponse EQueR et CLEF-QA. Les résultats obtenus pour la tâche médicale EQueR (cf. Section 6.2.1) confirment que le système Œdipe ne dispose d'aucune compétence lui permettant de trouver des réponses à des questions concernant le domaine médical. Cette difficulté nous a amené à développer un système de question-réponse capable de répondre à des questions sur les bonnes pratiques médicales.

# Cinquième chapitre
# Esculape : guider Œdipe par une ontologie du domaine médical

# 5. Esculape : guider Œdipe par une ontologie du domaine médical

*Ce cinquième chapitre est dédié principalement à la présentation de notre système de question-réponse médical « Esculape ». Le système Esculape est l'extension du système Œdipe, présenté dans le chapitre 4, spécialisé dans le traitement des questions portant sur le domaine médical. Nous décrivons ainsi les deux phases importantes du système Esculape, à savoir l'analyse des questions médicales et la méthodologie employée pour extraire les réponses. Enfin, nous terminons le chapitre par l'évaluation de ces deux phases de traitement.*

## 5.1 Introduction

Le système « Esculape » est un système de question-réponse devant permettre à des professionnels de la santé d'obtenir les informations qui leur sont nécessaires dans des délais compatibles avec leurs activités cliniques. Plus précisément, son rôle est de répondre aux questions portant sur les bonnes pratiques médicales, à l'instar de questions telles que : « Quel est le traitement prescrit dans le cas d'une cirrhose *?* ». Pour ce faire, le système Esculape se fonde à la fois sur la même architecture générale que le système de question-réponse Œdipe, présenté dans le chapitre 4, et sur les mêmes outils de base, à savoir le moteur de recherche du LIC2M pour la sélection des documents candidats et l'analyseur linguistique LIMA pour l'analyse des questions et des documents sélectionnés. Cependant, à la différence du système Œdipe, Esculape repose principalement sur l'exploitation de ressources sémantiques spécifiques au domaine médical (cf. Section 3.2.1) lui permettant de prendre en compte les connaissances de la spécialité nécessaires pour trouver des réponses à des questions médicales. Ces connaissances interviennent plus spécifiquement au niveau de l'analyse des questions et de la phase finale d'extraction des réponses.

Dans la section qui suit, nous évoquerons la prise en compte des questions médicales par le système Esculape en commençant par expliciter la classification des questions médicales sur laquelle nous nous sommes appuyé, puis en présentant la modélisation sémantique des

questions et enfin, en exposant la méthodologie employée par le système pour le typage de ces dernières.

## 5.2 Taxinomie des questions

De nombreux travaux ont étudié plus particulièrement la classification des questions dans des domaines variés. Cette classification s'appuie sur les éléments caractéristiques de la question (pronom interrogatif, entité nommée, …) et sert de support à la détermination du type de réponse attendu, ce qui permet en final de sélectionner une méthode appropriée de recherche de réponses et dans une moindre mesure de construire des requêtes d'interrogation pour un moteur de recherche. Dans le contexte du « question/réponse » en domaine ouvert, des taxinomies de questions ont été constituées afin de déterminer des stratégies de réponse spécifiques pour les différents types de questions définis dans la taxinomie (Harabagiu et al., 2002 ; Hovy et al., 2001b ; Yu et al., 2005). Parmi les études qui nous semblent les plus intéressantes concernant la classification des questions dans un domaine général, on notera celles de (Lehnert, 1978), (Harabagiu et al, 2002), (Woods et al., 2000) et enfin (Lavenus et al., 2004).

Wendy Lehnert (Lehnert, 1978) propose, dans le cadre de son système QUALM, une catégorisation des questions permettant de regrouper les questions selon des types conceptuels (cf. Tableau 5.1). Cette classification a été également reprise par (Graesser et al., 1985) et étendue avec les quatre catégories conceptuelles suivantes : « définition », « exemple », « comparaison » et « interprétation ». (Harabagiu et al., 2002) définissent une taxinomie des questions pour le domaine général fondée principalement sur le pronom interrogatif de la question, généralement révélateur du type de la réponse à rechercher :

- **What** : basic what, what who, what when, what where ;
- **Who** ;
- **How** ;
- **Where** ;
- **Which** : which who, which when, which what ;
- **Name** : name who, name where, name what ;
- **Why** ;
- **Whom**.

| | |
|---|---|
| **Causal antecedent** | Antécédent causal |
| **Goal orientation** | Orientation de but |
| **Enablement** | Capacité |
| **Causal consequent** | Conséquent causal |
| **Verification** | Vérification |
| **Disjunctive** | Disjonction |
| **Instrumental / procedural** | Instrumentale / procédurale |
| **Concept completion** | Complétion de concept |
| **Expectational** | Prévisionnel |
| **Judgmental** | Jugement |
| **Quantification** | Quantification |
| **Feature specification** | Spécification de propriétés |
| **Request** | Requête |

Tableau 5.1 La classification de Lehnert (Lehnert, 1978)

(Woods et al., 2000) présente une taxinomie des questions presque similaire à celle de (Harabagiu et al., 2002) en s'appuyant sur le principe qu'une classification fournit des indices sur le type de la réponse attendu (cf. Tableau 5.2). Partant de la même idée, (Lavenus et al., 2004) propose une typologie des questions, composée de six types de questions, fondée sur le type de la réponse attendue : « définitions », « explications », « entités nommées », « entités », « actions » et « autres ». Cette taxinomie s'appuie également sur le pronom interrogatif de la question ainsi que sur l'objet sur lequel porte la question.

| Question Word | Answer type |
|---|---|
| Whose | Person |
| (who, whom) | Person |
| (which, what) | Depend on subsequent words |
| Prep (which, what, whom) | Do case analysis |
| When | Date |
| Where | Location |
| Why | Reason |
| How (many, much, long) | Number |
| How (adjective) | Number |
| How | Way |
| Name | Name |
| Otherwise | Look for embedded question words |

Tableau 5.2 Taxinomie des questions selon (Woods et al., 2000)

Les différentes approches utilisées pour la classification des questions présentées ci-dessus montrent que la plupart des travaux s'appuient sur le pronom interrogatif pour déterminer le

type de la réponse attendue. Cependant, cette propriété n'est généralement pas suffisante, surtout si le type de la réponse recherchée n'est pas une entité nommée. Par exemple, la question « Quel métal fond à 1500°C ? » attend comme réponse une entité sémantique « métal ». Ainsi, (Hovy et al., 2001a) propose une classification fondée sur l'association de critères surfaciques, comme le pronom interrogatif utilisé, et d'entités sémantiques présentes dans la question.

Dans le domaine médical, on recense globalement les mêmes types de questions que dans le domaine général, excepté que l'information recherchée concerne une connaissance médicale. De ce fait, on peut considérer que les indices exploités pour parvenir à catégoriser les questions portant sur un domaine général peuvent également être utilisés dans le cadre d'une classification des questions médicales, c'est-à-dire en prenant en compte les pronoms interrogatifs des questions et les entités médicales présentes dans les questions. Dans ce contexte, les études menées par (Ely et al., 1999) et (Ely et al., 2000) constituent une référence pour de nombreux travaux en question-réponse dans le domaine médical. Ils proposent une approche fondée sur une taxinomie syntaxico-sémantique des questions médicales (cf. Figure 5.1).

Dans une première expérience, (Ely et al., 1999) ont proposé une classification fondée sur l'étude d'un ensemble de 1101 questions collectées auprès de 103 professionnels de santé. Les questions ont ainsi été classées en 69 catégories selon la spécialité sur laquelle portait la question. Un classement des dix questions les plus fréquemment posées est donné dans le Tableau 5.3. La seconde expérience menée par (Ely et al., 2000) consistait à analyser une collection de 4653 questions recueillies auprès d'une centaine de médecins généralistes dont les 10 premières questions revenant le plus souvent sont classées dans le Tableau 5.4. Ces deux travaux ont inspiré une classification des questions orientée par l'approche « médecine factuelle ». Lorsqu'il s'agit de questions portant sur la prise en charge des patients, les décisions à prendre reposent sur l'EBM (Evidence Based Medicine)[52] (Gorman et al., 1994). L'EBM est une approche permettant de déterminer le type de soin pour chaque patient en s'appuyant sur la meilleure pratique issue de la recherche médicale.

[52] Médecine factuelle fondée sur des preuves, i.e. sur les données actuelles de la science (Sackett, 1997).

| Questions | Nombre | % |
|---|---|---|
| What is the cause of symptom X? | 94 | 9 |
| What is the dose of drug X? | 88 | 8 |
| How should I manage disease or finding X? | 78 | 7 |
| How should I treat finding or disease X? | 75 | 7 |
| What is the cause of physical finding X? | 72 | 7 |
| What is the cause of test finding X? | 45 | 4 |
| Could this patient have disease or condition X? | 42 | 4 |
| Is test X indicated in situation Y? | 41 | 4 |
| What is the drug of choice for condition X? | 36 | 3 |
| Is drug X indicated in situation Y? | 36 | 3 |

Tableau 5.3 Classement des 10 questions les plus fréquentes selon (Ely et al., 1999)

| Questions | Nombre | % |
|---|---|---|
| What is the drug of choice for condition X? | 150 | 11 |
| What is the cause of symptom X? | 115 | 8 |
| What test is indicated in situation X? | 112 | 8 |
| What is the dose of drug X? | 94 | 7 |
| How should I treat condition X? | 82 | 6 |
| How should I manage condition X? | 67 | 5 |
| What is the cause of physical finding X? | 67 | 5 |
| What is the cause of test finding X? | 64 | 5 |
| Can drug X cause finding Y? | 59 | 4 |
| Could this patient have condition X? | 51 | 4 |

Tableau 5.4 Classement des 10 questions les plus fréquentes selon (Ely et al., 2000)

En allant plus loin dans l'analyse, (Ely et al., 2002) ont développé une classification des questions médicales fondée sur la preuve (Evidence Taxonomy) (cf. Figure 5.1). Cette taxinomie a pour but d'identifier les questions auxquelles il est possible de trouver une réponse. Selon les auteurs, seules les questions se référant à des données probantes sont

susceptibles de se voir apporter une réponse en s'appuyant principalement sur les ressources médicales existantes, par exemple les questions portant sur des situations cliniques. Ces questions sont considérées comme « bonnes », à l'instar d'une question telle que « *Quel est le médicament approprié pour une maladie X* ? ». En revanche, les questions portant sur des informations précises concernant un patient s'avèrent difficile à traiter car la réponse ne peut pas être trouvée dans la littérature médicale. Partant de la classification proposée par (Ely et al., 2002), (Yu et al., 2005) ont étudié la possibilité de classer automatiquement les questions médicales. Les résultats obtenus lors de cette étude ont montré que l'identification des concepts UMLS dans les questions permet d'augmenter la pertinence de la classification.

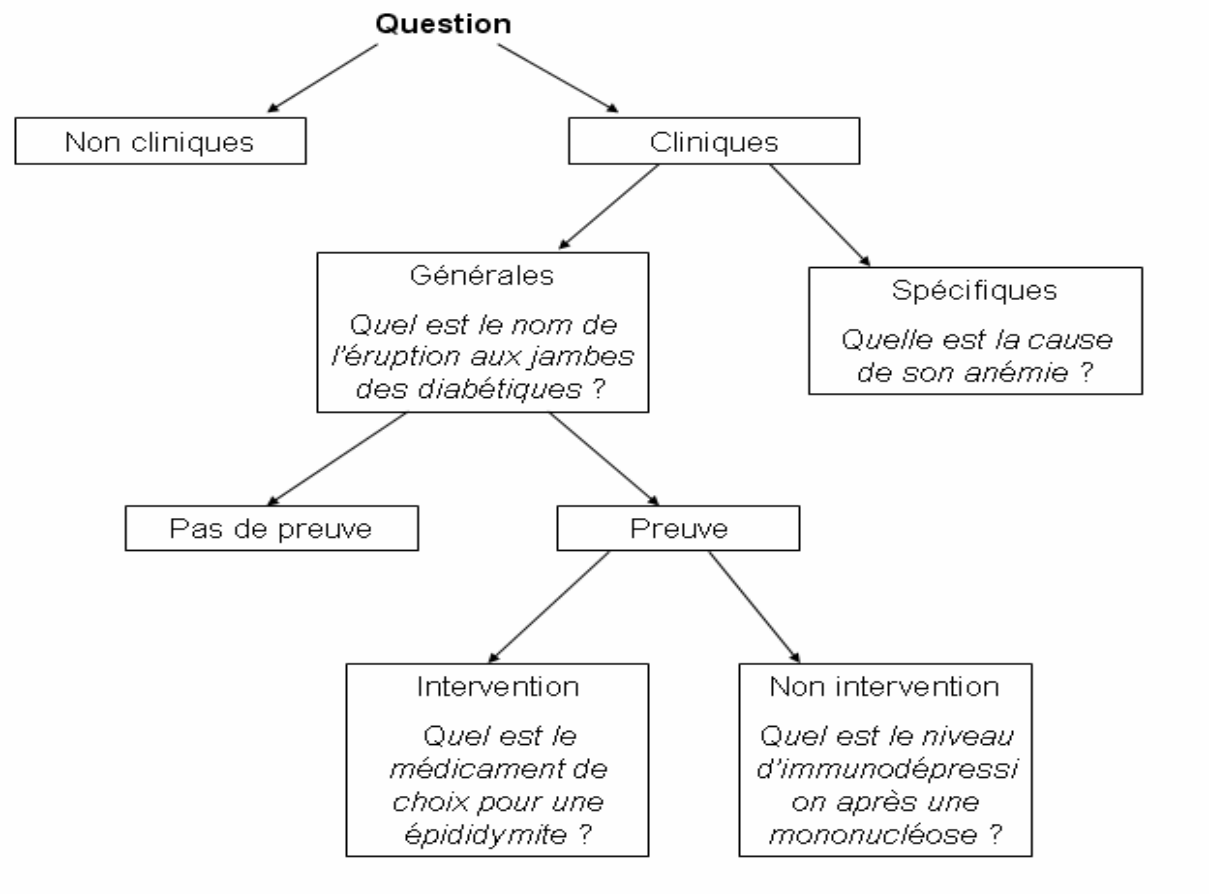


Figure 5.1 Classification fondée sur la preuve (Ely et al., 2002)

La taxinomie des questions que nous proposons dans le cas du système de question-réponse Esculape a été élaborée manuellement après analyse et synthèse des différentes études sur la classification et la nature des questions médicales. Pour ce faire, nous avons également exploité le jeu de questions concernant la tâche médicale proposé lors de la campagne d'évaluation des systèmes de question-réponse EQueR (collection de 200 questions) ainsi que

d'autres questions typiquement posées par des médecins généralistes, sollicités dans le cadre du projet GPM (Guide des bonnes pratiques médicales). Notre classification est largement inspirée des travaux réalisés par Ely et ses collègues (Ely et al., 1999 ; Ely et al., 2000). Cependant, nous nous sommes intéressé plus particulièrement aux types de questions (factuels) pour lesquels le système Esculape peut effectivement retourner une réponse, c'est-à-dire aux questions portant sur les entités médicales étudiées. De ce fait, nous avons décidé de classer les questions selon deux grandes catégories (voir Figure 5.2) : diagnostique vs non-diagnostique, où la catégorie « diagnostique », qui regroupe toutes les questions portant sur des entités médicales, est divisée en plusieurs sous-classes sémantiques reflétant le type de la réponse attendue : général, définition, booléen et enfin entité médicale, qui concerne uniquement les types sémantiques du domaine médical traités, à savoir Maladie, Médicament, Traitement, Examen clinique et Symptôme. La classe « général », quant à elle, regroupe toutes les questions médicales dont la réponse recherchée concerne les autres informations cliniques comme les effets secondaires, les contre-indications, les causes, etc.

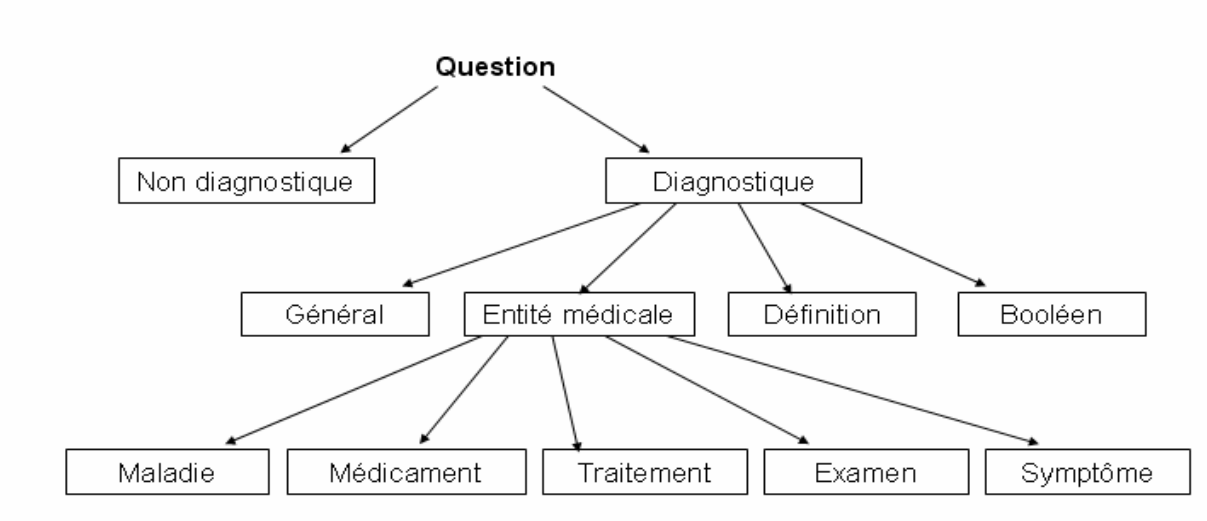


Figure 5.2 Classification des questions médicales du système Esculape

## 5.3 Modélisation des questions

Pour aller plus loin dans l'analyse des questions, certains travaux s'intéressent à la modélisation sémantique des questions (Jacquemart et al., 2003 ; Mendes et Moriceau, 2004). L'objectif de cette modélisation est d'aider le système à interpréter les questions formulées en langage naturel en construisant une représentation de leur sens correspondant au besoin du processus de recherche d'un système de question-réponse.

Afin de réaliser une modélisation sémantique des questions, nous adoptons une approche guidée par une ontologie. Cette approche consiste d'abord à reconnaître les entités médicales présentes dans la question ainsi que le type de la réponse attendue (entité médicale). Ensuite, nous déterminons le type de la relation candidate intervenant entre les occurrences des deux entités reconnues (le type de l'entité médicale de la question et le type de la réponse attendue) parmi les relations sémantiques de l'ontologie médicale (cf. Section 2.3) ; il s'agit, en effet, de faire correspondre la relation identifiée dans la question avec l'une des relations présentes dans l'ontologie.

Dans le système Esculape, nous représentons plus précisément les questions médicales par le triplet (Concept-Question, Relation-Sémantique, Concept-Réponse) :

**- Concept-Question :** c'est l'entité médicale présente dans la question, déterminée par l'application des règles de reconnaissance d'entités médicales (cf. Section 3.2.2).

**- Relation-Sémantique :** c'est la relation sémantique, sélectionnée dans l'ontologie médicale, compatible avec la relation exprimée dans la question correspondant au lien sémantique intervenant entre le type de l'entité de la question et le type sémantique de la réponse désirée.

**- Concept-Réponse :** c'est le type de la réponse attendue (entité médicale). Ce type est identifié en appliquant les règles de typage des questions médicales présentées à la Section 5.4.

La représentation des questions dans le système Esculape est illustrée par les exemples donnés ci-dessous :

1- Quel médicament faut-il utiliser pour soigner l'asthme ?
*Entité médicale de la question : Maladie « Asthme »*
*Type de la réponse attendue : « Médicament »*
*Relation sémantique intervenant entre les deux types : « Soigne »*
*→ Représentation sémantique de la question : (Asthme, Soigne, « Médicament »)*

2- Quel médicament est contre-indiqué en cas d'une galactosémie congénitale ?
*Entité médicale de la question : Maladie «* galactosémie congénitale *»*
*Type de la réponse attendue : « Médicament »*
*Relation sémantique intervenant entre les deux types : « Contre-indication »*
*→ Représentation sémantique : (*Galactosémie congénitale, *Contre-indication, « Médicament »)*

3- Quel est le traitement de la maladie de Parkinson ?
*Entité médicale de la question : Maladie « maladie de Parkinson »*
*Type de la réponse attendue : « Traitement »*
*Relation sémantique intervenant entre les deux types : « Traite »*
*→ Représentation sémantique : (Maladie de Parkinson, Traite, « Traitement »)*

4- Quel est le traitement prescrit à un patient dyslipidémique ?
*Entité médicale de la question : « None »*
*Type de la réponse attendue : « Traitement »*
*Relation sémantique intervenant entre les deux types : « None »*
*→ Représentation sémantique : (None, None, « Traitement »)*[53]

Bien que ce modèle par triplet soit suffisamment expressif pour couvrir la plupart des questions, certaines questions médicales ne rentrent pas dans ce modèle de représentation. Les questions portant sur des explications ou des justifications ne sont pas sous-tendues par une relation simple et il est alors impossible d'appliquer les patrons d'extraction de réponses. Dans ce cas, le système Esculape procède à une recherche classique qui consiste à attribuer un poids aux passages réponses par rapport à la présence des termes de la question dans le passage.

[53] NONE : Terme employé lorsque le type sémantique concerné n'est pas une entité médicale ou encore si la relation sémantique n'est pas identifiée dans l'ontologie médicale.

## 5.4 Analyse des questions

Dans le cadre du système de question-réponse Esculape, une analyse spécifique des questions médicales est réalisée, son résultat influençant directement la stratégie de recherche adoptée. L'objectif principal de cette étape est d'attribuer une catégorie sémantique à la question suivant une classification des questions médicales établies préalablement. Plus précisément, la spécificité de cette tâche dans Esculape est l'identification de la relation sémantique exprimée dans la question, avec comme référence les relations existantes dans notre ontologie médicale (voir Figure 2.1). Cette identification a pour but de sélectionner la liste de patrons d'extraction de réponses (cf. Section 5.5.2) à appliquer pour extraire l'information recherchée. En outre, l'analyse de la question permet de préciser le type de la réponse attendue habituellement formalisé sous la forme d'une entité médicale. Pour notre étude, nous nous sommes intéressé aux questions portant sur les types sémantiques du domaine médical que nous avons sélectionnés, c'est-à-dire les entités Maladie, Médicament, Traitement, Examen et Symptôme. Lors de cette analyse des questions, sont également extraites les entités médicales présentes dans la question.

Le but de l'analyse de la question est de fournir un modèle sémantique de la question (cf. Section 5.3) permettant de spécifier la liste de patrons d'extraction de réponses à employer. Cette analyse est réalisée en deux étapes. La première étape de l'analyse repose sur l'identification de l'entité médicale présente dans la question en utilisant le module de reconnaissance d'entités nommées médicales présenté précédemment (cf. Section 3.2.2). La seconde étape quant à elle permet d'attribuer une catégorie à la question. Pour ce faire, nous appliquons à toutes les questions un ensemble de règles de typage (cf. Section 4.5.1), écrites manuellement, fondées essentiellement sur la forme de la question et prenant plus particulièrement en compte le type de l'interrogatif et le focus (cf. Annexe 4 pour la liste de ces règles et le Tableau 5.5 pour une synthèse de leur volumétrie). Ces règles, qui expriment des patrons morpho-syntaxiques faisant intervenir les mots ou leur catégorie morpho-syntaxique, suivent le même schéma que celui adopté pour la reconnaissance des entités médicales dans les textes (cf. Section 3.2.2) : elles s'appuient principalement sur un élément déclencheur (pronom interrogatif, mot-clé, ...) ainsi que sur des contraintes s'appliquant aux contextes précédent et suivant de ce déclencheur. À chaque catégorie de questions est associé

un type de réponse attendu et un type de relation exprimée dans la question. Ainsi, l'analyse de la question doit produire :

- **Le focus de la question :** pour chaque question, la méthode d'analyse repère l'objet sur lequel porte la question (focus), c'est-à-dire l'objet qui a le plus de chances d'être présent dans la phrase réponse.

*Question : Quel est le traitement contre le paludisme ?*
*Le focus de la question : Maladie « paludisme »*

*Question : Qu'est ce que la schizophrénie ?*
*Le focus de la question : Maladie « schizophrénie »*

- **Le type de la question :** cette caractéristique de la question permet à la fois de définir les stratégies de réponse à utiliser et de distinguer les questions auxquelles le système Esculape peut répondre. Dans le cas d'une question factuelle ou d'une définition portant sur un type d'entité médicale traitée le système retourne comme réponse une entité médicale du type concerné tandis que pour les questions portant sur les autres types de questions (explication, justification, …), Esculape retourne un passage réponse de 250 caractères.

*Question : Quel est l'examen qui permet de repérer une sciatique ?*
*Type de la question : Factuel-Quel-Examen*

*Question : Comment se déroule la kinésithérapie respiratoire ?*
*Type de la question : Général-Comment*

- **Le type de la réponse attendue :** l'analyse de la question permet de déterminer si le type de la réponse attendue correspond à une entité médicale reconnue par le système de question-réponse Esculape. Les entités repérées sont : Maladie, Médicament, Traitement, Symptôme, et Examen. Dans le cas où la réponse attendue n'est pas une entité médicale traitée, le système Esculape attribue une classe sémantique générale correspondant, pour certaines questions, à l'entité exprimée dans la question. Les deux questions suivantes illustrent ce typage :

*Question : Quel est le médicament le plus efficace contre la polyarthrite ?*
*Type de la question : Factuel-Quel-Médicament*
*Type de la réponse attendue = Entité médicale « Médicament »*

*Question : Quel germe est responsable de la pneumonie ?*
*Type de la question : Factuel-Quel-Germe*
*Type de la réponse attendue = Entité générale « Germe »*

**- La relation de la question :** la détection de la relation à rechercher dans la question permet de sélectionner la liste des patrons d'extraction de réponses à appliquer aux phrases candidates. La relation exprime le lien sémantique intervenant entre l'entité médicale présente dans la question et le type de la réponse attendue correspondant aux relations étudiées. Cependant, la relation binaire est parfois inexistante car d'une part, les questions ne contiennent pas toujours une entité médicale et d'autre part, certaines questions n'attendent pas une entité médicale comme réponse, comme les questions portant sur des explications.

*Question : Comment traiter la varicelle ?*
*Type de la question : Factuel-Comment-Traiter*
*Le focus de la question : Maladie « varicelle »*
*Type de la réponse attendue : Entité médicale « Traitement »*
*Relation exprimée : Traite (Maladie - Traitement)*

*Question : Quel médicament est contre-indiqué en cas de maladie de Parkinson ?*
*Type de la question : Factuel-Quel-Médicament-Contre_indiqué*
*Le focus de la question : Maladie « maladie de Parkinson »*
*Type de la réponse attendue : Entité médicale « Médicament »*
*Relation exprimée : Contre-indication (Maladie - Médicament)*

Voici quelques exemples de règles utilisées pour catégoriser les questions médicales permettant ainsi de déterminer le type de la réponse attendue et la relation exprimée :

[Quel] **::** [est] [$L_DET] [@Traitement] [$L_DET] *{1-10} [\ ?] **:** F-Quel-Traitement[54]
→ Quel est le traitement du chérubisme ?

[54] @ regroupe une liste de termes marquant la présence du concept.

[Quel] **::** [@Médicament] [*{1-5}] [soigner$L_V] *{1-10} [\ ?] **:** F-Quel-Médicament
→ Quel médicament faut-il utiliser pour soigner la démangeaison ?

[Quel] **::** [@Symptôme] [caractériser$L_V] [$L_DET] *{1-10} [\ ?] **:** F-Quel-Symptôme
→ Quel symptôme caractérise la maladie infectieuse de la toxoplasmose ?

[Quel] **::** [@Examen] [permettre$L_V] [*{1-5}] [diagnostic] *{1-10} [\ ?] **:** F-Quel-Examen
→ Quel examen permet le diagnostic de cancer du côlon ?

Le tableau ci-dessous récapitule le nombre de règles construites permettant au système de question-réponse Esculape de réaliser le typage des questions médicales afin d'identifier le type de la réponse attendu. On remarquera le nombre élevé de règles concernant l'entité médicale « Traitement » par rapport aux autres types d'entités. Ce nombre est obtenu en raison de la diversité des questions portant sur ce type sémantique. Par ailleurs, pour le typage des questions définitoires, nous utilisons le même ensemble de règles définies précédemment (cf. Section 4.5.1).

| **Type de la réponse ⟶ Type de relation** | **Nombre de règles** |
|---|---|
| Traitement ⟶ Traite (Maladie-Traitement) | 24 |
| Médicament ⟶ Soigne (Maladie-Médicament) | 11 |
| Symptôme ⟶ Caractérise (Maladie-Symptôme) | 12 |
| Examen ⟶ Détecte (Maladie-Examen) | 11 |
| Définition | 29 |
| Général | 32 |
| **Total** | **119** |

Tableau 5.5 Nombre de règles de typage

## 5.5 Extraction des réponses

La phase d'extraction de réponses constitue le dernier traitement dans le processus de recherche de la réponse à une question dans un système de question-réponse. L'objectif de cette étape est de localiser dans un premier temps les passages pertinents pour la question

posée à partir d'un ensemble de passages candidats, puis d'extraire la réponse recherchée. Afin d'établir un lien entre la question et le passage contenant la réponse, un système de question-réponse doit s'appuyer sur les indices dérivés de la question au préalable. Pour extraire la réponse, différentes approches peuvent être appliquées dont la plupart se fondent sur les moyens d'analyse linguistique élaborés.

La détermination du type de la réponse attendue est la tâche de l'analyse des questions commune à la plupart des systèmes de question-réponse puisque l'information définie permet de caractériser le type de l'entité pouvant correspondre à la réponse souhaitée. Ce type peut intervenir à la fois dans la sélection des passages candidats susceptibles de contenir la réponse à partir des documents retournés par le moteur de recherche et dans le choix de la méthode à utiliser pour extraire la réponse concernée. Les types attendus peuvent différer d'un système à l'autre ; ils peuvent concerner des types d'entités nommées (noms propres, expressions temporelles et expressions numériques) ou des catégories particulières spécifiques à chaque système qui relèvent parfois d'un domaine restreint (Rosario, 2005 ; Embarek et al. 2007). La classification des types de réponse possibles est préalablement définie selon les types d'entités pouvant être reconnus par le système de question-réponse. Lorsque le type de la réponse concerne une entité nommée, cette dernière est facilement repérable dans les textes. En effet, il est possible d'identifier et de catégoriser les entités nommées par des modules spécialisés, et l'extraction de la réponse correcte s'avère alors plus facile. Cependant, dans tous les cas, l'extraction des réponses consiste à réaliser un appariement entre la représentation de la question et celle des phrases candidates sélectionnées à l'issue des étapes précédentes. Cette représentation vise à rechercher un lien entre les termes composant la question et les termes employés dans les passages réponses.

Dans le but d'apparier les questions et les passages candidats, diverses approches peuvent être adoptées. Ces techniques reposent sur des ressources et des modes de raisonnement variés. De ce fait, les systèmes de question-réponse se différencient notamment par leur mode de représentation et de sélection des passages pertinents. Parmi ces différentes approches, on peut en distinguer trois principales.

Une première approche se fonde sur les variations terminologiques. Il s'agit de repérer les variantes lexicales des termes de la question dans les textes pour sélectionner les passages

candidats (Ferret et al., 2002b ; Nyberg et al., 2003). Cette sélection des passages est également déterminée par la présence ou l'absence du type attendu pour l'entité réponse. Certains systèmes de question-réponse complètent ces informations par le calcul d'une mesure de proximité entre les différentes entités lexicales qui permet d'évaluer leur densité dans les extraits de textes (Gillard et al., 2006). (Yang et al., 2002) ont quant à eux proposé une méthode permettant d'identifier les variantes ou les co-occurrents des termes de la question dans les passages candidats en exploitant essentiellement des ressources sémantiques comme WordNet.

Une fois les passages sélectionnés, la réponse peut être extraite soit en identifiant directement dans le passage l'entité qui correspond au type de la réponse attendue dans le cas des entités nommées ou en appliquant des patrons d'extraction, souvent sous forme d'expressions régulières opérant au niveau morpho-syntaxique, lorsque la réponse attendue n'est pas typée. Dans de nombreux systèmes, ces patrons sont construits manuellement pour chaque type de questions puis appliqués dans les passages sélectionnés pour identifier des réponses (Soubbotin et al., 2001). Un autre mode d'extraction consiste tout simplement à reformuler la question sous une forme affirmative (Banko et al., 2002).

Une deuxième approche utilisée pour la sélection de la réponse repose sur l'appariement des représentations sémantiques de la question et des passages candidats. Cette représentation peut être sous la forme de représentations logiques (Ahn et al., 2004) ou de schémas prédicatifs construits à partir de bases de connaissances externes telles que FrameNet (Narayanan et al., 2004) ; mais elle peut également rester plus proche de la forme des textes (Ferret et al., 2001b).

Le recours à des bases de connaissances est généralement nécessaire pour effectuer certains appariements. Ces ressources sont très utiles pour fournir les différentes variantes lexicales et sémantiques des mots. Ainsi, (Moldovan et al., 2003b) s'appuient principalement sur le réseau sémantique WordNet afin de récolter pour chaque terme de la question un ensemble de variantes morphologiques. La question et les phrases candidates sont ensuite transformées en une représentation logique et appariées par le biais d'une démonstration automatique.

Une dernière approche, employée par un grand nombre de systèmes de question-réponse, s'intéresse plus particulièrement à des patrons d'extraction de réponses. Cette approche utilise un ensemble de patrons linguistiques (expressions régulières) afin d'extraire l'information recherchée en la localisant par rapport aux entités de la question présentes au niveau de la phrase candidate. Les patrons s'appuient sur les résultats d'outils du traitement automatique des langues allant de l'étiquetage morpho-syntaxique et de la lemmatisation des mots jusqu'à l'analyse syntaxique des phrases en passant par la reconnaissance d'entités nommées. Toutefois, cette approche est limitée par la couverture des patrons linguistiques construits. En effet, la majorité des systèmes se fondent sur l'application de patrons lexico-syntaxiques construits manuellement (Soubbotin et al., 2001), à partir d'une sélection d'exemples de question/réponse correspondant à une catégorie de question et ne prennent guère en compte les variations terminologiques au niveau des patrons.

Pour notre étude dans le domaine médical, nous avons retenu une approche utilisant des patrons linguistiques appris automatiquement, à l'image de ce que nous avons fait en domaine ouvert pour les questions définitoires. Outre le fait qu'une telle approche apparaît comme un bon compromis entre efficacité et difficulté de mise en œuvre, elle présente l'avantage de s'appuyer sur les mêmes ressources que celles constituées pour l'extraction de relations sémantiques. Les patrons appris pour cette extraction sont en effet directement applicables pour extraire des réponses. Du point de vue proprement dit de l'extraction de réponses, les patrons sont appris en utilisant l'algorithme proposé par Ravichandran (Ravichandran, 2005) concernant l'apprentissage de patrons multi-niveaux, ce qui représente une extension de (Ravichandran et al., 2002). Une liste de schémas lexico-syntaxiques est constituée pour chaque relation sémantique traitée. Pour sélectionner une liste de patrons linguistiques appropriée, le système Esculape repose sur le module d'analyse de la question qui permet de différencier dans la question la relation sémantique sur laquelle porte la question. Les patrons concernés sont ensuite appliqués dans les textes afin d'extraire la réponse à la question posée.

Dans la suite de cette section, nous présenterons dans un premier temps le processus employé pour acquérir les patrons d'extraction de réponses, puis nous exposerons leur utilisation pour trouver des réponses à des questions médicales dans les passages sélectionnés.

### 5.5.1 Apprentissage de patrons d'extraction de réponses

Dans Esculape, un patron d'extraction de réponses est une formule linguistique qui permet de localiser les réponses à extraire à partir des passages candidats et établit un lien entre le focus de la question et l'élément recherché. Il se caractérise par des contraintes lexico-syntaxiques (informations issues d'une analyse morphologique et d'un étiquetage morpho-syntaxique) et sémantiques (dans notre cas, des types d'entités médicales).

Comme nous l'avons vu ci-dessus, la plupart des approches fondées sur l'application des patrons d'extraction de réponses s'appuient sur des patrons écrits manuellement. Ces patrons, bien que performants, présentent souvent un problème de couverture car il est difficile d'être exhaustif quant aux variations terminologiques des éléments linguistiques qui composent le patron. Plus globalement, il n'est pas toujours facile de construire manuellement un ensemble de patrons représentant un compromis acceptable entre rappel et précision. En effet, une définition manuelle a naturellement tendance à conduire à un ensemble plutôt restreint de patrons assez généraux. Or, sans aller jusqu'à une approche de type memory-based learning (Daelemans et al., 2005), il semble qu'une approche axée sur la construction d'un nombre assez large de patrons issus de généralisations limitées soit plus à même de donner de bons résultats. C'est d'ailleurs ce que suggère le travail rapporté dans (Soubbotin et al., 2001), avec la limite dans ce cas que les multiples patrons étaient définis manuellement, impliquant ainsi un effort difficilement généralisable. De ce fait, de nombreux travaux sur l'apprentissage de patrons linguistiques s'intéressent de plus en plus à des méthodes fondées essentiellement sur des techniques d'apprentissage permettant d'apprendre d'une manière automatique ou semi-automatique des motifs d'extraction de réponses (Du et al., 2004 ; Cui et al., 2005).

Dans le cadre du système Esculape, l'extraction de la réponse s'effectue en utilisant les mêmes patrons lexico-syntaxiques présentés à la Section 3.3.2. En effet, ces patrons linguistiques, construits initialement pour extraire de nouvelles relations sémantiques, peuvent également permettre au système de trouver des réponses à partir des phrases candidates sélectionnées. De ce fait, leur contribution à Esculape pour fournir des réponses est double : soit la relation sémantique sous-tendant la question posée a été extraite d'un corpus grâce à ces patrons et la réponse est puisée dans la base de connaissances ainsi constituée *a priori* ; soit la relation n'est pas présente dans la base de connaissances d'Esculape et les patrons sont

utilisés pour extraire une réponse à partir du corpus mis à disposition pour ce faire. La méthode d'apprentissage des patrons d'extraction repose quant à elle sur l'approche proposée par Pantel et Ravichandran (Pantel et al., 2004 ; Ravichandran, 2005). Elle étend celle proposée dans (Ravichandran et al., 2002) en permettant d'extraire des patrons lexico-syntaxiques dits « multi-niveaux », c'est-à-dire pouvant faire référence à plusieurs types d'information sur les mots. En l'occurrence, ces patrons combinent leur forme fléchie, leur lemme et leur catégorie morpho-syntaxique.

Pour cette étude, nous avons considéré quatre relations sémantiques intervenant entre l'entité médicale « Maladie » et les autres types d'entités traités, à savoir :

1- la relation « Traite » entre les deux entités « Maladie » et « Traitement »,
2- la relation « Soigne » entre les deux entités « Maladie » et « Médicament »,
3- la relation « Détecte » entre les deux entités « Maladie » et « Examen »,
4- la relation « Caractérise » entre les deux entités « Maladie » et « Symptôme ».

Par ailleurs, pour répondre aux questions définitoires, nous utilisons l'ensemble des patrons linguistiques construits pour le domaine général (cf. Section 4.5).

### 5.5.2 Utilisation des patrons d'extraction de réponses

Afin de retourner des réponses à des questions médicales, le système Esculape repose sur une stratégie guidée par l'ontologie définie dans le deuxième chapitre. Plus précisément, à l'issue de l'analyse de la question, il commence par rechercher la réponse dans la base de connaissances regroupant les instances des relations de l'ontologie qu'il a déjà acquises : si la réponse est trouvée alors cette dernière est directement renvoyée à l'utilisateur. Dans le cas contraire, une réponse est recherchée dans les documents disponibles par application de la liste des patrons d'extraction de réponse appropriée sur les phrases candidates sélectionnées à l'issue des étapes précédentes. Il est à préciser que dans le cas où la question est sous-tendue par une relation, le processus de recherche des passages consiste à se focaliser sur les phrases contenant des couples d'entités du type de celles de la relation, avec l'une des entités correspondant à celle de la question. Une telle approche fondée sur l'utilisation d'une

ontologie se retrouve au niveau des systèmes d'extraction d'information (Le Roux, 2003 ; McDowell et al., 2006 ; Buitelaar et al., 2006) ou encore des systèmes d'annotation automatique (Cao et al., 2006).

Le choix des patrons à appliquer dépend généralement du type de la question posée (entité médicale, définition). De ce fait, le système Esculape s'appuie sur les caractéristiques importantes de la question comme le type de la réponse attendue, le focus et l'entité médicale identifiée. Une fois la liste des patrons déterminée, l'extraction de la réponse s'effectue en recherchant un appariement entre les schémas sélectionnés et les phrases candidates supposées contenir la réponse attendue : si un appariement est trouvé, une réponse est extraite à partir de son emplacement dans le patron. Plus explicitement, le processus implémenté pour trouver des réponses à des questions médicales est le suivant :

Considérons la question : « *Quel est le médicament le plus efficace contre la polyarthrite ?* »

1- analyse de la question : elle permet au système de déterminer le type de la réponse attendue, « Médicament », l'entité médicale présente dans la question, « Polyarthrithe », appartenant à la classe « Maladie », et donc de déduire grâce à un mapping entre le type de la question et le type de la relation présent dans l'ontologie que la relation cible intervenant entre ces deux types d'entités est « Soigne » ;
2- instanciation des patrons de la relation cible : cette opération consiste à remplacer dans tous les patrons le type de l'entité concerné par l'entité médicale de la question. Ici, on remplace « Maladie » par « Polyarthrithe » ;
3- application de l'analyseur linguistique LIMA à chaque phrase candidate sélectionnée par le système Esculape pour obtenir les trois niveaux d'information linguistique des mots ;
4- extraction des réponses : cette tâche consiste à aligner chaque patron instancié avec toutes les phrases candidates. L'alignement débute à partir de l'entité médicale dans la phrase puis une comparaison est effectuée terme par terme entre la phrase et le patron jusqu'à la balise correspondant au type de l'entité attendue. Si le groupe nominal correspond à une entité du même type que celui de la réponse souhaitée, il est extrait comme réponse à la question ;

5- toutes les réponses extraites sont classées selon le nombre de patrons appliqués pour les extraire puis la réponse avec le plus grand nombre est retournée.

Pour les questions définitoires, le système Esculape applique le même processus que celui adopté par le système de question-réponse Œdipe (cf. Section 4.5.2).

## 5.6 Évaluation

Pour évaluer le système de question-réponse tel qu'Esculape, il est nécessaire de disposer d'un ensemble de questions médicales représentatives de l'usage des professionnels de la santé, accompagnées de réponses de référence issues d'un corpus de documents médicaux. Cependant, la seule campagne d'évaluation des systèmes de question-réponse en domaine médical qui a été menée jusqu'à présent est la partie médicale de la campagne EQueR. Ce manque de ressources nous a incité à construire un jeu de questions spécifiques en s'inspirant principalement des questions factuelles (faisant appel à des entités médicales) proposées pour la tâche médicale de la campagne EQueR, à l'image des questions :

« *Quel est le traitement du chérubisme ?* »
« *Quels sont les trois types d'examens à réaliser en cas de suspicion d'un neuroblastome ?* »
« *Quels sont les 4 médicaments qu'il est possible de prescrire dans le cadre d'une ostéoporose corticosonique ?* »

Parallèlement, nous nous sommes appuyés sur un corpus de référence traitant du domaine médical un peu plus large que le corpus EQueR. Pour cela, nous avons décidé d'exploiter le corpus médical utilisé dans le cadre du projet Technolangue Atonant.

Le corpus concerné, d'une taille de 400 Mo, a été divisé en trois sous-corpus d'une taille de 133 Mo chacun. Un premier sous-corpus a été choisi pour apprendre des patrons lexico-syntaxiques caractéristiques des relations sémantiques visées. Le second sous-corpus a été utilisé quant à lui pour enrichir notre ontologie médicale en extrayant de nouvelles relations sémantiques grâce aux patrons appris à partir du premier sous-corpus. Enfin, le dernier sous-corpus a servi de référence pour construire une collection de questions/réponses médicales.

Cette liste de questions/réponses, établie par un médecin généraliste, constitue le corpus d'évaluation que nous avons utilisé pour évaluer le système Esculape quant à ses capacités d'extraction de réponses, Esculape adoptant la procédure de recherche en deux temps comme exposée à la section 5.3.2 en étant doté à la fois des patrons issus du premier sous-corpus et de la base de connaissances constituée grâce au deuxième sous-corpus.

Le nombre de questions de la collection test a été fixé à 100 (20 questions par type de relation) afin d'obtenir une diversité suffisamment grande (cf. Annexe 2 pour la liste de ces questions). Pour que cette évaluation soit véritablement significative vis-à-vis des capacités d'Esculape, nous avons sélectionné les questions pour qu'elles concernent des relations traitées par Esculape, c'est-à-dire des questions portant sur des traitements, des médicaments, des symptômes, des examens et des questions définitoires.

Dans ce qui suit, nous présentons l'évaluation du système Esculape réalisée sur le corpus de questions construit. Cette évaluation concerne à la fois l'analyse des questions et l'extraction des réponses.

### 5.6.1 Évaluation de l'analyse des questions

Les questions médicales définies nous ont servi dans un premier temps à évaluer la méthode d'analyse des questions du système Esculape. L'objectif est plus précisément d'évaluer le typage des questions, c'est-à-dire l'identification du type de la réponse attendue. Nous présentons dans le Tableau 5.6 une évaluation du typage réalisé par Esculape sur l'ensemble des 100 questions de notre corpus d'évaluation. Il est à noter que, mises à part les questions définitoires, toutes les questions factuelles attendent des réponses correspondant à l'un des types d'entité médicale étudiés. Pour ces questions, le type de l'entité médicale attendue comme réponse a été correctement identifié pour 79% des questions et globalement, le module de typage a parfaitement catégorisé 83% des questions. Il faut ajouter que le focus de la question a été reconnu dans 91% des cas.

Le Tableau 5.6 montre que la méthode d'analyse des questions est plus efficace pour typer les questions définitoires que les questions factuelles. Toutes les questions définitoires de la

collection de test ont en effet été correctement typées alors que le taux d'erreurs (21,3%) pour les 80 questions factuelles nécessite de revoir le jeu de règles de typage pour les améliorer et les compléter. Parmi les erreurs observées, on note des erreurs dues à l'identification des entités médicales dans les questions. Par exemple, dans la question « *Comment traiter le schwannome vestibulaire ?* », le terme « schwannome vestibulaire » n'a pas été reconnu comme une maladie. D'autres erreurs concernent les règles elles-mêmes dont la couverture incomplète ne permet pas de catégoriser convenablement des questions médicales telles que : « *Quel bilan doit être effectué pour une pancréatite chronique ?* » où le terme « bilan » n'a pas été identifié comme un examen clinique (la réponse attendue) mais comme un bilan comptable (dont la réponse attendue est une entité numérique).

| Type de question | Nombre de questions | Nombre de types incorrects |
|---|---|---|
| Définition (MD) | 20 | 0 |
| Factuel Traitement (MT) | 20 | 1 |
| Factuel Médicament (MM) | 20 | 4 |
| Factuel Symptôme (MS) | 20 | 6 |
| Factuel Examen (ME) | 20 | 6 |
| **Total** | **100** | **17** |

Tableau 5.6 Les résultats du module de typage des questions d'Esculape

### 5.6.2 Évaluation sur l'extraction des réponses

La deuxième évaluation menée concerne la capacité du système Esculape à trouver des réponses aux questions. Comme nous l'avons vu précédemment, cette tâche est réalisée d'abord en allant consulter une base de connaissances constituée au préalable puis, en cas d'échec de cette première méthode, en extrayant les réponses à partir des textes grâce à des patrons linguistiques (quelques exemples de ces patrons sont donnés à l'Annexe 5). Le même ensemble de patrons étant utilisé dans les deux cas, nous avons d'abord évalué son efficacité dans la constitution préalable d'une base de réponses possibles. Cette première évaluation a été réalisée en utilisant le premier tiers de notre corpus d'évaluation pour l'apprentissage des patrons. Le Tableau 5.7 résume, pour chaque relation étudiée, le nombre de phrases extraites et le nombre de patrons multi-niveaux appris à partir de ce corpus.

| Type de Relation | Nb phrases extraites | Nb patrons multi-niveaux | Nb patrons 2-niveaux[55] |
|---|---|---|---|
| Traite | 3 215 | 1 457 | 131 |
| Soigne | 582 | 149 | 33 |
| Caractérise | 4 850 | 1 697 | 156 |
| Détecte | 1 100 | 419 | 66 |

Tableau 5.7 Résultats de l'apprentissage de patrons lexico-syntaxiques

La constitution de la base de réponses possibles et l'évaluation de la validité des patrons extraits ont été conjointement réalisées en appliquant ces derniers à la deuxième partie de notre corpus d'évaluation. Pour mesurer la validité des nouvelles relations ainsi extraites, et donc indirectement juger de la pertinence des patrons appris, nous avons fait appel à la précision et au rappel, définies comme suit :

- la précision représente le nombre de relations extraites correctes sur le nombre total des relations extraites par notre système ;
- le rappel représente le nombre de relations extraites correctes par notre système sur le nombre total de relations annotées dans le corpus ;
- la F1-mesure représente la moyenne harmonique entre la précision et le rappel.

| Relations | Précision | Rappel | F1-mesure |
|---|---|---|---|
| Maladie – Examen | 0,81 | 0,54 | 0,65 |
| Maladie – Médicament | 0,85 | 0,58 | 0,69 |
| Maladie – Traitement | 0,90 | 0,61 | 0,73 |
| Maladie – Symptôme | 0,77 | 0,47 | 0,58 |
| **Moyenne** | **0,83** | **0,55** | **0,66** |

Tableau 5.8 Résultats de l'extraction de nouvelles relations sémantiques

Le Tableau 5.8 montre que l'extraction de relations obtient sur ce corpus une précision moyenne de 83% et un rappel de 55%. Ce faible taux de rappel s'explique par le fait que les

[55] C'est-à-dire en ne prenant en compte que la forme fléchie et la forme normalisée (lemme) des termes.

patrons appris ne permettent pas de couvrir toutes les formes par lesquelles les relations cibles apparaissent dans le corpus. Cependant, on peut considérer que les relations sémantiques extraites sont d'une bonne fiabilité. La comparaison avec les résultats de l'extraction de relations présentés à la Section 3.4.2 montre que les chiffres sont un peu moins bons mais la tendance générale est la même. Cela est dû en grande partie à la taille du corpus d'évaluation (133 Mo) qui est plus important que celui utilisé pour l'extraction de relations (65 Mo).

Le second volet de cette évaluation concerne la recherche d'une réponse et tout particulièrement l'utilisation des patrons appris précédemment pour l'extraction de réponses à partir de textes. Cette évaluation a été menée sur la troisième partie de notre corpus d'évaluation. Le Tableau 5.9 montre que le système Esculape a renvoyé une réponse correcte (réponse courte) pour 33% des questions soumises. La réponse représente la chaîne extraite par les patrons avec la plus grande fréquence, c'est-à-dire le nombre maximum de patrons utilisés pour l'extraire. À noter qu'aucune réponse n'a été trouvée dans l'ontologie médicale. Cela s'explique par le nombre réduit de relations extraites du deuxième sous-corpus pour peupler l'ontologie (cf. Figure 2.2)[56] et aussi par la nature des questions posées qui ne portaient pas nécessairement sur les entités extraites de ce sous-corpus. Ainsi, l'ontologie est constituée des types suivants : maladie (2128 instances), symptôme (613 instances), traitement (807 instances), examen (939 instances) et enfin médicament (2429 instances).

Les résultats obtenus par le système Esculape pour cette évaluation s'avèrent encourageants puisqu'ils représentent les résultats de la première évaluation du système. Ils nous permettent néanmoins de juger des capacités du système afin d'apporter des améliorations pour les prochaines évaluations. Parmi les sources d'améliorations à étudier, on note la classification des questions. L'analyse effectuée montre que la méthode d'analyse des questions n'a pas permis, pour un certain nombre de questions (17), d'identifier le bon type de la réponse attendue. Dans certains cas, le focus n'a en outre pas été détecté. De tels cas diminuent considérablement les chances du système Esculape de sélectionner la bonne liste des patrons d'extraction à appliquer. Un autre point à améliorer concerne la recherche documentaire et la sélection des passages candidats puisque le moteur de recherche LIC2M n'a renvoyé des documents susceptibles de contenir une bonne réponse que pour 72% des questions alors que

[56] Relation Traite (1155), relation Soigne (725), relation Caractérise (2707) et relation Détecte (593).

le système Esculape n'a retourné des passages réponses que pour 59 questions. Enfin, un dernier point d'amélioration à explorer a trait à l'apprentissage de nouveaux patrons d'extraction de réponses. En effet, l'utilisation des patrons construits n'a pas permis d'extraire certaines réponses présentes dans les passages candidats (9), notamment lorsque la réponse est trop éloignée de l'élément sur lequel porte la question (entité médicale de la question). Comparativement au système Œdipe (cf. Section 6.2.1), même si on se focalise ici sur un ensemble de types de questions limités, le système Esculape se distingue par sa capacité à trouver des réponses courtes aux questions médicales.

| Type de question | Nb réponses correctes | Nb réponses incomplètes | Nb réponses incorrectes | Nb absences de réponses | Total |
|---|---|---|---|---|---|
| Définition (MD) | 13 | 4 | 1 | 2 | 20 |
| Factuel (MT) | 6 | 3 | 7 | 4 | 20 |
| Factuel (MM) | 5 | 1 | 13 | 1 | 20 |
| Factuel (MS) | 4 | 0 | 10 | 6 | 20 |
| Factuel (ME) | 5 | 0 | 7 | 8 | 20 |
| Total | 33 | 8 | 38 | 21 | 100 |

Tableau 5.9 Résultats du module d'extraction de réponses du système Esculape[57]

## 5.7 Conclusion

Nous avons présenté dans ce chapitre le système de question-réponse Esculape que nous avons développé pour répondre à des questions portant sur les bonnes pratiques médicales. À la différence du système Œdipe, Esculape s'appuie principalement sur des connaissances sémantiques du domaine médical. Cette caractéristique lui permet de mieux analyser les questions posées, de repérer les entités médicales dans les passages candidats et d'extraire les réponses en fonction des relations intervenant entre ces entités. Afin d'évaluer ses capacités, il est apparu nécessaire de disposer d'un corpus de questions proche de l'usage réel des

[57] Réponse incomplète : une réponse a été jugée incomplète dans le cas où le système ne retourne qu'une partie de la réponse.
Réponse incorrecte : une réponse a été jugée incorrecte lorsque cette dernière est fausse.

professionnels de la santé. De ce fait, et par manque d'exemples de questions posées par les médecins, nous avons construit une collection de questions médicales représentatives de l'usage des praticiens accompagnées de leurs réponses, issues d'un corpus représentatif du domaine considéré. Nous nous sommes spécialement intéressé aux questions portant sur un ensemble limité de types d'entités médicales et sur les relations qu'ils entretiennent.

La méthode d'analyse des questions du système a été évaluée sur l'ensemble des questions élaborées. Les résultats obtenus montrent que la méthode réalise une bonne classification des questions en général et des questions définitoires en particulier. Les différentes informations issues de l'analyse des questions conditionnent de façon importante le succès de l'étape d'extraction de la réponse. Cette dernière se fonde sur une approche en deux temps s'appuyant d'abord sur la consultation d'une base de connaissances construite au préalable puis, si la réponse attendue n'est pas trouvée, sur l'application des patrons d'extraction de réponses spécifiques à la relation cible. Une réponse courte est alors retournée pour chaque question.

De même que pour l'analyse des questions, la tâche d'extraction de réponses a été testée sur les questions associées à notre corpus d'évaluation. L'évaluation montre que le système Esculape obtient des résultats satisfaisants (33 réponses courtes correctes sur les 100 questions soumises). Ce résultat, bien qu'étant obtenu sur un corpus de questions différent, affirme que le système Esculape est plus adapté pour répondre à des questions médicales que le système Œdipe (7 passages correctes sur les 200 questions proposées lors de la campagne EQueR). Cependant, certains points restent à approfondir, comme l'analyse des questions, qui échoue parfois dans la détermination du type attendu de la réponse, et la couverture des patrons d'extraction de réponses.

# Sixième chapitre
# Évaluation

# 6. Évaluation

*Le chapitre 4 et 5 de ce manuscrit nous ont permis de présenter les deux systèmes de question-réponse Œdipe et Esculape : le premier a été développé pour trouver des réponses à des questions en domaine ouvert tandis que le second a été développé spécifiquement pour répondre à des questions médicales. Dans la première partie de ce chapitre, nous présentons l'évaluation du système Œdipe dans deux campagnes d'évaluation des systèmes de question-réponse, à savoir EQueR et CLEF-QA. Dans une seconde partie, nous donnons les résultats de l'évaluation du système Esculape sur le jeu de questions de la tâche médicale EQueR.*

## 6.1 Les campagnes d'évaluation EQueR et CLEF-QA

L'intérêt pour les systèmes de question-réponse a connu un essor important depuis l'introduction de la tâche question-réponse dans différentes campagnes d'évaluation en recherche d'information, à commencer par la piste Question Answering de la campagne TREC (*Text REtrieval Conference*) du NIST, initiée en 1999. La campagne TREC-QA concerne les systèmes travaillant en anglais sur des documents en domaine ouvert. D'autres évaluations comme NTCIR (Test Collection for IR Systems) ou CLEF-QA (Cross Language Evaluation Forum) par exemple s'intéressent aux systèmes monolingues et multilingues[58]. En confrontant les résultats de différents systèmes obtenus dans les mêmes conditions, ces campagnes permettent de dresser un état des lieux des avancées au niveau méthodologique dans le domaine des systèmes de question-réponse en vue d'orienter la conception de ce type de systèmes. Du point de vue des processus de traitement automatique des langues, elles fournissent en outre un contexte d'évaluations de nature applicative.

Dans cette partie, nous nous intéressons aux deux campagnes d'évaluations CLEF-QA et EQUER, auxquelles le système Œdipe a participé, et plus particulièrement à EQueR qui a été, à ce jour, la seule campagne à avoir proposé une tâche médicale.

[58] Question dans une langue et la réponse attendue dans une autre.

### 6.1.1 La campagne d'évaluation EQueR

Le projet EVALDA-EQueR (Evaluation en Question-Réponse) (Ayache, 2005) a permis de réaliser, en 2004, une campagne d'évaluation des systèmes de question-réponse pour le français EQueR. Ce projet a été lancé à l'initiative du Ministère de la Recherche dans le cadre de l'action Technolangue. La campagne d'évaluation EQueR a été organisée et pilotée par ELDA (http://www.elda.org) avec pour responsable scientifique Brigitte Grau du laboratoire LIMSI. Elle a vu la participation de huit groupes : AP/HP-Paris XIII, LIMSI, LIA-iSmart, l'université de Neuchâtel, CEA-LIST/LIC2M, France Télécom R&D, Sinequa et Synapse Développement. La campagne offrait un cadre d'évaluation à des systèmes complets de question-réponse. Son objectif principal était de donner un aperçu des recherches sur les systèmes de question-réponse développés pour le français.

EQueR a proposé deux tâches de recherche automatique de réponses : une « tâche générale[59] » sur une collection hétérogène de textes, constituée principalement d'articles de journaux et des rapports d'information du Sénat, et une « tâche spécifique[60] » liée au domaine médical, sur une collection de textes de cette spécialité. Les deux corpus de questions ont été élaborés pour quatre types de questions prévus pour cette campagne, à savoir : les questions « factuelles » (*i.e.* qui attendent des entités nommées comme réponse)[61], les questions de type « définition », les questions de type « liste » (qui attendent un nombre bien précis de réponses) et enfin les questions « booléennes » (réponse de type oui/non).

Pour la tâche générale, ELDA a composé un corpus de 500 questions réparties comme suit : 407 questions factuelles, 32 questions définitoires, 31 de type liste et 30 de type booléen. Parmi ces 500 questions proposées, 100 questions étaient des reformulations de questions factuelles simples et 5 questions n'avaient pas de réponse dans le corpus d'évaluation[62]. Pour

[59] La collection du domaine général représente un volume de 1,5 Go provenant du journal Le Monde (1992-2000), Le Monde Diplomatique (1992-2000), de rapports d'information du Sénat (1996-2001), de rapports Interparlementaires d'Amitiés du Sénat (1992-2001), de lois et de rapports législatifs du Sénat (1996-2001).

[60] Le volume de la collection représente 140 Mo, constitués d'articles scientifiques médicaux et de recommandations de bonnes pratiques provenant du Sénat, de l'HAS, de la documentation française, d'agences de santé étatiques comme Santé Canada, de portails médicaux comme CISMeF, d'Orphanet, de la FNLCC et de l'université de Rouen.

[61] Les questions portaient sur des dates, des durées des distances ou des dimensions, des lieux, des personnes, des organisations, la manière ou le mode de déroulement d'évènements, des entités concrètes ou abstraites (Ayache, 2005).

[62] Dans ce cas, le système devait renvoyer la réponse « NIL ».

la tâche médicale, une collection de 200 questions dont 51 reformulations (cf. Annexe A) a été constituée par l'équipe du CISMeF. Elle est composée de :

- 81 questions factuelles simples « Quel est le traitement du chérubisme ? »,
- 70 questions définitoires « Quelle est la définition de la désinfection ? »,
- 25 questions de type liste « Citez 4 symptômes de l'AVF. »,
- 24 questions de type booléen « Un enfant peut-il être atteint de schizophrénie ? ».

Le type de la question était indiqué par un codage d'identification attribué à chaque question et donc connu des systèmes.

Pour chaque question, les systèmes participants pouvaient retourner soit des passages réponses de 250 caractères au maximum, soit des réponses courtes accompagnées d'un passage. Dans les deux cas, le document d'origine devait être spécifié en guise de justification des réponses. Le nombre de réponses à renvoyer pour chaque question était limité à cinq réponses ordonnées (20 pour les questions Liste). De plus, chaque participant avait la possibilité de soumettre jusqu'à 2 « runs » par tâche afin de tester différentes méthodes ou différents paramètres.

Chaque réponse a été jugée manuellement par 2 juges pour la tâche générale et un juge spécialiste de l'équipe CISMeF pour la tâche médicale. L'évaluation concernait à la fois les passages et les réponses courtes renvoyés par les systèmes participants. De ce fait, deux types d'évaluation ont été proposés :

- Pour les réponses courtes, une réponse est jugée correcte si la chaîne retournée contient exactement la bonne réponse et que celle-ci est justifiée par le document dont elle est extraite. Elle est jugée incorrecte si la chaîne ne correspond pas à la réponse attendue. Elle est jugée inexacte lorsque la chaîne contient la bonne réponse est trouvée dans un document la justifiant mais qu'elle n'est pas assez précise, c'est-à-dire incomplète. Enfin, une réponse courte est jugée non justifiée lorsque la chaîne retournée contient la bonne réponse mais que le document associé ne justifie pas cette réponse.
- Pour les passages, le jugement est seulement correct ou incorrect. Il est jugé incorrect s'il ne contient pas la réponse attendue ou celle-ci ne répond pas à la question.

Afin d'évaluer les réponses, deux mesures ont été adoptées. Les questions de type « factuel », « définition » et « booléen » sont évaluées par la MRR[63] (Mean Reciprocal Rank), qui correspond à la moyenne de l'inverse des rangs de la première bonne réponse. Les questions de type « liste » ont quant à elles été jugées en utilisant la précision moyenne non interpolée (Non Interpolated Average Precision, NIAP)[64].

$$MRR = \frac{1}{\#questions} \sum_{i=1}^{\#questions} \frac{1}{answer_i \ \ rank}$$

$$prec\_moy\ (q_i) = \frac{\sum_{j=1}^{j=n} I(rep_j) \cdot prec\,(j)}{R} \leq 1$$

$$I(rep_j) = \begin{cases} 1\ si\ rep_j\ est\ une\ bonne\ \ réponse \\ 0\ si\ rep_j\ est\ une\ mauvaise\ \ réponse\ \ ou\ une\ réponse\ \ déjà\ \ proposée \end{cases}$$

$$prec\,(j) = \frac{\sum_{k=1}^{j} I(rep_k)}{j} = \frac{Nombre\ \ de\ bonnes\ \ réponses\ \ différente\ s\ \ jusqu\ 'au\ rang\ \ j}{j} \leq 1$$

### 6.1.2 La campagne d'évaluation CLEF-QA

En 2000 est apparue la campagne CLEF (Cross Language Evaluation Forum) d'évaluation des systèmes de recherche d'information. La campagne se propose d'aborder la dimension du multilinguisme en recherche d'information et ce, pour les langues européennes. Ainsi, l'objectif principal de CLEF est d'évaluer des systèmes de recherche d'information crosslingues ou monolingues pour des langues européennes autres que l'anglais. Pour ce faire, elle propose un cadre d'évaluation fondé sur le modèle de TREC.

[63] Ce critère tient compte du rang de la première bonne réponse trouvée. Si la première réponse est correcte, la MRR est égale à 1 ; si la réponse correcte est en deuxième position, la MRR a pour valeur 1/2 ; si la bonne réponse est en troisième position, la valeur est 1/3…

[64] Cette mesure tient compte du rappel et de la précision mais aussi de la position des bonnes réponses dans la liste. La précision correspond au pourcentage des bonnes réponses trouvées parmi toutes les réponses trouvées, alors que le rappel représente le pourcentage de bonnes réponses présentes dans la liste parmi toutes les bonnes réponses à trouver.

La quatrième édition de l'évaluation CLEF (CLEF 2003) a vu l'introduction de la tâche question/réponse (CLEF-QA). La tâche proposée se focalise également sur la dimension multilingue. Le but est de promouvoir le développement de systèmes de question-réponse capables, à partir d'une question posée dans une langue source donnée, de retourner une réponse extraite d'une base documentaire dans une langue cible différente. Des évaluations pour des systèmes monolingues existent également pour des langues autres que l'anglais. Lors de la campagne CLEF-QA 2006, on dénombrait dix langues sources : anglais, allemand, français, espagnol, italien, portugais, bulgare, néerlandais, indonésien et roumain. Ces langues étaient également les langues cibles à l'exception de l'indonésien et du roumain. Dans cette même compétition, deux nouvelles sous-tâches ont été introduites : une sous-tâche d'évaluation des systèmes de question-réponse utilisant Wikipédia « WiQA » et une autre, « AVE : Answer Validation Exercise», concernant la justification des réponses.

Pour chaque langue, CLEF propose un corpus de 200 questions, principalement factuelles (10% sont des questions définitoires). Les questions sont d'abord élaborées dans la langue cible, puis traduites dans la langue source. Ainsi, toutes les tâches ayant la même langue cible partagent les mêmes questions. Les participants doivent trouver des réponses courtes à ces questions dans une collection fournie au préalable de documents dans la langue cible. Les réponses retournées par les systèmes participants sont ensuite jugées manuellement. Le jugement de la réponse courte retournée est fondé sur le même principe que celui présenté pour la campagne EQueR (correcte, incorrecte, inexacte, non justifiée).

Afin d'évaluer les réponses, trois mesures ont été utilisées en plus de la précision (nombre de réponses correctes sur le nombre de réponses retournées). Ces trois mesures sont : la MRR, la K1-measure et le CWS (Confident Weighted Score), qui permet de donner un poids plus important aux réponses correctes apparaissant en tête de classement.

$$K1(sys) = \frac{\sum_{R \in \text{réponses}(sys)} score(R) \cdot eval(R)}{\#\ questions}$$

$$K1(sys) \in IR \ \wedge \ K1(sys) \in [-1,1]$$

Où : score(R) est le score de confiance attribué par le système à la réponse R et eval(R) dépend du jugement manuel de la réponse R.

$$eval(R) = \begin{cases} 1 \ si\ R\ est\ jugée\ correcte \\ -1 \ dans\ les\ autres\ cas \end{cases}$$

## 6.2 Évaluation du système Œdipe

Dans cette section, nous exposons l'évaluation obtenue par le système de question-réponse Œdipe dans les campagnes d'évaluation EQueR et CLEF-QA (2005 et 2006).

### 6.2.1 Le système Œdipe dans EQueR

Dans le cadre de la campagne d'évaluation EQueR, le laboratoire LIC2M a participé à la fois à la tâche générale et à la tâche médicale, en utilisant dans les deux cas exactement le même système et en traitant tous les types de questions. Cependant, seules des réponses prenant la forme de passages de 250 caractères ont été renvoyées. Le Tableau 6.1 synthétise les résultats du système Œdipe. Pour la tâche générale, la MRR globale se situe à 0,7 pour le meilleur système et aux alentours de 0,3 pour la majorité des systèmes, à comparer à 0,5 et 0,1 pour la tâche médicale. Comparativement, les résultats d'Œdipe sont donc faibles et même très faibles pour le domaine médical. Ceci peut s'expliquer bien sûr par le fait que la version d'Œdipe pour EQueR était minimaliste[65]. En particulier, aucun traitement spécifique des questions définitoires n'était présent, traitement que nous avons ajouté par la suite. Mais le Tableau 6.1 montre également un nombre anormalement élevé de questions jugées sans réponse dans le corpus d'évaluation par le système (ce qui explique d'ailleurs le relatif bon score obtenu pour les questions booléennes). Une analyse *a posteriori* a montré la présence de deux bogues au niveau d'Œdipe ayant eu comme conséquence d'entraîner un traitement uniforme des questions, indépendamment de leur type. La deuxième ligne du Tableau 6.1 donne les résultats d'Œdipe pour les questions factuelles et les questions de définition après correction de ces deux bogues.

En dehors de ces problèmes de développement, afin d'éclaircir l'origine des insuffisances d'Œdipe, nous avons mené une analyse manuelle concernant les performances du typage des questions, analyse dont les résultats sont reportés dans le Tableau 6.2. Celle-ci laisse apparaître que le typage des questions effectué par le système Œdipe se révèle assez efficace[66]. Si l'on prend en compte à la fois les cas dans lesquels Œdipe trouve le type d'entité

[65] Cette version d'Oedipe est la version initiale développée spécifiquement pour EQueR et ne représentait qu'à peine un mois-homme de développement.

[66] Ce que confirment d'ailleurs indirectement les résultats corrigés d'Œdipe puisqu'en tenant compte de ce typage, on obtient un quasi-doublement des performances.

nommée attendue comme réponse (première ligne du tableau) et les cas dans lesquels il considère que la réponse n'est pas une entité nommée (seconde ligne du tableau), on constate qu'il se trompe dans 28,6% des cas pour les 469 questions analysées (hors questions booléennes) du domaine général et dans 10,8% des cas pour les 176 questions analysées du domaine médical.

| **Tâche** | **Passages corrects / questions** | **MRR Sauf listes** | **MRR Sauf listes et polaires** | **MRR polaires** | **Précision moyenne Listes** | **Détection d'absence de réponse (Nb)** |
|---|---|---|---|---|---|---|
| Générale (officiel) | 113 / 464 | 0,18 | 0,17 | 0,38 | 0,13 | 236<br>Précision : 0<br>Rappel : 0,4 |
| Générale (après corrections) | 196 / 440 | nc | 0,31 | nc | nc | nc |
| Médicale | 7 / 175 | 0,02 | 0,02 | 0 | 0 | n/a |

Tableau 6.1 Résultats du système Œdipe pour l'évaluation EQueR

| **Typage** | **Jugement manuel** | **Général** | **Médical** |
|---|---|---|---|
| Type identifié par Œdipe | Correct | 215 / 254 (45,8%) | 17 / 29 (9,7%) |
| | Incorrect | 39 / 254 (8,3%) | 12 / 29 (6,8%) |
| Type non identifié par Œdipe = réponse non factuelle | Correct | 120 / 215 (25,6%) | 140 / 147 (4%) |
| | Incorrect | 95 / 215 (20,3%) | 7 / 147 (79,5%) |

Tableau 6.2 Résultats de l'analyse manuelle du typage des questions par Œdipe

Plus globalement, l'étude des résultats de la campagne EQueR montre que les résultats obtenus pour la tâche générale sont nettement supérieurs à la tâche médicale. En effet, selon (Ayache, 2005), le meilleur système de question-réponse pour la tâche médicale n'obtient que 40% de bonnes réponses contre 67% pour la tâche générale. Le classement des systèmes de question-réponse pour la tâche spécialisée lors de la campagne EQueR/EVALDA 2004 (Ayache, 2005) est dans l'ordre : pour les passages (cf. Tableau 6.3), le système de Synapse Développement (groupe 4), celui de l'Université de Neuchâtel (groupe 2), ex-æquo, les systèmes de l'AP/HP- Paris XIII (groupe 3) et de France Télécom (groupe 1) et enfin le système Œdipe (groupe 5) ; pour les réponses courtes (cf. Tableau 6.4), le système de Synapse

Développement, ex-æquo, les systèmes de l'AP/HP-Paris XIII et de l'Université de Neuchâtel et le système France Télécom.

Le premier constat à tirer de cette évaluation est que les techniques adoptées pour la tâche générale s'avèrent moins performantes dans un domaine spécialisé comme le domaine médical. Parmi tous les systèmes participants pour la tâche médicale, seul le système construit par l'équipe de l'AP/HP- Paris XIII proposait une approche fondée sur des connaissances sémantiques spécifiques au domaine. Cependant, le peu de temps consacré au développement du système ne permet pas de porter un réel jugement sur sa compétence. Leurs concepteurs confirment néanmoins que le recours à des connaissances sémantiques sur le domaine permet de rendre les systèmes plus efficaces. On peut aussi supposer que les résultats plus faibles obtenus pour la tâche médicale sont dus à la nature des questions factuelles proposées. Ces dernières se sont révélées plus difficiles et plus complexes que celles de la tâche générale. En effet, certaines questions de la tâche médicale, classées factuelles, n'attendent pas forcément une entité nommée comme réponse et dans d'autres cas, il est même difficile de déterminer le type de la réponse attendue, comme par exemple pour des questions telles que :

« *MF1 Pour quelles raisons une consultation diététique est-elle préconisée ?* »

« MF19 *Comment le poids corporel est-il déterminé ?* »

| **Participants** | **Nb** | **PC (Nb)** | **PI (Nb)** | **PC (%)** | **MRR1**[67] | **MRR2**[68] | **MRR3**[69] | **MRR4**[70] | **MRR5**[71] | **NIAP (ML)** |
|---|---|---|---|---|---|---|---|---|---|---|
| Groupe 4 | 175 | 110 | 65 | 62,85 | 0,49 | 0,51 | 0,42 | 0,62 | 0,37 | 0,02 |
| Groupe 4 | 175 | 102 | 73 | 58,28 | 0,47 | 0,48 | 0,41 | 0,57 | 0,41 | 0,02 |
| Groupe 2 | 175 | 26 | 149 | 14,85 | 0,11 | 0,13 | 0,19 | 0,05 | 0,04 | 0,02 |
| Groupe 2 | 175 | 27 | 148 | 15,42 | 0,13 | 0,13 | 0,23 | 0,02 | 0,08 | 0,02 |
| Groupe 1 | 166 | 23 | 143 | 13,14 | 0,09 | 0,09 | 0,11 | 0,07 | 0,04 | 0 |
| Groupe 3 | 112 | 16 | 96 | 9,14 | 0,09 | 0,05 | 0,02 | 0,08 | 0,33 | 0,01 |
| **Groupe 5** | **175** | **7** | **168** | **4** | **0,02** | **0,02** | **0,04** | **0** | **0** | **0** |

Tableau 6.3 Résultats de l'évaluation des runs pour les passages (tâche médicale)[72]

[67] MRR sur (MF, MD, MRF, MRD, MB)
[68] MRR sur (MF, MD, MRF, MRD)
[69] MRR sur (MF, MRF)
[70] MRR sur (MD, MRD)
[71] MRR sur MB
[72] PC pour passages corrects et PI pour passages incorrects

| Participants | Nb | RC[73] (Nb) | RC (%) | MRR1 | MRR2 | MRR3 | MRR4 | MRR5 | NIAP (ML) |
|---|---|---|---|---|---|---|---|---|---|
| Groupe 4 | 174 | 71 | 40,57 | 0,31 | 0,3 | 0,3 | 0,31 | 0,37 | 0 |
| Groupe 4 | 175 | 59 | 33,71 | 0,27 | 0,25 | 0,25 | 0,24 | 0,41 | 0,01 |
| Groupe 2 | 175 | 8 | 4,57 | 0,03 | 0,03 | 0,05 | 0 | 0,04 | 0,01 |
| Groupe 2 | 175 | 13 | 7,42 | 0,06 | 0,06 | 0,11 | 0 | 0,08 | 0 |
| Groupe 3 | 35 | 12 | 6,85 | 0,06 | 0,02 | 0 | 0,05 | 0,33 | 0 |
| Groupe 1 | 117 | 6 | 3,42 | 0,03 | 0,02 | 0,03 | 0,01 | 0,34 | 0 |

Tableau 6.4 Résultats de l'évaluation des runs pour les réponses courtes (tâche médicale)

### 6.2.2 Le système Œdipe dans CLEF-QA

Le système Œdipe a participé à CLEF-QA 2005 et 2006. La version 2005 était globalement identique à la version d'Œdipe ayant participé à l'évaluation EQueR avec néanmoins les modifications suivantes :

- la correction des deux bogues évoqués ci-dessus permettant en particulier de prendre en compte de façon effective le typage des questions ;
- l'extraction de réponses exactes au lieu de passages réponses. Pour les questions attendant une entité nommée en tant que réponse, la réponse extraite était l'entité nommée correspondant au type attendu autour de laquelle le passage de plus haut score était centré. Pour les questions définitoires, la réponse était extraite grâce à un ensemble très restreint de patrons écrits manuellement.

La version 2006 reprenait quant à elle la version 2005 en substituant aux patrons écrits manuellement des patrons appris automatiquement à partir de textes, comme nous l'avons vu à la Section 4.5.

[73] Réponses correctes.

### 6.2.2.1 Les résultats du système Œdipe dans CLEF-QA 2005

Pour l'évaluation CLEF-QA 2005, un seul run du système de question-réponse Œdipe a été soumis. Pour les 200 questions proposées, le système a retourné 28 réponses correctes ; toutes étaient des réponses à des questions factuelles parmi lesquelles 6 comprenaient un contexte temporel. Parmi les questions proposées, 20 questions n'avaient pas de réponses dans le corpus. Œdipe a détecté 3 questions sans réponses possibles dont une seule était correcte. La deuxième colonne du Tableau 6.7 récapitule les meilleurs résultats obtenus par les sept participants à la tâche monolingue français-français (Vallin et al., 2006). Le Tableau 6.5 montre que les résultats du système Œdipe (système 7) sont insuffisants mais restent tout de même proches des résultats de la moitié des participants.

| Systèmes | Réponses correctes | Score avec la difficulté de la question |
|---|---|---|
| 1 | 128 | 67,5 |
| 2 | 70 | 30,75 |
| 3 | 46 | 17,75 |
| 4 | 35 | 15,25 |
| 5 | 33 | 17,75 |
| 6 | 29 | 15 |
| **7** | **28** | **16,75** |

Tableau 6.5 Les résultats de CLEF-QA 2005 pour la tâche monolingue Français

Pour tenir compte du fait que toutes les questions n'ont pas le même niveau de difficulté, nous avons calculé un score tenant spécifiquement compte de cet aspect à partir des données fournies par les organisateurs de la campagne (cf. troisième colonne du Tableau 6.5). Le niveau de difficulté d'une question est évalué par le nombre de systèmes n'ayant pas retourné une bonne réponse pour la question concernée. Nous avons calculé la moyenne (dénotée M*diff*) et l'écart type (dénoté SD*diff*) des valeurs de difficulté pour les 200 questions, puis nous déterminons le score d'une bonne réponse à une question comme suit :

$$\begin{cases} \text{Score} = 0{,}25 & \text{si} \quad \text{difficulté} \leq M_{diff} - SD_{diff} \\ \text{Score} = 0{,}5 & \text{si} \quad \text{difficulté} \leq M_{diff} \\ \text{Score} = 0{,}75 & \text{si} \quad \text{difficulté} \leq M_{diff} + SD_{diff} \\ \text{Score} = 1 & \text{si} \quad \text{difficulté} \leq M_{diff} + SD_{diff} \end{cases}$$

Comme tout autre système de question-réponse, les erreurs du système Œdipe peuvent provenir d'un ou de plusieurs modules qui composent sa chaîne de traitement. De ce fait, nous avons mené une étude manuelle afin d'identifier les lacunes du système. Les résultats de cette étude ont été obtenus en prenant comme référence les runs évalués de tous les participants à CLEF-QA 2005 travaillant pour le français comme langue cible. Cette référence s'avère tout de même incomplète puisque d'une part, toutes les réponses n'ont pas été trouvées, et d'autre part, il n'y a aucune garantie que toutes les occurrences d'une réponse aient été trouvées dans le corpus. Cependant, la pratique montre que c'est une approche fiable pour calculer automatiquement le score minimal d'un système de question-réponse sur ce corpus.

La première source de réponses non trouvées concerne la récupération des documents par le moteur de recherche. Dans notre cas, nous avons constaté que le moteur de recherche du LIC2M a renvoyé au moins un document avec une réponse pour 132 questions parmi les 200 questions proposées, ce qui représente 66% des questions. Ainsi, le système Œdipe a trouvé 21,2% des réponses possibles après l'étape de recherche documentaire. Plus globalement, le moteur de recherche du LIC2M a renvoyé 262 des 383 documents contenant une réponse trouvée par au moins un participant, c'est-à-dire 68,4%.

Une autre partie importante d'un système de question-réponse est le module d'analyse des questions puisque celui-ci détermine généralement la stratégie à adopter pour rechercher une réponse à une question. Le Tableau 6.6 résume les résultats du module d'analyse des questions du système Œdipe sur le corpus des 200 questions de la campagne CLEF-QA 2005. La première chose à noter est que le taux d'erreur de la classification (10,5%) est faible. D'ailleurs, toutes les questions définitoires ont été bien typées, ce qui signifie que le module des questions n'est pas responsable des mauvais résultats d'Œdipe pour cette catégorie de questions.

L'influence des autres modules d'Œdipe sur ses résultats globaux est illustrée par le Tableau 6.9, qui donne le nombre de réponses et de passages corrects trouvés dans les « *R* » premières réponses aux questions CLEF-QA 2005. Plus particulièrement, le Tableau 6.7 montre que 44,7% des réponses qui pouvaient être trouvées après l'étape de recherche documentaire sont présentes dans les 10 premiers passages réponses extraits par Œdipe. Ce pourcentage est réduit à 37,1% pour les 5 premiers passages réponses. Pour les réponses courtes, il est égal à

28% dans le premier cas et à 25,8% dans le second. Cependant, la différence la plus évidente entre les passages réponses et les réponses courtes concerne les questions définitoires : bien que des passages réponses aient été trouvés pour certaines questions définitoires, aucune bonne réponse courte n'a pu être extraite pour eux. Cela signifie que les heuristiques utilisées pour extraire des réponses courtes pour les questions définitoires étaient inefficaces, ce qui n'est pas surprenant compte tenu de la rapidité de leur développement et de leur manque de test. C'est d'ailleurs sur ce point que les efforts ont porté lors de l'évaluation qui a suivi, CLEF-QA 2006, en remplaçant ces heuristiques par des patrons d'extraction de réponse appris de manière supervisée.

| Type de question | Nb questions | Nb types corrects | Nb types incorrects |
|---|---|---|---|
| Définition (D) | 50 | 50 | 0 |
| Factuel (F) | 120 | 106 | 14 |
| Factuel temporel (T) | 30 | 23 | 7 |
| **Total** | **200** | **179** | **21** |

Tableau 6.6 Résultats du module d'analyse des questions dans le cadre de CLEF-QA 2005

| | Réponse juste | | | | | Passage juste | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Nb réponses/question | MRR | Nb réponses correctes | | | | MRR | Nb Passages corrects | | | |
| | | Total | T | D | F | | Total | T | D | F |
| 1 | 0,140 | 28 | 6 | 0 | 22 | 0,170 | 34 | 7 | 2 | 25 |
| 2 | 0,147 | 31 | 7 | 0 | 24 | 0,182 | 39 | 8 | 2 | 29 |
| 3 | 0,151 | 33 | 7 | 0 | 26 | 0,193 | 45 | 8 | 5 | 32 |
| 4 | 0,152 | 34 | 8 | 0 | 26 | 0,194 | 46 | 9 | 5 | 32 |
| 5 | 0,152 | 34 | 8 | 0 | 26 | 0,197 | 49 | 9 | 7 | 33 |
| 10 | 0,154 | 37 | 8 | 0 | 29 | 0,203 | 59 | 9 | 11 | 39 |

Tableau 6.7 Résultats détaillés d'Œdipe pour la tâche monolingue français de CLEF-QA 2005

#### 6.2.2.2 Les résultats du système Œdipe dans CLEF-QA 2006

Comme pour CLEF-QA 2005, un seul run du système Œdipe a été soumis pour l'évaluation CLEF-QA 2006. Pour les 200 questions proposées, Œdipe a retourné 30 réponses correctes, 3

réponses non justifiées[74] et 6 réponses inexactes, ce qui donne une moyenne globale de 16% de réponses. De plus, la détection des réponses non trouvées par Œdipe était exacte pour seulement une question parmi les trois repérées.

Les résultats obtenus par Œdipe à l'évaluation CLEF-QA 2006 sont comparables, avec une légère amélioration, à ceux obtenus lors de la campagne CLEF-QA 2005, dont la moyenne globale était égale à 0,14 avec 28 réponses correctes. Cependant, le Tableau 6.8 montre que les distributions des réponses correctes sont différentes pour les deux évaluations. L'utilisation des patrons « définitions » apporte une amélioration très significative pour les questions définitoires. En revanche, les résultats concernant les questions factuelles (simples et temporelles), qui ont été traitées par la même version du système Œdipe que celle utilisée lors de CLEF-QA 2005, diminuent de manière significative. L'amélioration des questions définitoires était prévue mais il n'y a aucune explication évidente à la diminution des résultats pour les questions factuelles, en dehors du fait que leur formulation était peut-être plus difficile à traiter que celle des questions factuelles proposées lors de CLEF-QA 2005[75].

| | **Factuelle (F + T)** | | **Définition (D)** | |
|---|---|---|---|---|
| | **Nb réponses correctes** | **% réponses correctes** | **Nb réponses correctes** | **% réponses correctes** |
| CLEF-QA 2005 | 28 | 18,7 | 0 | 0 |
| CLEF-QA 2006 | 15 | 10,3 | 15 | 36,6 |

Tableau 6.8 Comparaison des distributions des réponses correctes du système Œdipe lors de CLEF-QA 2005 et CLEF-QA 2006

Par ailleurs, le Tableau 6.9 montre que le module d'analyse des questions n'est pas responsable de la baisse des résultats du système Œdipe pour les questions factuelles puisque son exactitude pour les questions de CLEF-QA 2006 est plus élevée que pour les questions de CLEF-QA 2005. Il est intéressant de noter que le focus a été correctement identifié pour toutes les questions définitoires.

[74] Une réponse correcte est extraite mais le document associé ne justifie pas cette réponse.
[75] En outre, contrairement à CLEF-QA 2006, dans CLEF-QA 2005 le type de la question était fourni avec la question.

| Type de question | Nb questions | Nb types incorrects | Exactitude (2006) | Exactitude (2005) |
|---|---|---|---|---|
| Factuel (F + T) | 146 | 9 | 93,8 | 86 |
| Définition (D) | 41 | 4 | 90,2 | 100 |

Tableau 6.9 Résultats du module d'analyse des questions du système Œdipe pour CLEF-QA 2006 et la comparaison avec CLEF-QA 2005

## 6.3 Évaluation du système Esculape

Dans cette section, nous présentons l'évaluation du système de question-réponse Esculape sur le corpus de questions de la tâche médicale EQueR, la seule évaluation qui offrait la possibilité d'étudier les différentes solutions proposées en question/réponse médical. Cette évaluation a porté plus précisément sur l'analyse de deux étapes importantes du système, à savoir la classification des questions du domaine médical et l'extraction des réponses courtes attendues. Le Tableau 6.10 montre que la méthode d'analyse des questions du système Esculape est plus performante que celle du système Œdipe pour la tâche médicale, puisque seules 10% des 200 questions ont été mal typées (contre 22% pour le système Œdipe). L'analyse des résultats du typage réalisé par Esculape a montré que la majorité des erreurs observées concernaient les questions factuelles dont les pronoms interrogatifs font référence à des notions différentes de celles attendues dans un domaine plus général. Par exemple, dans la question « Quand doit-on procéder au dosage de la créatininémie ? », le pronom « quand » représente une condition clinique et non une temporalité.

| Type de question | Nb questions | Nb type corrects | % types corrects |
|---|---|---|---|
| Factuel (F) | 81 | 66 | 81,5 |
| Définition (D) | 70 | 69 | 98,5 |
| Liste (L) | 25 | 21 | 84 |
| Booléen (B) | 24 | 24 | 100 |
| Total | 200 | 180 | 90 |

Tableau 6.10 Résultats de l'analyse des questions par le système Esculape pour la tâche médicale EQueR

La deuxième partie de l'évaluation consiste à étudier la capacité du système Esculape à extraire les réponses aux questions médicales. Pour cela, nous avons tout d'abord utilisé les passages réponses corrects trouvés par le système Œdipe lors de la campagne EQueR. Cependant, le nombre réduit de passages retournés par Œdipe (7/175) ne permet pas d'évaluer convenablement la compétence d'Esculape. En effet, en utilisant ces passages corrects d'Œdipe, le système Esculape n'a pu extraire qu'une bonne réponse pour les 7 questions concernées. Ainsi, afin de juger Esculape sur une collection plus large de passages réponses, nous nous sommes appuyés sur le fichier de jugement des réponses construit à partir des résultats renvoyés par tous les systèmes participants à la tâche médicale EQueR. L'idée était donc d'exploiter tous les passages jugés corrects par un spécialiste de l'équipe CISMEF de Rouen.

Le Tableau 6.11 montre que les résultats obtenus par le système Esculape se rapprochent des meilleurs systèmes mais reste tout de même très loin du premier système (71 réponses trouvées) (cf. Tableau 6.4). On note aussi le pourcentage élevé des réponses correctes trouvées pour les questions définitoires par rapport aux questions factuelles. Ce résultat s'explique assez directement par la nature des questions factuelles posées dans le cadre de la compétition, questions qui ne se limitaient pas au cadre d'une consultation médicale[76] et faisaient intervenir un ensemble d'entités et de relations plus large que celui que nous avons pris en compte dans notre travail. On notera ainsi que l'utilisation des patrons d'extraction de réponses s'est avérée plus performante pour les questions définitoires, dont le traitement repose sur des patrons « généraux », que pour les questions factuelles, pour lesquelles les patrons dépendent des relations sous-jacentes aux questions. Il est à préciser que sur les 60 questions factuelles, le système Esculape ne pouvait traiter que 38 questions.

| | **Total** | **Nb Type de question** | | | | | |
|---|---|---|---|---|---|---|---|
| | | **Factuel (F)** | **%** | **Définition (D)** | **%** | **Liste (L)** | **%** |
| Nb questions | 130 | 60 | 46,15 | 58 | 44,60 | 12 | 9,25 |
| Nb réponses correctes | 24 | 5 | 20,80 | 18 | 75 | 1 | 4,20 |

Tableau 6.11 Résultats du système Esculape sur les passages des participants EQueR

[76] Le système Esculape a été développé pour répondre à des questions sur les bonnes pratiques médicales.

Enfin, on remarque que les questions de la tâche médicale étaient globalement plus difficiles à traiter que celles de la tâche générale. En effet, les questions médicales proposées étaient de type variés et attendaient parfois comme réponses des explications, des conséquences ou des manières, telles que :

« Quels sont les éléments qui distinguent la migraine de l'adulte de celle de l'enfant ? »
« Quelle est la conséquence de la corticothérapie sur l'os ? »
« Comment organiser le suivi d'un patient atteint d'insuffisance rénale ? »
« Comment le degré d'immobilité d'un patient peut-il être évalué ? »

## 6.4 Synthèse

Nous avons présenté dans ce chapitre l'évaluation du système Œdipe dans les différentes campagnes d'évaluation des systèmes de question-réponse auxquelles il a participé, soit les campagnes EQueR, CLEF-QA 2005 et CLEF-QA 2006. Pour la campagne EQueR, Œdipe a été évalué pour les deux tâches proposées : la tâche générale et la tâche médicale. Toutefois, le système n'a été jugé que pour les passages corrects puisqu'il n'avait pas la compétence nécessaire pour extraire les réponses courtes. Les résultats obtenus par le système lors de cette campagne sont faibles, et même très faibles pour la tâche médicale, mais ne sont en réalité pas très représentatifs du fait de ses bogues. Une version du système Œdipe légèrement améliorée (pour extraire des réponses courtes) a participé à la tâche monolingue « français » de l'évaluation CLEF-QA 2005. L'analyse de ses résultats montre qu'un tel système « minimaliste » est capable de répondre à au moins 20% des questions factuelles. En revanche, son absence de traitement spécifique des questions définitoires le rend globalement peu performant. Pour faire face à ce problème, la version d'Œdipe ayant participé à la compétition CLEF-QA 2006 intégrait l'utilisation de patrons lexico-syntaxiques appris automatiquement pour extraire des réponses courtes à des questions définitoires. Cette démarche a donné des résultats satisfaisants mais peut encore être améliorée. Une première amélioration concerne l'intégration de l'analyse syntaxique de LIMA dans Œdipe, ce qui permettrait d'extraire des groupes nominaux et de rendre les patrons linguistiques plus généraux. Une autre amélioration consiste à étendre l'application des patrons d'extraction de

réponses au traitement des questions factuelles. Les résultats d'Œdipe dans CLEF-QA 2006 pour ce type de questions ont en effet sensiblement diminué par rapport à l'évaluation CLEF-QA 2005.

Enfin, une étude comparative a été réalisée entre les systèmes Esculape et Œdipe. Cette analyse, sur les questions de la tâche médicale EQueR, montre que le système Esculape, dédié au traitement de questions médicales, obtient de meilleurs résultats que le système Œdipe. Cette amélioration s'explique par l'usage de connaissances spécifiques au domaine médical, usage qui permet au système d'effectuer un meilleur typage des questions et de repérer les concepts médicaux dans les passages susceptibles de contenir une réponse. Malgré leur niveau encore modeste, les résultats du système Esculape apparaissent comme encourageants compte tenu de la complexité des questions médicales proposées en comparaison avec les questions factuelles généralement posées dans d'autres campagnes d'évaluation. De plus, ces questions ne se restreignaient pas au contexte applicatif prévu pour la version testée d'Esculape.

# Conclusion et perspectives

# Conclusion

Dans ce travail, nous avons abordé la problématique de l'accès à l'information précise et plus spécialement à la connaissance médicale. Nous nous sommes intéressé plus particulièrement aux systèmes de question-réponse, qui visent à retourner une réponse précise à un besoin d'information exprimé en langage naturel. Cette étude nous a amené au développement d'un système de question-réponse dédié au domaine médical, « Esculape ». Le système Esculape doit permettre de répondre à des questions sur les bonnes pratiques médicales en utilisant des stratégies de recherche adaptées au domaine de la médecine. L'objectif est de donner la possibilité aux professionnels de la santé de rechercher une réponse à une question dans une base de connaissances médicales lors d'une consultation.

Pour ce faire, nous avons tout d'abord étudié la problématique des systèmes de question-réponse. En premier lieu, nous avons présenté l'architecture d'un tel type de systèmes ainsi que les différents modules intervenant dans la chaîne de traitement. Puis, nous avons distingué les différentes approches adoptées pour la classification des questions posées et l'extraction des réponses attendues par la présentation de quelques systèmes existants. Cette analyse nous a permis de constater l'importance d'une base de connaissances structurée pour le bon fonctionnement d'un système de question-réponse dans un domaine particulier, notamment pour le domaine médical. Elle nous a également apporté une certaine vision globale de notre problème et nous a permis de définir le type d'approche à employer qui nous semble la plus adaptée à notre problématique.

En réalisant un tour d'horizon sur les ressources sémantiques existantes pour le domaine médical, nous avons constaté le manque de relations spécialisées de nature syntagmatique dans ces différentes sources de données, relations telles que : « Une maladie X peut être soignée par le médicament Y ». En accord avec le cadre applicatif que nous nous sommes fixés, nous avons ensuite proposé une ontologie médicale composée essentiellement de concepts médicaux pouvant intervenir dans une consultation de médecine générale. Dans le cadre de ce travail, nous avons choisi d'étudier cinq entités médicales - Maladie, Médicament, Traitement, Examen et Symptôme - et quatre relations sémantiques de nature syntagmatique : la relation « Traite » entre Maladie et Traitement, « Soigne » entre Maladie et Médicament, « Détecte » entre Maladie et Examen et enfin la relation « Signe » entre Maladie et Symptôme. Pour identifier les entités médicales, le système dispose d'une base de connaissances médicales appropriée construite à partir de ressources médicales constituées à

partir de sources ouvertes. Le peuplement de l'ontologie est réalisé en s'appuyant principalement sur des patrons lexico-syntaxiques appris automatiquement à partir de textes annotés et permettant d'extraire de nouveaux couples de termes correspondant à la relation recherchée. Les avantages de cette approche sont doubles : les patrons construits permettant de peupler notre ontologie sont en effet les mêmes aidant le système de question-réponse à repérer des réponses candidates. L'évaluation concernant l'extraction de nouvelles relations sémantiques a donné des résultats satisfaisants, preuve de la fiabilité des relations extraites.

L'approche fondée sur l'utilisation des patrons lexico-syntaxiques pour extraire des réponses a été employée par le système de question-réponse Œdipe lors de l'évaluation CLEF-QA 2006. Le système Œdipe a été développé initialement pour répondre, par des extraits de textes, à des questions en domaine ouvert. Il se fonde principalement sur l'analyseur linguistique LIMA. Les résultats obtenus par Œdipe dans les différentes campagnes d'évaluation des systèmes de question-réponse (EQueR et CLEF-QA) sont globalement modestes en raison de sa conception minimaliste. Cependant, l'évaluation CLEF-QA 2006 a montré une nette amélioration du système pour le traitement des questions définitoires. Cette amélioration est plus particulièrement due à l'application de patrons linguistiques construits spécifiquement pour repérer des réponses à ce type de questions. Par ailleurs, les résultats obtenus par Œdipe pour la tâche médicale EQueR ont confirmé que le système ne disposait d'aucune compétence lui permettant de trouver des réponses à des questions médicales. Cette difficulté nous a amené à développer un système de question-réponse capable de répondre à des questions médicales, le système Esculape. À la différence du système Œdipe, Esculape permet de retourner des réponses courtes à des questions concernant le domaine médical. Ces deux systèmes partagent néanmoins une même architecture.

Afin d'offrir au système Esculape la compétence nécessaire pour répondre à des questions portant sur le domaine médical, nous avons défini dans un premier temps une méthode d'analyse des questions permettant d'une part, de catégoriser les questions médicales et d'autre part, d'en extraire les éléments importants pour faciliter la recherche des réponses attendues. Cette analyse se fonde sur des règles de typage écrites manuellement afin de caractériser le type de la réponse attendue et sur la reconnaissance des entités médicales présentes dans la question. Elle permet à Esculape de déterminer la relation cible, c'est-à-dire la relation sur laquelle porte la question parmi les relations sémantiques traitées. En ce qui

concerne l'extraction des réponses, nous avons proposé une stratégie de recherche fondée sur l'exploitation en séquence d'une base de connaissances acquise *a priori*, puis de patrons d'extraction de relations. Le processus que nous avons défini recherche donc, dans un premier temps la réponse souhaitée dans une base de connaissances construite automatiquement à partir de textes, base structurée par une ontologie médicale. Dans le cas où aucune réponse n'a pu être trouvée dans cette base, un ensemble de patrons lexico-syntaxiques de la relation cible sont appliqués pour extraire la réponse à partir de passages extraits du corpus de textes considéré.

L'implémentation du système Esculape nous a permis d'effectuer une évaluation des différentes méthodes adoptées pour un ensemble de questions que nous avons constitué afin d'illustrer de façon significative l'intérêt des approches que nous avons choisies et mises en œuvre. Les résultats obtenus sont encourageants et illustrent la capacité du système Esculape à trouver des réponses pour des questions médicales. Cependant, les patrons lexico-syntaxiques utilisés se sont révélés insuffisants pour couvrir toutes les formes par lesquelles les réponses apparaissent dans le corpus considéré.

Enfin, dans le dernier chapitre du manuscrit, nous avons exposé les différentes évaluations auxquelles le système Œdipe a participé, à savoir la campagne EQueR pour les deux tâches proposées (générale et médicale) et les campagnes CLEF-QA (2005 et 2006). Les résultats obtenus par Œdipe montrent la difficulté d'un système classique à trouver des réponses aux questions portant sur un domaine de spécialité. Nous avons également présenté l'évaluation du système Esculape sur le corpus de questions proposé pour la tâche médicale EQueR. Les résultats obtenus par Esculape sont meilleurs que ceux obtenus par Œdipe mais doivent encore être améliorés. Cela s'explique par la nature des questions médicales d'EQueR, plus difficiles que les questions en domaine ouvert, mais aussi surtout par le fait qu'Esculape, dans sa forme actuelle, ne couvre qu'un sous-ensemble du domaine médical en termes de ressources. Toutefois, l'évaluation a permis d'étudier les potentialités du système de question-réponse Esculape dans le traitement des questions médicales afin d'apporter d'éventuelles améliorations pour le rendre plus performant par la suite.

Globalement, cette étude a permis d'exposer concrètement la problématique des systèmes de question-réponse dans le domaine médical, pouvant en cela contribuer *a posteriori* à trouver les solutions adéquates aux questions auxquelles est confronté ce domaine de recherche.

Après avoir rappelé les grandes lignes de notre travail, nous allons maintenant énoncer quelques perspectives. Ces perspectives concernent en premier lieu les axes importants de notre approche.

Comme nous avons pu le voir, pour guider la recherche d'une réponse, il est primordial de bien analyser la question posée en déterminant les éléments importants de la question à transmettre au module d'extraction de réponses. L'évaluation de la méthode d'analyse des questions a montré quelques faiblesses au niveau du typage des questions factuelles, notamment pour la détermination du focus de la question. Cette limite est principalement due à la structure des règles de typage pour ce type de questions. Il nous semble donc nécessaire de revoir l'écriture de ces règles de typage pour les rendre plus efficaces. Dans ce même esprit d'amélioration des performances de l'analyse des questions, nous envisageons d'intégrer un niveau plus élaboré de traitement linguistique, jusque là limité à la lemmatisation des mots, en prenant en compte les résultats d'une analyse syntaxique des questions. Cette analyse permettra d'abord de reconnaître les groupes nominaux dans les questions et les passages candidats et ainsi de construire des patrons plus généraux. Au-delà, l'intégration de relations de dépendance syntaxique dans les patrons, à la manière de celle réalisée dans (Snow et al., 2004), est aussi un moyen d'apprendre des patrons d'extraction moins dépendants de la structure linéaire des phrases, et donc plus généraux.

Une autre partie importante des extensions envisagées de notre travail concerne l'extraction des réponses. Le processus de recherche proposé s'appuie sur une ontologie médicale et sur l'utilisation des patrons d'extraction de réponses. Cependant, l'ontologie est actuellement pauvre en entités médicales. Pour cela, nous souhaiterions à la fois compléter notre base de connaissances médicales, en utilisant d'autres ressources sémantiques existantes du domaine médical, comme le thésaurus MeSH, et élargir le nombre de règles de reconnaissance des entités médicales. L'utilisation des connaissances externes s'avère très importante pour étendre les listes d'entités et introduire des relations paradigmatiques (synonymes, hyperonymes), les relations plus syntagmatiques (maladie-traitement, …) étant à acquérir par des patrons lexico-syntaxiques. Par ailleurs, le peuplement de notre ontologie repose sur l'application des patrons linguistiques, construits automatiquement, pour extraire de nouvelles relations sémantiques. Toutefois, ces patrons ne couvrent pas toutes les formes par lesquelles les relations se manifestent dans les textes. Pour améliorer à la fois la couverture des patrons

linguistiques et l'identification des entités médicales, nous envisageons d'adopter une démarche itérative classiquement utilisée dans un tel cas : au lieu de restreindre l'usage des patrons linguistiques à la seule validation des relations extraites, il est aussi possible de les utiliser pour extraire de nouvelles entités en ne fixant qu'une seule des entités d'une relation. Ces entités viennent à leur tour enrichir la reconnaissance des entités médicales et peuvent ainsi servir à acquérir de nouveaux patrons linguistiques.

Dans la même perspective d'améliorer la couverture des patrons lexico-syntaxiques, nous envisageons plusieurs extensions. Une première extension consiste à introduire un niveau supplémentaire, plus sémantique, pour exploiter des synonymes ou des hyperonymes. Une deuxième extension est d'étudier les possibilités de généralisation des patrons en leur appliquant l'algorithme de généralisation. Dans ce cadre, un accent particulier devra être mis sur la définition d'un critère d'arrêt en évaluant la précision de ces patrons généralisés pour l'extraction de nouvelles relations sémantiques. Enfin, une dernière extension envisagée concerne la transformation de la base de patrons en base d'exemples (cf. Memory-Based Learning) dans laquelle des patrons de différents niveaux (y compris restant au niveau d'exemples) pourraient apparaître avec une même représentation. L'utilisation de la distance d'édition pour sélectionner les patrons va dans ce sens.

Un autre point envisagé concerne l'extension de la couverture des relations de notre ontologie médicale, c'est-à-dire l'application de la méthode des patrons lexico-syntaxiques à d'autres relations de l'ontologie, comme la relation « contre-indication », la relation « étiologie (cause) » ou encore la relation « effets-secondaires ». Nous sommes persuadé que ces extensions amélioreront la performance du système Esculape en lui permettant de bénéficier d'une couverture plus large des connaissances du domaine médical.

Enfin, à plus long terme, le travail réalisé pourra être amélioré par d'autres voies. Dans un premier temps, nous souhaiterions fournir au système de question-réponse la compétence nécessaire pour lui permettre de gérer les dérivations morphologiques du vocabulaire médical. En effet, le lexique médical a largement recours à des mots construits par dérivation ou composition savante à partir de bases connues (Zweigenbaum, 2001). Dans un second temps, nous pensons élargir l'utilisation des patrons linguistiques à la justification de réponses sur plusieurs passages de textes issus de différents documents. La problématique de la

justification des réponses est en effet un des aspects les plus avancés des systèmes de question-réponse dans la mesure où son objectif est de mettre en évidence la chaîne d'inférences permettant de faire le lien entre une question et une réponse. Elle fait par ailleurs l'objet de nombreux travaux actuellement comme en atteste l'introduction de la tâche « Answer Validation Exercise » au cours de l'évaluation CLEF-QA 2006, cette tâche se donnant précisément pour but de justifier automatiquement les réponses retournées par les systèmes de question-réponse participant à CLEF-QA. Dans cette perspective, nous envisageons plus précisément de mettre à profit la possibilité d'extraire grâce à des patrons de nouvelles relations à partir des documents traités pour répondre à une question afin de réaliser des inférences permettant de justifier des réponses. Cette extension s'inscrit dans le cadre des travaux du projet CONIQUE sur les systèmes de question-réponse avancés.

# Bibliographie

**Aberdeen J. & Burger J. & Day D. & Hirschman L. & Robinson P. & Vilain M. -** MITRE: Description of the Alembic system as used for MUC-6. *In: proceedings of the 6th Message Understanding Conference (MUC-6),* Morgan Kaufmann, San Francisco, p. 141-155, 1995.

**Ahn K. & Bos J. & Clark S. & Curran J. R. & Dalmas T. & Leidner J. L. & Smillie M. B. & Webber B. -** Question answering with QED and Wee at TREC 2004. *Voorhees E.M., Buckland L.P., Eds, 13th Text REtrieval Conference (TREC 2004),* Gaithersburg, MD, USA, 2004.

**Alper B. S. & Stevermer J. J. & White D. S. & Ewigman B. G. -** Answering family physicians clinical questions using electronic medical databases. J Fam Pract, vol. 50, n° 11, p. 960-965, 2001.

**Assadi H. -** Construction d'ontologies à partir de texts techniques : Applications aux systèmes documentaries. *Thèse de doctorat,* Université Paris 6, 1998.

**Ayache C. -** Campagne EVALDA/EQUER : Evaluation en question-réponse, rapport final de la campagne EVALDA/EQUER. *Rapport interne, ELDA,* Paris, 2005. Disponible à (http://www.technolangue.net/IMG/pdf/rapport_EQUER_1.2.pdf).

**Ayache C. & Grau B. & Vilnat A. & -** EQueR: the French evaluation campaign of question answering system EQueR/Evalda. *In: Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC 2006),* Genoa, Italy, 2006, p. 1157-1160.

**Balvet A. & Embarek M. & Ferret O. –** Minimalisme et question-réponse : le système Œdipe. In : 12ème Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2005), p. 77-80, Dourdan, France, 2005.

**Baneyx A. -** Construire une ontologie de la pneumologie. *Thèse de doctorat,* Université de Paris 6, France, 2007.

**Banko M. & Brill E. & Dumais S. -** An analysis of the AskMSR question answering system. *In: Proceedings of the 2002 Conference on Empirical Methods in natural language processing,* 2002.

**Basili R. & Pazienza M. T. & Stevenson M. & Velardi P. & Vindigni M. & Wilks Y. -** An empirical approach to lexical tuning. *In: Proceedings of the Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications (First International Conference on Language Resources and Evaluation LREC 1998),* P. Velardi (ed.), Grenada, 1998.

**Béchet F. & Nasr A. & Genet F. -** Tagging unknown proper names using decision trees. *In: proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'2000),* Hong-Kong, p. 77-84, 2000.

**Benamara F. -** Cooperative question answering in restricted domains: the Webcoop experiment. *In: ACL-Portability of systems,* Barcelona, MIT Press, 2004, p. 98-110.

**Berthelin J.-B & de Chalendar G. & El Kateb F. & Ferret O. & Grau B. & Hurault-Plantet M. & Illouz G. & Monceaux L. & Robba I. & Vilnat A. -** Trouver des réponses sur le Web et dans une collection fermée. *Document de travail dans le cadre de l'action RIP-WEB.* Journée RIP-WEB du 01 décembre 2003, Orsay, France, 2003.

**Besançon R. & De Chalendar G. & Ferret O. & Fluhr C. & Mesnard O. & Naets H.** - Concept-Based Searching and Merging for Multilingual Information Retrieval: First Experiments at CLEF 2003. *LNCS 3237, Springer Verlag*, p. 174-184, 2004.

**Besançon R. & Embarek M. & Ferret O. -** Integrating new language in a multilingual search system based on a deep linguistic analysis. *In: Multilingual Information Access for Text, Speech and Images – 5th Workshop of the Cross-Language Evaluation Forum (CLEF'2004), vol. 3491 of Lecture Notes in Computer Science, Springer Berlin*, p. 83-83, 2005a.

**Besançon R. & Embarek M. & Ferret O. -** The Œdipe system at CLEF-QA 2005. In: 6th *Workshop of the Cross-Language Evaluation Forum (CLEF'2005)*, volume 4022 of Lecture Notes in Computer Science, p. 337-346, Springer Verlag, 2005b.

**Besançon R. & Embarek M. & Ferret O. -** Finding answers in the Œdipe system by extracting and applying linguistic patterns. In: 7th *Workshop of the Cross-Language Evaluation Forum (CLEF'2006), Selected revised papers, Lecture Notes in Computer Science, Springer Verlag*, 2006.

**Bikel D. & Miller S. & Schwartz R. & Weischedel R. -** *In: Proceedings of the 5th Conference on Applied Language Processing (ANLP'97),* Washington, p. 195-201, 1997.

**Blaschke C. & Andrade M. & Ouzounis C. & Valencia A. -** Automatic extraction of biological information from scientific text: Protein-protein interactions. *In: Proceedings of ISMB*, 1999.

**Bouaud J. & Bachimont B. & Charlet J. & Zweigenbaum P. -** Methodological principles for structuring an "ontology". In: *IJCAI'95 Workshop on "Basic Ontological Issues in Knowledge Sharing"*, 1995.

**Bourigault D. -** UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN'02),* Nancy, p. 75-84, 2002.

**Bourigault D. -** LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquisition de connaissances à partir de texts. *Thèse de doctorat,* EHESS, 1994.

**Buitelaar P. & Cimiano P. & Racioppa S. Siegel M. -** Ontology-based information extraction with SOBA. *In: proceedings of Language Resources and Evaluation Conference (LREC 2006),* Genoa, Italy.

**Bunescu R. & Ge R. & Kate R. & Marcotte E. & Mooney R. J. & Ramani A. K. & Wong Y. W. -** Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine,* 33(2), 2005.

**Burger J. & Cardie C. & Chaudhri V. & Gaizauskas R. & Harabagiu S. & Israel D. & Jacquemin C. & Lin C. & Maiorano S. & Miller G. & Moldovan D. & Ogden B. & Prager J. & Riloff E. & Singhal A. & Shriari R. & Strzalkowski T. & Voorhees E. & Weishedel R. -** Issues, tasks and program structures to roadmap research in Question & Answering (Q&A). *Rapport technique*, NIST, 2003.

**Cao T. D. & Dieng-Kuntz R. & Fiès B. & Bourdeau M. -** Vers un système d'aide à la veille technologique guidé par une ontologie. In : *Actes de la Conférence Francophone de reconnaissances des Formes et Intelligence Artificielle (RFIA'2006),* Tours, p. 25-27, 2006.

**Chandrasekaran B. & Josephson J. R. & Benjamins V. R. -** What are ontologies and why do we need them?. *IEEE Intelligent Systems.* Vol. 14, p. 20-26, 1999.

**Charlet J. & Bachimont B. & Bouaud J. Zweigenbaum P. -** Ontologie et réutilisabilité : expérience et discussion. *In : Acquisition et Ingénierie des Connaissances, N. Aussenac, P. Laubelet and C. Reynaud (éd.),* p. 69-87, Cépaduès-Editions, Toulouse, 1996.

**Chu-Carroll J. & Prager J. & Welty C. & Czuba K. & Ferrucci D. -** A Multi-Strategy and Multi-Source approach to question answering. *In: Proceedings of the 11th Text Retrieval Conference (TREC-11),* 2002.

**Cimino J. J. -** From data to knowledge through concept-oriented terminologies: experience with the Medical Entities Dictionary. *Journal of the American Medical Informatics Association (JAMIA),* p. 288-297, 2000.

**Claveau V. & Sébillot P. -** Extension de requêtes par lien sémantique nom-verbe acquis sur corpus. *In: Traitement Automatique des Langues Naturelles (TALN 2004),* Fès, Maroc, 2004.

**Collier N. & Nobata C. & Tsujii J. -** Extracting the names of genes and gene products with a hidden markov model. *In: Proceedings of COLING 2000,* p. 201-207, 2000.

**Condamines A. & Amsili P. -** Terminology between language and knowledge: an example of terminological knowledge base. *In: Proceedings of the 3rd International Congress on Terminology and Knowledge Engineering,* Cologne, Germany, 1993.

**Craven M. -** Learning to extract relations from Medline. *In: AAAI-99 Workshop on Machine Learning for Information Extraction,* Orlando, Florida, USA, 1999.

**Crestan É & Lemaire É & de Loupy C. -** Ressources pour un système de question/réponse. *In: Traitement Automatique des Langues Naturelles (TALN 2004),* Fès, Maroc, 2004.

**Cruse D. A. -** Lexical semantics. *Textbooks in Linguistics.* Cambridge University Press, 1986.

**Cucchiarelli A. & Velardi P. -** Unsupervised named entity recognition using syntactic and semantic contextual evidence. *Computational Linguistics*, n° 27 (1), p.123-131, 2001.

**Cui H. & Kan M. Y. & Chua T. S. -** Generic soft pattern models for definitional question answering. *In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development of Information Retrieval (SIGIR'2005),* Salvador, Brazil, 2005.

**Daelemans W. & Van den Bosch A. -** Memory-Based language processing. *Studies in Natural Language Processing*, Cambridge University Press, 2005.

**Darmoni S. J. & Consortium VUMeF -** VUMeF : extending the French involvement in the UMLS Metathesaurus. *AMIA annual symposium proceedings,* 2003.

**Delbecque T. & Jacquemart P. & Zweigenbaum P. -** Utilisation du réseau sémantique de l'UMLS pour la définition de types d'entités nommées médicales. *In : CORIA (Conférence en Recherche d'Informations et Applications),* Grenoble, p. 101-115, 2005.

**Du Y. & Huang X. & Li X. & Wu L. -** A novel pattern learning method for open domain question answering. *In: International Joint Conference on Natural Language Processing (IJCNLP'04),* p. 81-89, 2004.

**Durme B. V. & Huang Y. & Kupsc A. & Nyberg E. -** Towards light semantic processing for question answering. *HLT-NAACL 2003 Workshop on Text Meaning,* Edmonton, Canada, p. 54-61, 2003.

**Ehrmann M.** - Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation. *Phd thesis*, Université Paris Diderot, 2008.

**Ely J. W. & Osheroff J. A. & Ebell M. H. & Bergus G. R. & Levy B. T. & Chambliss M. L. & Evans E. R.** Analysis of questions asked by family doctors regarding patient care. *British Medical Journal,* 319, p. 358-361, 1999.

**Ely J. W. & Osheroff J. A. & Gorman P. N. & Ebell M. H. & Chambliss M. L. & Pifer E. A. & Stavri P. Z. -** A taxonomy of generic clinical questions: classification study. *British Medical Journal,* 321, p. 429-432, 2000.

**Ely J. W. & Osheroff J. A. & Ebell M. H. & Chambliss M. L. & Vinson D. & Stevermer J. & Pifer E. -** Obstacles to answering doctor's questions about patient care with evidence: qualitative study. *British Medical Journal,* 324, p. 710-713, 2002.

**Embarek M. & Ferret O. -** Extraction de relations sémantiques à partir de textes dans le domaine médical. *In : JOBIM 2006,* session poster, Bordeaux, France, July, 2006.

**Embarek M. & Ferret O. -** Une expérience d'extraction de relations sémantiques à partir de textes dans le domaine médical. In : *Traitement Automatique des Langues Naturelles (TALN 2007),* p. 37-46, Toulouse, France, 2007.

**Fellbaum C. -** WordNet : An Electronic Lexical Database. *The MIT Press*. 1998.

**Ferret O. & Grau B. & Hurault-Plantet M. & Illouz G. & Jacquemin C. & Masson N. & Lecuyer P. -** QALC-the question answering system of LIMSI-CNRS. *In: Technical report: LIMSI-CNRS TREC 9 evaluation,* 2000.

**Ferret O. & Grau B. & Hurault-Plantet M. & Illouz G. & Jacquemin C. -** Utilisation des entités nommées et des variantes terminologiques dans un système de question-réponse. *In: Traitement Automatique des Langues Naturelles (TALN 2001),* Tours, 2001.

**Ferret O. & Grau B. & Hurault-Plantet M. & Illouz G. & Jacquemin C. -** Document selection refinement based on linguistic features for QALC, a question answering system. *In: Recent Advances in Natural Language Processing (RANLP 2001),* Tzigov Chark, Bulgaria, 2001b.

**Ferret O. & Grau B. & Hurault-Plantet M. & Illouz G. & Monceaux L. & Robba I. & Vilnat A.** - Recherche de la réponse fondée sur la reconnaissance du focus de la question. In *Traitement Automatique des Langues Naturelles (TALN 2002),* Nancy : TALN, 2002a.

**Ferret O. & Grau B. & Hurault-Plantet M. & Illouz G. & Jacquemin C. & Monceaux L. & Robba I. & Vilnat A.** - How NLP can improve question answering. *Knowledge organization,* vol. 29 (3-4), p. 135-155, 2002b.

**Ferret O. & Zweigenbaum P.** - Représentation sémantique des connaissances pour les systèmes de question-réponse. *In : Brigitte Grau and Jean-Pierre Chevallet, editors, La recherche d'informations précises : traitement automatique de la langue, apprentissage et connaissances pour les systèmes de question-réponse,* chapitre 4, p. 133-169. Hermès-Lavoisier, Paris, 2007.

**Finkelstein-Landau M. & Morin E. -** Extracting semantic relationships between terms: Supervised vs. unsupervised methods. *In: International Workshop on Ontological Engineering on the Global Information Infrastructure,* p. 71-80, 1999.

**Fluhr C. & Schmit D. & Ortet P. & Elkateb F. & Gurtner K. –** Spirit-w3, a distributed crosslingual indexing and retrieval engine. *In : INET'97,* 1997.

**Fourour N. & Morin E. -** Apport du Web dans la reconnaissance des entités nommées. *Revue Québécoise de Linguistique (RQL),* Vol. 32, n° 1, p. 63-92, 2003.

**Fox M. & Gruninger M. -** Ontologies for enterprise integration. *In : Proceedings of the 2nd Conference on Cooperative Information Systems,* Toronto, 1994.

**Fukuda K. -** Toward information extraction: identifying protein names from biological papers. *In: Proceedings of the Pacific Symposium on Biocomputing,* p. 705-716, 1998.

**Gangemi A. & Galanti M. & Galeazzi E. & Rossi Mori A. -** Beyond UMLS: Computational semantics for medical records. In: *Proceedings of MEDINFO 1992* edited by LUN K. C., DEGOULET P., PIEMME T. RIENHOFF O., Geneva, p. 703-708, 1992.

**Gillard L. & Bellot P. & El-Bèze M. -** Le LIA à EQueR. *In : TALN 2005,* p. 81-84, Dourdan, France, 2005.

**Gillard L. & Bellot P. & El-Bèze M. -** Influence de mesures de densité pour la recherche de passages et l'extraction de réponses dans un système de questions-réponses. *In : Actes de la Troisième Conférence en Recherche d'Information et Applications (CORIA 2006),* éditeur ARIA, p. 193-204, 2006.

**Girju R. & Badulescu A. & Moldovan D. -** Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1), 2006.

**Giuliano C. & Lavelli A. & Romano L. -** Relations extraction and the influence of automatic named-entity recognition. *ACM Transactions on Speech and Language Processing (TSLP),* Volume 5, p. 1-26, New York, USA, 2007.

**Golbreich C. & Dameron O. & Gibaud B. & Burgun A. -** Standards et ontologies biomédicales pour un Web sémantique. Rapport interne, université de Rennes-1, 2002.

**Gomez-Pérez A. & Fernandez-Lopez M. & Corcho O. -** Ontology development methods and methodologies. *Ontological Engineering,* Springer Verlag, Madrid, Spain, p. 113-153, 2004.

**Gorman P. & Ash J. & Wykoff L. -** Can primary care physicians' questions be answered using the medical journal literature? *Bulletin of the Medical Library Association,* 82(2), p. 140-146, 1994.

**Graesser A. & Person N. & Huber J. -** Mechanisms that generate questions. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1992.

**Grau B. & Ligozat A. & Robba I. & Vilnat A. & Monceaux L. -** FRASQUES : A quesion-answering system in the EQueR evaluation campaign. *In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006),* 2006a.

**Grau B. & Ligozat A. & Robba I. & Vilnat A. & Bagur M. & Séjourné K. -** The bilingual system MUSCLEF at QA@CLEF 2006. *In: Working Notes, CLEF Crosss-Language Evaluation Forum, Alicante, Espagne, 2006b.*

**Green B. & Wolf A. & Chomsky C. & Laughery K. –** Baseball: an automatic question answerer. In*: Proceedings of the Western Joint Computer Conference,* p. 219-224, 1961.

**Grefenstette G. -** SEXTANT: Exploring unexplored contexts for semantic extraction from syntactic analysis. *In: Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL'92),* Newark, Delaware, 1992.

**Grefenstette G. -** Explorations in automatic thesaurus discovery. *Kluwer Academic Publishers,* Boston, 1994.

**Grefenstette G. -** The WWW as a resource for example-based MT tasks. *In: Proceedings of ASLIB Translating and the Computer Conference*, London, 1999.

**Grishman R. & Sundheim B. -** Design of the MUC-6 evaluation, *Actes de Message Understanding Conferences (MUC-6),* NIST, Eds, Morgan Kauffmann Publisher, Columbia, MD, 1995.

**Gruber T. R. -** A translation approach to portable ontology specifications. *Knowledge acquisition,* Vol. 5, p. 199-220, 1993.

**Gu H. & Perl Y. & Geller J. & Halper M. & Singh M. –** A methodology for partitioning a vocabulary hierarchy into trees. *In: Journal of the Artificial Intelligence in Medicine (JAIM),* Vol. 15, n° 1, p. 77-98, 1999.

**Habert B. & Nazarenko A. -** La syntaxe comme marche-pied de l'acquisition de connaissances : Bilan critique d'une expérience. *Actes des septièmes Journées Acquisition des Connaissances (JAC'96),* Sète, p. 137-148, 1996.

**Hakimpour F. & Geppert A. -** Resolving semantic heterogeneity in schema integration: an ontology based approach. *In: Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS-01),* Ogunquit, Maine, USA, p. 297-308, 2001.

**Harabagiu S. A. & Miller A. G. Moldovan D. I. -** WordNet 2: A morphologically and semantically enhanced resource, *In: Proceedings of SIGLEX-99,* University of Maryland, 1999.

**Harabagiu S. & Moldovan D. & Psca M. & Mihalcea R. & Surdeanu M. & Bunescu R. & Gîrju R. & Rus V. & Morarescu P. -** Falcon: Boosting knowledge for answer engines. *In: Proceedings of the 9th Text Retrieval Conference (TREC-9),* NIST, p. 479-488, 2001.

**Harabagiu S. & Moldovan D. -** Tutorial on open-domain textual question answering. *In: Proceedings of the 19th International Conference on Computational Linguistics (COLING'02),* Taipei, Taiwan, 2002.

**Harabagiu S. & Moldovan D. & Clark C. & Bowden M. & Hickl A. & Wang P. -** Employing two question answering systems in TREC-2005. *NIST, ED., 14th Text REtrieval Conference (TREC 2005),* Gaithersburg, MD, USA, 2005.

**Harris Z. S. -** Mathematical structures of language. *Wiley*, New York, 1968.

**Hearst M. -** Automatic acquisition of hyponyms from large text corpora. *In: Proceedings of the 14th International Conference on Computational Linguistics (COLING'92),* Nantes, France, 1992.

**Hendrix G. –** Human engineering for applied natural language processing. *In: Proceedings of the 5th International Joint Conference on Artificial Intelligence (IJCAI),* 1977.

**Hermjakob U. -** Parsing and question classification for question answering. *In: Proceedings of the Association for Computational Linguistics 2001 Workshop on Open-Domain question answering,* p. 17-22, 2001

**Hirshberg D. S. -** Algorithms for the Longest Common Subsequence problem. *Journal of the ACM,* vol. 24, 1977.

**Hovy E. & Gerber L. & Hermjakob U. & Junk M. & Lin C. Y. -** Question answering in Webclopedia. NIST, Ed., 9th Text REtrieval Conference (TREC-9), Gaithersburg, MD, USA, 2001a.

**Hovy E. & Gerber L. & Hermjakob U. & Lin C. Y. & Ravichandran D. -** Toward semantics-based answer pinpointing. *In: Proceedings of the Human Language Technology Conference (HLT'01),* 2001b.

**Jacquemart P. & Zweigenbaum P. -** Towards a medical question-answering system: a feasibility study. *In: R. Baud, M. Fieschi, P. Le Breux & P. Ruch, Rédacteurs, Actes Medical Informatics Europe, vol. 95 of Studies in Health Technology and Informatics,* p. 463-468, Amsterdam: IOS press, 2003.

**Jacquemin C. -** Syntagmatic and paradigmatic representations of term variation, *Actes de ACL 1999,* p. 341-348, 1999.

**Jacquemin C. & Bush C. -** Fouille du Web pour la collecte d'entités nommées. *Actes de la 7ème Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2000),* p. 187-196, 2000a.

**Jacquemin C. & Zweigenbaum P. -** Traitement automatique des langues pour l'accès au contenu des documents. *In : Le document multimédia en science du traitement de l'information,* éditeurs : Charlet J. Le Maitre J. et Grabay C., Éditions CÉPADUÈS, p. 71-110. Toulouse, 2000b.

**Jousse F. & Tellier I. & Tommasi M. & Marty P. -** Learning to extract answers in question answering: Experimental studies. *In : CORIA (Conférence en Recherche d'Informations et Applications),* p. 85-100, Grenoble, 2005.

**Katz B. & Lin J. J. & Felshin S. -** The Start multimedia information system : Current technology and future directions. *In: Proceedings of the International Workshop on Multimedia Information Systems,* p. 117-123, 2002.

**Khoo C. S. G. & Chan S. & Niu Y. -** Extracting causal knowledge from a medical database using graphical patterns. *In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'2000),* p. 336-343, Hong Kong, 2000.

**Kubala F. & Schwartz R. & Stone R. & Weischedel R. -** Named entity extraction from speech. *In: Proceedings of the DARPA Broadcast News Workshop,* Herndon, p. 287-292, 1999.

**Kwok C. & Etzioni O. & Weld D. S. -** Scaling question answering to the Web. *In: Tenth World Wide Web Conference,* Hong Kong, China, 2001.

**Laurent D. & Séguéla P. -** Qristal, système de Question-Réponse. *In: Traitement Automatique des Langues Naturelles (TALN 2005),* Dourdan, 2005.

**Le Roux E. -** Extraction d'information dans des textes libres guidée par une ontologie. *Thèse de doctorat Sciences du Langage,* Université de Paris X, Nanterre, France, 2003.

**Lehnert W. -** Human and computational question answering. Cognitive Science, 1, p. 47-63, 1977.

**Lehnert W. -** The process of question answering: A computer simulation of cognition. *Lawrence Erlbaum Associates,* 1978.

**Lenat D. B. -** Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM,* vol. 38, no. 11, p. 33-38, 1995.

**Levenshtein V. -** Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory,* 10(8), p. 707-710, 1966.

**Lin J. -** The Web as a resource for question answering: Perspectives and challenges. *In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002),* 2002.

**Lin J. -** An exploration of the principles underlying redundancy-based factoid. *In: ACM Transactions on Information Systems (TOIS),* Vol. 27, n° 2, 2007.

**Lindberg D. A. B. & Humphreys B. L. -** The UMLS knowledge sources: Tools for building better user interfaces. *In: Miller RA, ed. Proceedings of the 14th annual SCAMC.* Washington, D.C. IEEE Computer Society Press, p. 121-125, 1990.

**Lindberg D. A. B. & Humphrey B. L. & McCray A. T. -** The Unified Medical Language System. *Methods of Information in Medicine,* 1993, p. 81-91.

**Lopez Garcia V. & Motta E. & Uren V. -** AquaLog: An ontology-driven question answering system to interface the semantic Web. *In: Proceedings of the 9th International Conference on Applications of Natural Language to Information Systems (NLDB),* 2004.

**Malaisé V. & Zweigenbaum P. & Bachimont B. -** Repérage et exploitation d'énoncés définitoires en corpus pour l'aide à la construction d'ontologie. *In: Traitement Automatique des Langues Naturelles (TALN 2004),* Fès, Maroc, 2004.

**Malaisé V. -** Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles à partir de corpus textuels. *PhD thesis,* Université Paris 7 – Denis Diderot, France, 2005.

**Malaisé V. & Delbecque T. & Zweigenbaum P. -** Recherche en corpus de réponses à des questions définitoires. *In: Traitement Automatique des Langues Naturelles (TALN 2005),* Dourdan, 2005.

**McCray A. T. -** The UMLS semantic network. *In: 13th Annual Symposium on Computer Applications in Medical Care,* Washington DC, USA, 1989, p. 475-480.

**McCray A. T. & Srinivasan S. & Browne A. C. -** Lexical methods for managing variation in biomedical terminologies. *In: Proceedings of the Annual SCAMC,* p. 235-239, 1994.

**McDowell L. & Cafarella M. J. -** Ontology-Driven information extraction with OntoSyphon. *In: International Semantic Web Conference,* p. 428-444, 2006.

**Mendes S. & Moriceau V. -** L'analyse des questions : intérêts pour la génération des réponses. TALN 2004 Workshop Question-Réponse, Fès, Maroc, 2004.

**Miller G. –** Wordnet : an On-line lexical database. *International Journal of Lexicography,* 3(4). 1990.

**Moldovan D. & Harabagiu S. & Girju R. & Morarescu P. Lacatusu F. & Novischi A. Badulescu A. & Bolohan O. -** LCC tools for question answering. *In: Proceedings of the 11th Text Retrieval Conference (TREC-11),* 2002.

**Moldovan D. & Clark C. & Harabagiu S. & Maiorano S. -** COGEX: A logic prover for question answering, HLT-NAACL 2003, Edmondton, Canada, p. 87-93, 2003a.

**Moldovan D. & Pasca M. & Harabagiu S. & Surdeanu M. -** Performance issues and error analysis in an open-domain question answering system. *In: ACM transactions on Information Systems,* volume 21, p. 133-154, 2003.

**Monceaux L. & Robba I. -** Les analyseurs syntaxiques : atouts pour une analyse des questions dans un système de question-réponse. *Actes de Traitement Automatique des Langues Naturelles (TALN 2002),* Nancy, 2002.

**Morin E. -** Extraction de liens sémantiques entre termes à partir de corpus de textes techniques. *Thèse de doctorat,* Université de Nantes, France, 1999.

**MUC-7 -** *In: Proceedings of the Seventh Message Understanding Conference,* MUC-7, 1998.

**Mukherjea S. & Sahay S. -** Discovering biomedical relations utilizing the World Wide Web. *In: Pacific Symposium on Biocomputing 11,* p. 164-175, 2006.

**Narayanan S. & Harabagiu S. -** Question answering based on semantic structures. *In: 20th International Conference on Computational Linguistics (COLING 2004),* Geneva, Switzerland, p. 22-29, 2004.

**Nazarenko A. & Zweigenbaum P. & Bouaud J. & Habert B.** - Corpus-based identification and refinement of semantic classes. *Journal of the American Medical Informatics Association,* p. 585-589, 1997.

**Nédellec C. -** Machine Learning for Information Extraction in Genomics - State of the art and perspectives. *In: S. Sirmakessis, Ed., Text Mining and its Applications: Results of the NEMIS Launch Conference,* Springer Verlag, 2004.

**Ng S. & Wong M. -** Toward routine automatic pathway discovery from on-line scientific text abstracts. *In: Genome Informatics,* vol. 10, p. 104-112, 1999.

**Nyberg E. & Mitamura T. & Carbonell J. & Callan J. & Collins-Thompson K. & Czuba K. & Duggan M. & Hiyakumoto L. & Hu N. & Huang Y. & Ko J. & Lita L. & Murtagh S. & Pedro V & Svoboda D. -** The Javelin question answering system at TREC 2002. *In: Proceedings of the 11th Text Retrieval Conference (TREC-11),* 2002.

**Nyberg E. & Mitamura T. & Carbonell J. & Callan J. & Carbonell J. G. & Frederking R. E. & Collins-Thompson K. & Hiyakumoto L. & Huang Y. & Huttenhower C. & Judy S. & Ko J. & Kupsc A. & Lita L. & Pedro V & Svoboda D. & Van Durme B. -** The Javelin question answering system at TREC 2003. *In: Proceedings of the 12th Text Retrieval Conference (TREC-12),* 2003.

**Pantel P. & Ravichandran D. & Hovy E. -** Towards terascale knowledge acquisition. *In: International Conference on Computational Linguistics (COLING'04),* p. 771-777, Geneva, Switzerland, 2004.

**Pearson J. -** Terms in context, Amsterdam/Philadelphia: John Benjamins Publishing Company, 1998.

**Plamondon L. & Kosseim L. -** Quantum: A function-based question answering system. *In: Proceedings of the 15th Conference of the Canadian Society for Computational Studies of Intelligence (AI 2002),* Calgary, Canada, 2002.

**Poibeau T. & Zweigenbaum P. & Nazarenko A. -** Traitement automatique des langues pour les systèmes de question/réponse. Document de travail dans le cadre de l'action RIP-WEB. Journée RIP-WEB du 30 septembre 2003, 2003.

**Proux D. & Rechenmann F. & Julliard L. & Pillet V. & Jacq B. -** Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction. *In: Genome Informatics,* vol. 9, p. 72-80, 1998.

**Pustejovsky J. & Castaño J. & Sauri R. & Rumshisky A. & Zhang J. & Luo W. -** Medstract: Creating large-scale information servers for biomedical libraries. *In: ACL 02 Workshop on Natural Language Processing in the biomedical domain,* Philadelphia, USA, 2002a.

**Pustejovsky J. & Castaño J. & Zhang J. -** Robust relational parsing over biomedical literature: Extract inhibit relations. *In: Pacific Symposium on Biocomputing (PSB'02),* p. 362-373, 2002b.

**Ramani C. & Marcotte E. & Bunescu R. & Mooney R. -** Using biomedical literature mining to consolidate the set of known human protein-protein interactions. *In: Proceedings ISMB/ACL Biolink 2005,* 2005.

**Ravichandran D. & Hovy E. -** Learning surface text patterns for a question answering system. *In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics Conference (ACL 2002),* Philadelphia, USA, 2002.

**Ravichandran D. -** Terascale knowledge acquisition. *Ph.D. Thesis,* University of Southern California, 2005.

**Rebeyrolle J. -** Forme et fonction de la définition en discours, *Thèse de doctorat,* Université de Toulouse II – Le Mirail, France, 2000.

**Rector A. L. & Rogers J. E. & Pole P. A. -** The GALEN high level ontology. *In: Proceedings of MIE 96, IOS press*, p. 174-178, 1996.

**Rector A. L. & Bechhover S. & Goble C. A. & Horrocks I. & Nowlan W. & Solomon W. -** The GRAIL concept modeling language for medical terminology. *In: Artificial Intelligence in Medicine,* vol. 9, p. 139-171, 1997.

**Riloff E. -** Information extraction as a basis for portable text classification systems. *Ph.D. de l'Université du Massachusetts Amherst,* 1994.

**Rinaldi F. & Dowdall J. & Schneider G. -** Answering questions in the genomics domain. *In: Proceedings of ACL 04 : Workshop on question answering in restricted domains,* Barcelona, Spain, 2004.

**Rosario B. & Hearst M. -** Classifying semantic relations in bioscience texts. *In: Proceedings of the 42nd Annual Meeting of Association of Computational Linguistics (ACL'04),* Barcelona, Spain, 2004.

**Rosario B. -** Extraction of semantic relations from bioscience text. *Ph.D. Thesis,* University of California, Berkeley, 2005.

**Sackett N. -** Evidence-based medicine: how to how to practice and teach EBM. *Churchill Livingstone Inc.*, New York, 1997.

**Sasaki Y. & Matsuo Y. -** Learning semantic-level information extraction rules by type-oriented ILP. *In: Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000),* Sarrebrück, p. 698-704, 2000.

**Séguéla P. & Aussenac-Gilles N. -** Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. *In : Actes de la conférence Ingénierie des Connaissances (IC'99),* p. 79-88, Palaiseau, 1999.

**Skuce D. R. & Meyer I. -** Terminology and knowledge acquisition: Exploring a symbiotic relationship. *In: Proceedings of the 6th Knowledge Acquisition for Knowledge Based Systems Workshop,* Banff, Canada, 1991.

**Snow R. & Jurafsky D. & Ng A. Y. -** Learning syntactic patterns for automatic hypernym discovery. *In: Neural Information Processing Systems (NIPS),* 2005.

**Soubbotin M. M. & Soubbotin S. M. -** Patterns of potential answer expressions as clues to the right answer. *In: In: Proceedings of the Text REtrieval Conference (TREC-10), NIST*, editor, p. 175-182, Gaithersburg, USA, 2001.

**Soubbotin M. M. & Soubbotin S. M. -** Use of patterns for detection of answer strings: A systematic approach. *In: Proceedings of the Text REtrieval Conference (TREC-11),* Gaithersburg, USA, 2002.

**Sowa J. F. -** Knowledge representation: logical, philosophical and computational foundations. *Brooks Cole Publishing Company*, Pacific Grove, USA, 1999.

**Staab S. & Studer R. -** Handbook on ontologies. *Springer,* Berlin, Germany, 2003.

**Stapley B. J. & Benoit G. -** Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *In: Proceedings of the Pacific Symposium of Biocomputing*, Vol. 5, p. 529–540, 2000.

**Stephens M. & Palakal M. & Mukhopadhyay S. & Raje R. -** Detecting gene relations from Medline abstracts, *Pac Symp Biocomput,* 2001.

**Thieulle J. -** Pratique du mot médical. Cahier d'exercice, édition Lamarre, 1993.

**Uschold M. & Gruninger M. -** Ontologies: principles, methods and applications. *In: Knowledge Engineering Review*, Vol. 11, n° 2, p. 93-155, 1996.

**Vallin A. & Giampiccolo D. & Aunimo L. & Ayache C. & Osenova P. & Peñas A. & de Rijke M. & Sacaleanu B. & Santos D. & Sutcliffe R. -** Overview of the CLEF 2005 multilingual question answering track. *In: 6th Workshop of the Cross-Language Evaluation Forum CLEF 2005,* 2006.

**Van Zaanen M. -** Bootstrapping structure into language: Alignment-Based learning. *Ph.D. de l'Université de Leeds*, 2001.

**Vargas-Vera M. & Motta E. -** AQUA: An ontology-based question answering system. *In: Proceedings of Mexican International Conference on Artificial Intelligence (MICAI 2004),* Mexico City, Mexico, p. 468-477, 2004.

**Voorhees E. M. -** Query expansion using lexical-semantic relations. *In: Proceedings of ACM SIGIR'9*, Dublin, Irlande, 1994.

**Voorhees E. M. -** The TREC-8 question answering track report. *In: Proceedings of the Text REtrieval Conference (TREC-8),* 1999.

**Voorhees E. M. -** Overview of the TREC 2002 question answering track. *In: Proceedings of the Text REtrieval Conference (TREC-11),* 2002.

**Welty C. -** The ontological nature of subject taxonomies. *In: Proceedings of the 1st International Conference on Formal ontologies in Information Systems,* FOIS'98, Trento, Italy, p. 317-327, 1998.

**Wilensky R. -** Talking to Unix in English: an overview of an on-line Unix Consultant. Technical Report, Université de Californie à Berkeley, 1982.

**Winograd T. -** Understanding Natural Language. *Academic Press,* 1972.

**Winograd T. -** A procedural model of language understanding. *Readings in natural language processing,* 1986, p. 249-266, 1973.

**Woods W. A. -** Progress in natural language understanding: An application to lunar geology. *In AFIPS Conference Proceedings,* vol. 42, 1973, p. 441-450, 1973.

**Woods W. A. & Green S. & Martin P. & Houston A. -** Halfway to question answering. *In: Proceedings of the 9th Text Retrieval Conference (TREC 2000),* 2000.

**Yang H. & Chua T. S. -** The integration of lexical knowledge and external resources for question answering. *NIST, Ed., 11th Text REtrieval Conference (TREC 2002),* Gaithersburg, USA, 2002.

**Yu H. & Sable C. Zhu H. -** Classifying medical questions based on an evidence taxonomy. *In: AAAI-2005 Workshop,* Pittsburgh, Pennsylvania, 2005.

**Zweigenbaum P. & Consortium MENELAS -** MENELAS: an access system for medical records using natural language. *Computational Methods Programs Biomed,* vol. 45, p. 117-120, 1994.

**Zweigenbaum P. & Bachimont B. & Bouaud J. & Charlet J. & Boisvieux J-F -** Le rôle du lexique sémantique et de l'ontologie dans le traitement automatique de la langue médicale. *In : Le Beux P, Burgun A, editors. Actes du Colloque CRISTAL'S.* Saint-Malo, 1996.

**Zweigenbaum P. -** Traitements automatiques de la terminologie médicale. Revue française de linguistique appliquée, VI(2), p. 47-62, 2001.

**Zweigenbaum P. & Baud R. & Burgun A. & Namer F. & Jarrousse É. & Grabar N. & Ruch P. & Le Duff F. & Thirion B. & Darmoni S. -** UMLF : construction d'un lexique médical francophone unifié. *Journée Francophone d'Informatique médicale,* Tunis, 2003.

**Zweigenbaum P. -** Question answering in biomedicine. *In: Proceedings of EACL 03 Workshop: Natural Language Processing for Question Answering,* Budapest, Hungary, 2003.

**Zweigenbaum P. -** l'UMLS entre langue et ontologie : une approche pragmatique dans le domaine médical. *Revue d'Intelligence Artificielle,* p. 111-137, 2004.

# Annexes

# Annexe 1 Questions de la tâche médicale EQueR

MF1 Pour quelles raisons une consultation diététique est-elle préconisée ?
MF2 Quand rechercher une insuffisance rénale ?
MF3 Quel est le rôle des stations de base dans les communications mobiles ?
MF4 Comment le degré d'immobilité d'un patient peut-il être évalué ?
MF5 Quel est le gène responsable de l'aniridie ?
MF6 À partir de quel âge le dépistage des troubles de l'acuité visuelle est-il possible chez l'enfant ?
MF7 Où doit se dérouler une consultation diététique ?
MD8 Par quoi est caractérisée l'aspergillose bronchopulmonaire allergique ?
MF9 Quand doit-on procéder au dosage de la créatininémie ?
MF10 Comment organiser le suivi d'un patient atteint d'insuffisance rénale ?
MF11 Quand parle-t-on de crise dans le cadre d'une AVF ?
MD12 De quelle façon définir le sarcome Kaposi ?
MF13 À quel type de migraineux prescrit-on le traitement au méthysergide ?
MF14 Quel est le traitement du chérubisme ?
MD15 De quelle façon est-il possible de définir l'ostéosynthèse ?
MD16 Que veut dire "noyade sublétale" ?
MD17 Qu'est-ce qu'une infection opportuniste liée au SIDA ?
MF18 Quand est apparu le diéthylstilbestrol en France ?
MF19 Comment le poids corporel est-il déterminé ?
MF20 À quelle date le Livre Blanc sur la sécurité alimentaire a-t-il été adopté par la Commission Européenne ?
MF21 Quel pourcentage de risque existe-t-il d'avoir un enfant atteint du syndrome Coffin-siris pour un couple ayant déjà un enfant atteint ?
MF22 Quand la leucocyturie est-elle considérée comme pathologique ?
MF23 Qui prend en charge la ventilation manuelle ?
MD24 Que peut-on attribuer comme définition à l'Échelle de Glasgow ?
MD25 Quelle est la signification de "cordocentèse" ?
MF26 Quel est le traitement de l'artériopathie oblitérante des membres inférieurs ?
MF27 Comment prévenir l'engorgement mammaire ?
MD28 Quelle est la définition de "chimiothérapie" ?
MF29 Quel est le traitement approprié de l'état d'anxiété chez un patient en soins palliatifs ?
MF30 Comment rechercher l'hématurie ?
MF31 À quoi sont dues les fractures de l'ostéoporose cortisonique ?
MF32 Comment prendre en charge une aniridie ?
MF33 Qui doit réaliser une consultation de diététique ?
MF34 À quelle tranche d'âge peut-on être touché par le neuroblastome ?
MF35 Quelle est la deuxième cause de mortalité en France ?
MF36 Quel est le traitement de la schizophrénie ?
MF37 Quelle est la durée moyenne de l'agonie ?
MF38 Quel est l'effet essentiel de la corticothérapie sur l'os ?
MF39 Quelle est la cause du syndrome de CACH ?
MF40 Dans quel contexte le diéthylstilbestrol fut-il utilisé ?

MF41 Quel est le cancer féminin le plus fréquent ?
MF42 Quels sont les éléments qui distinguent la migraine de l'adulte de celle de l'enfant ?
MF43 Quand fait-on appel à la technique chirurgicale Tension-free Vaginal Tape (TVT) ?
MF44 En quoi consiste la préparation cutanée préopératoire ?
MF45 Quel est le traitement de première intention prescrit à un patient dyslipidémique ?
MF46 Quel était le coût global du cancer en France en 1994 pour l'Assurance Maladie ?
MF47 Quel est le traitement de l'acrodermatite entéropathique ?
MF48 À quoi est due la trisomie 21 ?
MF49 Par quelle bactérie est causée la méningite à méningocoque ?
MF50 Combien de décès par cancer ont-ils été dénombrés en 1995 ?
MF51 Quel est le traitement de l'algie vasculaire de la face ?
MD52 Qu'est-ce qu'une aniridie ?
MD53 Qu'est-ce qu'une maladie mentale ?
MD54 Que veut dire "maladie de Bell" ?
MD55 Qu'est-ce qu'une cholangite sclérosante ?
MD56 Qu'est ce qu'un mésothéliome ?
MD57 Qu'est-ce qu'une anomalie congénitale ?
MD58 Qu'est-ce que la fluorose dentaire ?
MD59 Que signifie "adénite" ?
MD60 Qu'est-ce qu'une hypoplasie du cœur gauche ?
MD61 Qu'est-ce qu'une anorexie ?
MD62 Quelle est la définition de la désinfection ?
MD63 Quelle est la définition du syndrome de CACH ?
MD64 Qu'est-ce que la radiothérapie ?
MD65 Comment peut-on définir la communication cellulaire ?
MD66 Quelle est la définition de l'asthme ?
MD67 Qu'est-ce qu'une perfusion parentérale ?
MD68 Qu'est-ce qu'une carie dentaire ?
MD69 Qu'est-ce que le séquençage ?
MD70 Qu'est-ce que le syndrome du décalage horaire ?
MD71 Qu'est-ce qu'un trouble dépressif ?
MD72 Quelle est la définition du chérubisme ?
MD73 Comment définir la néonatologie ?
MD74 Qu'est-ce qu'un trouble bipolaire ?
MD75 Qu'est-ce que l'Index de Pression Systolique ?
MD76 Qu'est-ce que la boulimie ?
MD77 Qu'est-ce que la schizophrénie ?
MD78 Qu'est-ce que l'hydrargyrisme ?
MD79 Qu'est-ce qu'une amblyopie ?
MD80 Quelle est la définition de la polyurie ?
MD81 Qu'est-ce que l'hyperoxie ?
MD82 Qu'est-ce qu'un scanner ?
MD83 Quelle est la définition de l'acrodermatite entéropathique ?
MD84 Qu'est-ce qu'une AVF ?
MD85 Qu'est-ce qu'une ischémie critique chronique ?
MD86 Qu'est-ce que l'HBP ?
MD87 Quelle est la définition de la génomique ?
MD88 Qu'est-ce que l'acide folique ?
MD89 Quelle est la définition du neuroblastome ?
MD90 Qu'est-ce qu'un bilan urodynamique ?

MD91 Qu'est-ce que la thérapie génique ?
MD92 Qu'est-ce qu'une mastite ?
MD93 Qu'est-ce que l'antisepsie ?
MD94 Qu'est-ce que la virémie ?
MD95 Qu'est-ce qu'une anomalie réductionnelle des membres ?
MD96 Qu'est-ce que l'oxygénothérapie ?
MD97 Qu'est-ce qu'une artériopathie oblitérante des membres inférieurs ?
MD98 Qu'est-ce qu'une sialographie ?
MD99 Qu'est-ce qu'une fente labio-palatine ?
MD100 Qu'est-ce qu'un abcès abdominal ?
MB101 Le chérubisme est-il une affection génétique ?
MB102 Le cancer peut-il être transmissible par voie sexuelle ?
MB103 Le sumatriptan est-il indiqué dans le traitement de l'AVF ?
MB104 L'aniridie peut-elle s'accompagner d'un retard mental ?
MB105 L'infirmière joue-t-elle un rôle dans la communication en direction de l'enfant malade ou de ses parents ?
MB106 Le décalage horaire peut-il engendrer une baisse des capacités motrices ?
MB107 L'amblyopie est-elle liée aux troubles de la réfraction ?
MB108 Une sensation de poids au niveau du ventre peut-elle constituer un symptôme du cancer de l'ovaire ?
MB109 La pose d'amalgame dentaire peut-elle provoquer des allergies ?
MB110 L'insémination artificielle post-mortem est-elle autorisée par la loi en Allemagne ?
MB111 Les antiseptiques sont-ils capables d'inhiber la croissance des micro-organismes ?
MB112 Les antécédents familiaux de strabisme ou de troubles de la réfraction exposent-ils à un risque accru d'apparition de l'amblyopie ?
MB113 La méningite peut elle entraîner rapidement la mort ?
MB114 Un enfant peut-il être atteint de schizophrénie ?
MB115 Le diagnostic anténatal est-il possible ?
MB116 L'apparition d'une fibrose pulmonaire peut-elle être liée à une exposition à l'amiante ?
MB117 Le mercure est-il un métal toxique ?
MB118 Est-ce que le neuroblastome est un cancer de l'enfant ?
MB119 La chimiothérapie consiste-t-elle à traiter le neuroblastome par des médicaments ?
MB120 La méningite peut-elle se transmettre directement par des gouttelettes de mucus provenant de la gorge et du nez d'une personne infectée ?
MB121 L'acrodermatite entéropathique est-elle une maladie récessive autosomique ?
MB122 L'allaitement maternel est-il contre-indiqué chez une femme portant une prothèse mammaire ?
MB123 L'alimentation parentérale est-elle indiquée durant l'agonie ?
MB124 L'AVF est-elle une pathologie touchant essentiellement l'enfant ?
ML125 Quelles sont les 2 situations où l'allaitement maternel peut être contre-indiqué ?
ML126 Citez 5 critères diagnostics de l'aniridie.
ML127 Citez 7 symptômes de l'hypertrophie bénigne de la prostate.
ML128 Citez 5 causes pouvant jouer un rôle dans l'apparition d'une maladie mentale.
ML129 Citez 10 symptômes de l'aniridie.
ML130 Quels sont les 3 pays européens où l'accès au dossier médical est prévu par la loi ?
ML131 Quels sont les 4 stades du cancer de l'ovaire ?
ML132 Quels sont les trois types d'examens à réaliser en cas de suspicion d'un neuroblastome ?
ML133 Quels sont les 6 facteurs de risque (supérieurs à 4,0) du cancer du sein ?

ML134 Quelles sont les 4 localisations possibles des neuroblastomes ?
ML135 Citez 7 situations pour lesquelles une exposition au diéthylstilbestrol devra être recherchée.
ML136 Citez 5 effets secondaires d'une corticothérapie.
ML137 Citez 4 causes possibles d'une infection du site opératoire.
ML138 Citez 9 éléments à prendre en compte pour définir l'état nutritionnel d'un patient.
ML139 Quels sont les 5 critères diagnostics de l'Algie vasculaire de la face selon l'International Headache Society ?
ML140 Citez 5 facteurs de risque possibles des troubles de l'alimentation.
ML141 Quelles sont les 7 méthodes d'évaluation de la fonction rénale ?
ML142 Citez 4 symptômes de l'AVF.
ML143 Quels sont les 7 objectifs de la consultation de diététique ?
ML144 Quels sont les 4 principaux symptômes du cancer de l'ovaire ?
ML145 Citez 3 complications de l'hypertrophie bénigne de la prostate.
ML146 Citez 7 symptômes de l'agonie.
ML147 Citez 5 symptômes possibles d'une mastite.
ML148 Quels sont les 4 médicaments qu'il est possible de prescrire dans le cadre d'une ostéoporose corticosonique ?
ML149 Quelles sont les trois principales complications induites par le diéthylstilbestrol ?
MRF150 Quelle est la bactérie causant la méningite à méningocoque ?
MRF151 Quelle est la cause de la trisomie 21 ?
MRF152 Comment l'algie vasculaire de la face peut-elle être traitée ?
MRD153 Que signifie le terme "chimiothérapie" ?
MRD154 Comment peut-on définir l'artériopathie oblitérante des membres inférieurs ?
MRD155 Comment l'IPS peut-il être défini ?
MRF156 Quel suivi proposer Ã  un patient atteint d'insuffisance rénale ?
MRF157 Quand une consultation diététique doit-elle être préconisée ?
MRD158 Quelle définition peut-on donner au bilan urodynamique ?
MRD159 Que signifie le sigle HBP ?
MRD160 De quelle façon la cholangite sclérosante peut-elle être définie ?
MRF161 De quelle façon l'engorgement mammaire peut-il être évité ?
MRF162 Comment un insuffisant rénal doit-il être suivi ?
MRF163 À quoi servent les stations de base dans les communications mobiles ?
MRD164 Comment l'amblyopie peut-elle être définie ?
MRF165 Quelle prise en charge proposer en cas d'aniridie ?
MRF166 L'aniridie est causée par quel germe ?
MRF167 De quelle façon détermine-t-on le poids corporel ?
MRF168 Quelle est la conséquence de la corticothérapie sur l'os ?
MRF169 À quel moment doit-on rechercher une insuffisance rénale ?
MRF170 À quelle date est apparu le diéthylstilbestrol en France ?
MRF171 Par qui la ventilation manuelle est-elle prise en charge ?
MRF172 Comment l'aniridie peut-elle être prise en charge ?
MRF173 À quel âge un neuroblastome peut-il apparaître ?
MRF174 L'agonie dure combien de temps en moyenne ?
MRF175 De quelle façon peut-on rechercher l'hématurie ?
MRF176 De quelle façon l'artériopathie oblitérante peut-elle être traitée ?
MRF177 Quelles sont les causes des fractures de l'ostéoporose cortisonique ?
MRF178 Une consultation en diététique doit être réalisée par qui ?
MRD179 Quelle est la définition de l'artériopathie oblitérante des membres inférieurs ?
MRF180 Quel a été le contexte d'utilisation du diéthylstilbestrol ?

MRD181 Que signifie le sigle AVF ?
MRF182 Quelle est la solution pour prévenir l'engorgement mammaire ?
MRF183 À quel moment la leucocyturie doit-elle être considérée comme pathologique ?
MRF184 Qu'engendre la corticothérapie sur l'os ?
MRF185 À quel chiffre s'élevait le coût global du cancer en France en 1994 pour l'Assurance maladie ?
MRF186 Quelles sont les causes de la trisomie 21 ?
MRF187 Quel est l'âge à partir duquel il est possible de dépister les troubles de l'acuité visuelle chez l'enfant ?
MRF188 Comment peut-on traiter l'algie vasculaire de la face ?
MRD189 Comment peut-on définir un trouble dépressif ?
MRF190 Dans quel endroit une consultation en diététique doit-elle se dérouler ?
MRF191 Par quel germe l'aniridie est-elle causée ?
MRF192 Comment l'artériopathie oblitérante des membres inférieurs peut-elle être traitée ?
MRF193 Quand procède-t-on au dosage de la créatininémie ?
MRF194 Dans quels cas le traitement au méthysergide est-il prescrit ?
MRD195 Comment définit-on la fluorose dentaire ?
MRF196 Comment peut-on distinguer la migraine de l'adulte de celle de l'enfant ?
MRF197 À quel moment doit-on procéder au dosage de la créatininémie ?
MRD198 Comment le syndrome de CACH peut-il être défini ?
MRF199 Quand le Livre Blanc sur la sécurité alimentaire a-t-il été adopté par la Commission Européenne ?
MRD200 Quelle est la définition du sarcome Kaposi ?

# Annexe 2 Corpus de questions utilisé pour évaluer le système Esculape[77]

MD1 Qu'est ce qu'une encéphalite Japonaise ?
MD2 Qu'est ce que la greffe de cornée ?
MD3 Comment définir le syndrome de Gitelman ?
MD4 Qu'est ce que la pyélonéphrite aigüe ?
MD5 Qu'est ce que la pyélonéphrite chronique ?
MD6 Qu'est ce que l'OMA ?
MD7 Comment définir la surdité ?
MD8 Quelle est la définition du pharynx ?
MD9 Qu'est ce que l'IRM ?
MD10 Que veut dire ODF ?
MD11 Qu'est ce que l'hyperparathyroïdie primitive ?
MD12 Comment définir l'ostéoporose ?
MD13 Quelle est la définition de l'anatomie ?
MD14 Comment l'hémovigilance peut-il être défini ?
MD15 Qu'est ce que la "méthionine" ?
MD16 De quelle façon définir la scintigraphie gastrique double phase ?
MD17 Comment peut-on définir l'hépatite C ?
MD18 Quelle est la signification de PAN ?
MD19 Que veut dire ANAES ?
MD20 Qu'est ce que "la maladie de Vaquez" ?
MT21 Quel est le traitement du rachitisme carentiel ?
MT22 Quel est le traitement administré en cas de diabète ?
MT23 Dans quel cas le traitement par corticoïdes est-il prescrit ?
MT24 Quel est le traitement proposé dans le cas d'une hémorragie digestive ?
MT25 Comment peut-on traiter les complications de l'angioplastie à la phase aiguë de l'infarctus du myocarde ?
MT26 Comment traiter le schwannome vestibulaire ?
MT27 Quel est le traitement des métastases cérébrales ?
MT28 Comment peut-on traiter les AOMI ?
MT29 Comment traiter une ascite infectée ?
MT30 Comment une otite moyenne aiguë peut-elle être traitée ?
MT31 Quel est le traitement de l'hypernatrémie ?
MT32 Comment traiter une poussée tensionnelle asymptomatique ?
MT33 Quelle thérapie proposer en cas d'atteinte bronchique ?
MT34 Quel traitement est proposé pour le choléra?
MT35 Quels sont les trois traitements possibles contre le cancer du colorectum ?
MT36 Comment peut-on traiter l'anémie ?
MT37 Comment traiter un cancer du rein ?

[77] Le type de la question précise l'entité médicale attendue : Traitement (MT), Médicament (MM), Symptôme (MS), Examen (ME) et enfin pour les questions définitoires (MD).

MT38 Quel traitement est utilisé pour soigner la pneumonie ?
MT39 Comment soigner la maladie de Parkinson ?
MT40 Par quel traitement le syndrome de Fanconi peut-il être traité ?
MM41 Quel est le médicament conseillé pour traiter une insuffisance cardiaque ?
MM42 Quel est le médicament à prescrire en cas de varicelle ?
MM43 Quel est le médicament qu'il est possible de prescrire dans le cadre d'une dysfonction érectile masculine ?
MM44 Quel est le médicament prescrit en cas de thrombose artérielle ?
MM45 Quel médicament administrer pour un œdème aigu du poumon ?
MM46 Comment guérir la vaginose bactérienne ?
MM47 Quel remède est indiqué pour la sclérose en plaques ?
MM48 Quel médicament prescrit-on contre la syphilis ?
MM49 Comment la maladie de Crohn peut-elle être soignée ?
MM50 Quel est le médicament à prescrire dans le cas d'une polyarthrite rhumatoïde ?
MM51 Quel est le choix du médicament pour une infection à toxoplasm ?
MM52 Quel est le choix du médicament pour une insuffisance rénale ?
MM53 Le médicament lamivudine est prescrit dans quel contexte ?
MM54 Comment l'épilepsie généralisée idiopathique peut-elle être soignée ?
MM55 Quel médicament est indiqué dans le cas d'une narcolepsie ?
MM56 Que peut-on prescrire contre les mycoses ?
MM57 Quel médicament proposer dans le cas d'une toxoplasmose cérébrale ?
MM58 Quel remède peut soigner l'arthrose ?
MM59 Dans quel contexte le collyre est-il utilisé ?
MM60 Que peut-on utiliser pour lutter contre le cancer de la prostate ?
MS61 Quels sont les symptômes de la maladie de Vaquez ?
MS62 Comment se manifeste une rhinite ?
MS63 Comment la bronchite chronique se manifeste-elle ?
MS64 Comment se caractérise les méningites chez le jeune enfant ?
MS65 Citez un symptôme de la lymphangiomatose pulmonaire ?
MS66 Quelles sont les manifestations de la neurofibromatose de type 2 ?
MS67 Quel est le principal symptôme de la CIVD ?
MS68 Quel symptôme accompagne une tumeur du médiastin ?
MS69 Quels sont les signes cliniques de l'encéphalopathie ?
MS70 Comment se révèle une pyélonéphrite aiguë ?
MS71 De quelle manière se manifeste le choléra ?
MS72 Citez trois symptômes de la maladie de Gélineau ?
MS73 Quand une leucémie chronique peut-elle être évoquée ?
MS74 De quelle façon se manifeste la maladie de Still ?
MS75 Quels sont les symptômes du syndrome de Reiter ?
MS76 Par quels symptômes se caractérise la fièvre jaune ?
MS77 Par quels signes cliniques se manifeste une rhinopharyngite ?
MS78 Comment se manifeste l'anémie ?
MS79 Quels sont les principaux signes cliniques du syndrome hépatorénal ?
MS80 À quels symptômes une infection urinaire compliquée peut-elle être associée ?
ME81 Quel est l'examen à réaliser pour confirmer une croissance tumorale ?
ME82 Comment peut-on examiner une hypertrophie ventriculaire ?
ME83 Comment peut-on conclure à une méningite ?
ME84 Comment peut-on diagnostiquer un lymphome ?
ME85 Quelle technique peut permettre de déceler une tumeur des tissus ?
ME86 Quel examen permet de dépister le cancer du sein ?

ME87 De quelle façon peut-on rechercher la tuberculose ?
ME88 Comment peut-on suspecter une cataracte unilatérale ?
ME89 Quel est l'examen à réaliser dans le cas d'une thrombose ?
ME90 Comment faire le bilan d'une affection de la thyroïde ?
ME91 Comment faire le diagnostic d'une ostéoarthrite ?
ME92 Comment pourrais-je conclure une ostéoporose ?
ME93 Quels sont les examens à réaliser en cas de suspicion d'ulcère ?
ME94 Comment rechercher un cancer colorectal ?
ME95 Quelle méthode permet de déterminer la présence d'un cancer des poumons ?
ME96 Comment peut-on suspecter une sarcoïdose ?
ME97 Comment certifier la détection d'une pneumonie ?
ME98 Quel examen diagnostique la rougeole ?
ME99 Quel examen permet de détecter l'arthrose ?
ME100 Quel bilan doit être effectué pour une pancréatite chronique ?

# Annexe 3 Exemples de règles de reconnaissance d'entités médicales

- La forme d'une règle :

**déclencheur : contexte_précédent : contexte_suivant : type_d'expression**

- **Reconnaissance des noms de maladie**

[maladie]**:**$L_NC [être$L_V] [$L_DET]**:**[*{1-20}]**:**DISEASE
[syndrome]**:**$L_NC [être$L_V] [$L_DET]**:**[*{1-20}]**:**DISEASE
[pathologie]**:**$L_NC [être$L_V] [$L_DET]**:**[*{1-20}]**:**DISEASE
maladie**::**de $L_NC**:**DISEASE
maladie**::**de $L_NP**:**DISEASE
syndrome**::**de $L_NC**:**DISEASE
cancer**::**(du|de la|des) $L_NC**:**DISEASE
[maladies]**::**[comme] [\:] [*{0-1}] $L_NC**:**DISEASE
[maladies]**::**[suivantes] [\:] [*{0-1}] $L_NC**:**DISEASE
syndrome**::**(du|de la|de l') $L_NC**:**DISEASE
@Disease**::**chronique**:**DISEASE

- **Reconnaissance des noms de traitement**

[traitement]**:**$L_NC [être$L_V] [$L_DET]**:**[*{1-20}]**:**TREATMENT
[traitement]**:**$L_NC [$L_PONCTU] [être$L_V] [$L_DET] [*{0-1}]**::**TREATMENT
[traitements]**::**[comme] [\:] [*{0-1}] $L_NC**:**TREATMENT
[thérapie]**:**$L_NC [*{0-3}]**:**[pour] [traiter$L_V]**:**TREATMENT
[traitement]**:**$L_NC [*{1-3}]**:**[servir$L_V] [à] [traiter$L_V]**:**TREATMENT
[traitement]**:**$L_NC [*{1-3}] [dans le]**:**[*{1-3)}]**:**TREATMENT
[traitement]**::**[pour] [(traiter$L_V|guérir$L_V] [*{0-5}] [$L_NC] [être$L_V] [*{0-1}] $L_NC**:** TREATMENT
[ou]**:**[@Treatment] [\(]**:**$L_NC**:**TREATMENT
traitement**::**au $L_NC**:**TREATMENT
traitement**::**par $L_NC**:**TREATMENT

- **Reconnaissance des noms de médicament**

[médicament]**:**$L_NC [être$L_V] [$L_DET]**:**[*{0-20}]**:**DRUG

[médicament]:$L_NP [$L_PONCTU] [être$L_V] [$L_DET] [*{0-1}]:[*{1-20}]:DRUG
[médicaments]::[comme] [\:] [*{0-1}] $L_NC:DRUG
[médicament]:$L_NC [*{0-5}]:[pour] [traiter$L_V] [*{1-10}]:DRUG
[traiter]:$L_NC [\,] [médicament] [pour]:[*{1-10}]:DRUG
[médicament]:$L_NC [*{0-5}]:[servir$L_V] [à ] [traiter$L_V]:DRUG
[médicament]:$L_NC [\,]:[*{1-20}]:DRUG
[médicaments]::[$L_NC] [\(] $L_NC:DRUG
T_A1:[traitement] [par]:[*{0-1}]:DRUG[78]
T_A1::[est] [utiliser$L_V] [pour] [traiter$L_V]:DRUG

- **Reconnaissance des noms de symptôme**

[symptôme]:$L_NC [être$L_V] [$L_DET]:[*{1-10}]:SYMPTOM
[clinique]:$L_NC [être$L_V] [$L_DET] [signe]:[*{1-10}]:SYMPTOM
[cliniques]:[signes]:[suivants] [\:] [*{0-1}] $L_NC:SYMPTOM
[signes]::[suivants] [\:] [*{0-1}] $L_NC:SYMPTOM
[signe]:$L_NC [être$L_V] [$L_DET]:[*{1-20}]:SYMPTOM
[symptome]::[$L_NC] [\(] $L_NC:SYMPTOM
trouble::(de la|du) $L_NC:SYMPTOM
@Symptom::(de la|de l') @Disease:SYMPTOM
@Symptom::$L_NC [\(]:SYMPTOM
@Symptom::(0-2) @Anatomy:SYMPTOM

- **Reconnaissance des noms des examens cliniques**

[examen]:$L_NC [être$L_V] [$L_DET]:[*{1-20}]:EXAM
[examen]:$L_NC [$L_PONCTU] [être$L_V] [$L_DET] [*{0-1}]:[*{1-20}]:EXAM
[examens]::[suivants] [\:] [*{0-1}] $L_NC:EXAM
[tests]::[suivants] [\:] [*{0-1}] $L_NC:EXAM
[examen]:$L_NC [*{0-5}]:[pour] [détecter$L_V]:EXAM
[examen]:$L_NC [*{1-3}]:[servir$L_V] [à] [détecter$L_V]:EXAM
[examen]:$L_NC [\,]:[*{1-20}]:EXAM
[test]:$L_NC [\,]:[*{1-20}]:EXAM
[examen]::[$L_NC] [\(] $L_NC:EXAM
@Exam::$L_NC [\(]:EXAM

[78] T_A1 signifie que la première lettre du mot est en majuscule.

# Annexe 4 Règles de typage des questions médicales

- **Réponse attendue : « Traitement »**

[@Quel] : : [être$L_V] [$L_DET] [*{0-1}] [@Traitement] [@De] *{1-30} [\?] : ESCULAPE :F_Traite_Traitement

[@Quel] : : [être$L_V] [$L_DET] [*{1-3}] [(pour)] [(prévenir)] *{1-30} [\?] : ESCULAPE :F_Traite_Traitement

[@Quel] : : [*{0-1}] [@Traitement] [$L_V] *{1-30} [\?] :ESCULAPE :F_Traite_Traitement

[Par] : : [@quel] [@Traitement] [$L_DET] *{1-30} [\?] : ESCULAPE:F_Traite_Traitement

[@Quel] : : [être$L_V] [$L_DET] [*{0-2}] [@Manière] [*{1-5}] [traiter$L_V|éviter$L_V] *{1-30} [\?] : ESCULAPE :F_Traite_Traitement

[De] : : [@quel] [@Manière] *{1-20} (être) (traiter$L_V) [\?] : ESCULAPE :F_Traite_Traitement

[De] : : [@quel] [@Manière] [traiter] *{1-30} [\?] : ESCULAPE :F_Traite_Traitement

[@Quel] : : [être$L_V] [$L_DET] [(traitement|traitements)] [contre-indiqué] *{1-30} [\?] : ESCULAPE :F_Deconseiller_Traitement

[Dans] : : [@quel] [cas] [$L_DET] [@Traitement] *{1-10} [être$L_V] [contre-indiqué] [\?] : ESCULAPE :F_Deconseiller_Maladie

- **Réponse attendue : « Symptôme »**

[@Quel] : : [être$L_V] [$L_DET] [*{0-5}] [@Symptome] *{1-30} [\?] : ESCULAPE :F_Caracterise_Symptome

[@Quel] : : [@Symptome] *{1-30} [\?] : ESCULAPE :F_Caracterise_Symptome

[Comment] : : [se caractériser$L_V] [*{0-2}] [@Maladie] *{1-20} [\?] : ESCULAPE : F_Caracterise_Symptome

[@Quel] : : [*{0-5}] [@Symptome] *{1-30} [\?] : ESCULAPE :F_Caracterise_Symptome

[Par] : : [@quel] [@Symptome] *{1-30} [\?] : ESCULAPE : F_Caracterise_Symptome

[@Quel] : : [être$L_V] [$L_DET] [(effets secondaires)] *{1-30} [\?] : ESCULAPE :F_Deconseiller_Symptome

- **Réponse attendue : « Médicament »**

[@Quel] : : [être$L_V] [$L_DET] [*{0-5}] [@Médicament] *{1-30} [\?] : ESCULAPE :F_Soigne_Medicament

[@Quel] : : [sont] [$L_DET] [@Médicament] *{1-30} [\?] : ESCULAPE :L_Soigne_Medicament

[@Quel] : : [@Médicament] *{1-30} [\?] : ESCULAPE :F_Soigne_Medicament

[Par] : : [@quel] [@Médicament] *{1-30} [\?] : ESCULAPE :F_Soigne_Medicament

[@quel] : [Dans] : [cas] [*{0-2}] [@Médicament] [être$L_V] [$L_V] *{1-30} [\?] : ESCULAPE :F_Soigne_Maladie

[@quel] : [Dans] : [cas] [*{0-2}] [@Médicament] [$L_V] *{1-30} [\?] : ESCULAPE :F_Soigne_Maladie

[@quel] : [Dans] : [cas] [*{0-2}] [@Médicament] *{1-30} [\?] : ESCULAPE :F_Soigne_Maladie

- **Réponse attendue : « Examen »**

[@Quel] : : [*{0-2}] [(examen|test)] *{1-30} [\?] : ESCULAPE :F_Detecte_Examen

[@Quel] : : [être$L_V] [$L_DET] [(examen|test)] [à] *{1-30} [\?] : ESCULAPE :F_Detecte_Examen

[@Quel] : : [être$L_V] [$L_DET] [*{0-3}] [(examen|test)] *{1-30} [\?] : ESCULAPE :F_Detecte_Examen

[@Quel] : : [être$L_V] [*{0-3}] [(examen|test)] *{1-30} [\?] : ESCULAPE :F_Detecte_Examen

[Par] : : [@quel] [(examen|test)] *{1-30} [\?] : ESCULAPE :F_Detecte_Examen

[Dans] : : [@quel] [(cas)] [*{0-2}] [$L_DET] [(examen|test)] *{1-30} [\?] : ESCULAPE :F_Detecte_Maladie

[Comment] : : [rechercher$L_V|détecter$L_V] [*{0-2}] [@Maladie] *{1-20} [\?] : ESCULAPE :F_Detecte_Examen

- **Réponse attendue : « Définition »**

[Qu'] : : [est] [$L_DET] [@Que] *{1-30} [\?] : ESCULAPE :D_Definition_NONE[79]

[Que] : : [@Signifier] [$L_DET] [@sigle] *{1-10} [\?] : ESCULAPE :D_Definition_NONE

[@Pronoms_interrogatifs] : : [@Signifier] [$L_DET] [symbole] *{1-10} [\?] : ESCULAPE :D_Definition_NONE

[@Pronoms_interrogatifs] : : [@Signifier] *{1-10} [\?] : ESCULAPE :D_Definition_NONE

[@Pronoms_interrogatifs] : : [(vouloir$L_V dire)] *{1-10} [\?] : ESCULAPE :D_Definition_NONE

[De] : : [@quel] [@Manière] [définir] *{1-30} [\?] : ESCULAPE :F_Definition_NONE

[@Quel] : : [est] [$L_NC] [@Definition] [@De] *{1-30} [\?] : ESCULAPE :D_Definition_NONE

[79] None : spécifie que la réponse attendue n'est pas une entité nommée.

# Annexe 5 Exemples de patrons lexico-syntaxiques appris automatiquement

**- La relation « Traite » (X : Maladie, Y : Traitement)**

X ( Y

X par Y

Y utilisé pour traiter le X

Y est le traitement de choix pour la X

Y est souvent utilisée en pratique courante dans le traitement de X

Y recommandés dans le traitement des X

X peut être prévenue efficacement par un traitement Y

X <*s*> traitée par <*s*> Y

Y <*s*> contre la X

X <*g*> être <*s*> traités par Y

Y <*g*> L_CONJ_COORD L_PREP_GENERAL la X

Y L_NC_GEN pour traiter <*g*> X

Y L_PREP_GENERAL le traitement <*s*> du X

X peut être <*g*> <*s*> par un <*s*> Y

X <*g*> après Y

**- La relation « Soigne » (X : Maladie, Y : Médicament)**

X par Y

X , Y

X ( Y

Y est indiqué dans le traitement de fond de la X

Y pour contrôler le X

X est suspectée , le traitement par Y

Y pour un X

Y chez un patient ayant une X

Y est actuellement indiquée dans la prévention secondaire de l' X

X peut s'effectuer par l’administration de Y

Y <*s*> pour traiter <*g*> X

X <*s*> traité par l' Y

Y <*s*> pour L_VERBE_PRINCIPAL_INFINITIF le X

X L_NC_GEN à la Y

Y <*s*> contre le X

**- La relation « Détecte » (X : Maladie, Y : Examen)**

X ( Y

X , une Y

Y révèle une X

Y pouvoir révéler une X

Y est recommandée comme examen essentiel dans l'évaluation de l' X

Y est de confirmer une X

Y montre une X

Y qui signe le diagnostic de la X

Y dans le diagnostic pratique des X

X , il est recommandé de faire une Y

Y <*g*> <*s*> permet de L_VERBE_PRINCIPAL_INFINITIF le X

Y ( <*s*> X

X , <*g*> Y

Y pour exclure <*s*> X

X L_VERBE_PRINCIPAL_INDICATIF <*s*> <*g*> Y

**- La relation « Caractérise » (X : Maladie, Y : Symptôme)**

X ( Y

X ( apparition soudaine de Y

X par une Y

X sont le Y

X caractérisée par des Y

X ( absence de Y

X , avec Y

Y vient compliquer une X

Y révélatrice d’un X

Y pouvoir être une manifestation d'une X

X être <*g*> Y

X être un L_NC_GEN <*g*> <*s*> des Y

X débute <*s*> par un <*s*> Y

X avec <*g*> Y

Y <*s*> peuvent évoluer L_PREP_GENERAL <*s*> X

**- La relation « Définition » (X : focus de la question, Y : Réponse)**

Y ( X

Y , X

X ( Y

X , un Y

X , le Y

Y <*g*> X

X <*g*> Y

X L_VERBE_PRINCIPAL_INDICATIF L_DET_ARTICLE_INDEF Y

X être <*g*> <*s*> Y

X être L_DET_ARTICLE_INDEF Y

Y , <*g*> X

Y <*g*> L_NC_GEN X

Y ( <*s*> X

X , <*s*> Y

X <*g*> comme un Y