

Indexation et interrogation automatiques de textes médicaux : application à la base A.D.M.

SEKA L.-P.^a, POULIQUEN B.^a, LE BEUX P.^a

^a *Laboratoire d'Informatique Médicale, faculté de Médecine, Université de Rennes I,
Avenue du Professeur Léon Bernard, 35043 Rennes cedex, France*

Résumé : L'ADM-INDEX est un système utilisant les méthodes linguistiques (morphologie, syntaxe et sémantique) pour l'indexation et l'interrogation de comptes-rendus médicaux. Sa base de connaissances est une restructuration du système A.D.M. qui est une vaste base de connaissances médicales développée au Centre Hospitalier Régional de Rennes depuis 1972 et qui a été le premier videotex professionnel d'aide au diagnostic médical en France. Dans cet article, nous présentons d'une part la base de connaissances de l'ADM-INDEX qui est constituée d'un dictionnaire, d'un ensemble de règles et d'un thésaurus, et d'autre part, le processus d'indexation et de recherche d'informations. Un prototype du système est consultable sur World Wide Web, qui est le système multimédia le plus utilisé sur Internet.

Mots clés: Indexation et interrogation automatiques, Base de connaissances, Concept, Thésaurus et nomenclature, compte rendu médical.

Abstract : ADM-INDEX is a system for indexing and retrieval of Patients Discharge Summaries (PDSs) by using linguistic methods (morphologic, syntactic and semantic processing). The ADM-INDEX knowledge base is a restructuring of a diagnostic aid knowledge base (ADM) in order to allow the linguistic analysis of medical texts. The ADM system is a comprehensive medical knowledge base which has been developed since 1972 at the University Hospital of Rennes and which has been the first professional videotex medical diagnostic aid in France. ADM-INDEX, after linguistic analysis, build the index table with thesaurus wording, medical words, concepts and phrases, unknown words contained in each PDS. The benefit of using those different elements is to improve retrieval. Although our system is constructed with the ADM dictionnary, it can be easily applied to other medical nomenclature or thesaurus. In this paper, we present on the one hand the ADM-INDEX knowledge base which is constituted by rules, a dictionary and a thesaurus, and on the other hand, the process of indexing and retrieval information. A prototype of the system is available on World Wide Web, which is now the most used multimedia information system on the Internet.

1. Introduction

Le domaine médical est très vaste, très riche et très complexe sémantiquement. Un grand nombre de ses activités (radiologies, coronarographie, etc ..) se traduit par la production de comptes rendus médicaux rédigés en langage naturel.

Les comptes rendus médicaux décrivent l'état de santé des patients. Ce sont des documents importants dans la mesure où ils servent principalement [1] d'outil de suivi du malade et secondairement d'outil de synthèse et d'auto enseignement, d'outil médico-légal, de communication, de recherche clinique, de gestion hospitalière, d'études épidémiologiques, d'évaluation de la qualité de soins et d'enseignement.

Vu l'abondance des informations que contient chaque compte rendu, il s'avère nécessaire de pouvoir les stocker et les retrouver de manière sélective et pertinente par l'intermédiaire d'un système rapide, efficace et ergonomique. Ce système apportera une aide considérable aux médecins dans l'accomplissement de leur tâche.

La mise en place d'un tel système suppose non seulement la résolution des problèmes liés au discours médical (élipses, paraphrases, ambiguïtés, etc) [2, 3, 4] car celui-ci doit sélectionner les concepts pertinents permettant de représenter le contenu des documents

mais également l'utilisation d'une nomenclature¹ médicale. Il serait donc intéressant de pouvoir utiliser les bases de connaissances des systèmes médicaux tels que INTERNIST [5] (qui comprend 500 maladies) et son dérivé QMR [6], RECONSIDER [7] (qui comprend 3.000 maladies), l'A.D.M. [8,9] (qui comprend 12.000 maladies).

Nous avons opté pour la base A.D.M. (Aide au Diagnostic Médical) car d'une part elle a été développée au Laboratoire d'Informatique Médicale de la Faculté de Médecine de l'Université de Rennes I et d'autre part, elle est très riche qualitativement et quantitativement (12.000 maladies, syndromes, effets indésirables et formes cliniques décrits par une nomenclature de 130.000 entités, un dictionnaire de 60.000 entrées). Malgré cette richesse, la base A.D.M. se prête mal à l'analyse des textes médicaux (manque d'informations syntaxiques, sémantiques, etc) rédigés en langage naturel.

Après une présentation de la base A.D.M. et le constat effectué de son incapacité à traiter efficacement les documents médicaux, nous présenterons la base de connaissances du système ADM-INDEX qui est une restructuration de la base A.D.M. afin de l'adapter à l'analyse linguistique de documents. Nous montrerons comment nous détectons les concepts et/ou expressions médicales, comment nous constituons les index et comment nous nous en servons pour la recherche d'informations. Nous terminerons par la description de l'implantation et l'évaluation du système.

2. La base de connaissances du système A.D.M.

La base du système A.D.M. a été développée dans le but de fournir une aide documentaire et diagnostique aux médecins en utilisant des méthodes de recherches combinatoires. Elle est constituée d'une nomenclature de libellés, d'un dictionnaire ainsi que des descriptions de maladies. Nous présentons en détail les deux premiers éléments.

2.1 La nomenclature des libellés A.D.M.

La nomenclature A.D.M. contient environ 130.000 entités. Une entité A.D.M. est un terme ayant un sens médical et servant à décrire, entre autres, les pathologies. Ces entités concernent tous les domaines de la sémiologie ainsi que les professions exposées et les pays d'endémie. Une entité appartient à l'un des champs sémantiques suivants :

P = Physiopathologie N = Nosologie E = Examen L = Anatomie G = Pronostic
 S = Signes M = Maladies D = Médicaments T = Evolution Y = Syndromes
 C = Commentaires A = Environnement / Terrain F = Formes cliniques

Les liens existants entre les entités sont des liens de type Père-fils (EST_UN).

Cette large nomenclature à l'avantage d'être finement constituée. Le niveau de précision est très élevé. Un exemple est donné dans le tableau suivant :

S00197	Rachialgie	S07385	Douleur rachis dorsal aigue
S07384	Dorsalgie	S40344	Douleur rachis lombaire aigue
S17028	Douleur rachis aigue	S34943	Douleur rachis lombaire médiane
S00195	Douleur rachis lombaire	S71424	Douleur rachis lombaire très aigue
S15708	Douleur rachis permanente		

Tableau n°1 : Extrait de la terminologie A.D.M.

¹. La nomenclature est définie comme l'ensemble des termes employés dans une science, sans définition explicite de ces termes, mais méthodiquement classés.

2.2 Le dictionnaire A.D.M.

Le dictionnaire de l'A.D.M. (environ 60.000 mots) est un dictionnaire particulier. Il ne donne ni la définition des mots ni leur contexte d'utilisation. Il permet de traduire les phrases des médecins ou étudiants utilisateurs de l'application, en phrase A.D.M., c'est-à-dire en langage quasi-naturel mais "laconique".

Le dictionnaire contient tous les mots des entités A.D.M. Les mots sont regroupés par famille. Le père d'une famille de mots est appelé synonyme de référence ou mot de référence. Les liens existants entre les différents mots d'une famille sont des liens de synonymie ou de flexion. Le dictionnaire A.D.M. compte environ 24.000 familles. Au sein d'une famille, on distingue deux catégories de mots : les mots simples et les mots complexes qui sont subdivisés en mots composés et mots associés.

- Mots simples : Ils sont composés d'un mot unique (chaîne de caractères). Parmi les mots simples, on distingue les mots nuls ou mots vides de sens (le, la, les) et les mots ignorables (les adverbess essentiellement). Les mots simples sont au nombre de 45.000 environ.

- Mots complexes : Ils sont composés par un ensemble de mots simples. Parmi les mots complexes, on distingue deux sous catégories :

- Le mot composé a été créé afin d'empêcher de dissocier les mots le constituant. Il permet d'exprimer un lien très fort entre ses constituants d'une part et d'autre part, de mieux gérer les synonymies entre par exemple FIEVRE JAUNE (terme multimots) et le concept AMARIL (mot simple). Le mot composé est très rigide car lors de sa reconnaissance, il n'accepte ni synonyme (synonyme des mots le constituant), ni aucune permutation de l'ordre de ses constituants. Ceci dans le but d'éviter de reconnaître de faux termes. Les mots composés sont étiquetés dans le dictionnaire par le symbole "\$". Ils sont actuellement au nombre de 900 environ.

- Mots associés. Ils sont similaires aux mots composés mais ils sont beaucoup plus souples que ces derniers. En effet, les mots associés permettent de tenir compte des synonymies entre les mots composants (Ex : *Absence congénitale* et *Déficit congénital* sont similaires sinon sémantiquement identiques). Les mots associés sont étiquetés dans le dictionnaire par le symbole "&". Ils sont actuellement au nombre de 750 environ.

Le dictionnaire A.D.M. tel qu'il est structuré ne peut pas être utilisé dans un système d'indexation et d'interrogation basé sur les techniques du traitement du langage naturel. En effet, dans un tel système, le dictionnaire joue un rôle primordial car c'est lui qui contient les informations morphologiques, syntaxiques, sémantiques et mêmes pragmatiques. Ces informations sont nécessaires aux différentes phases d'analyse pour la reconnaissances exactes des expressions. En outre, il doit servir de tremplin pour l'inférence et la déduction dans la mesure où c'est lui qui fournira les éléments nécessaires pour le déclenchement des processus de déduction et/ou d'inférence. L'inférence et la déduction sont deux éléments importants dans un système d'indexation utilisant les techniques linguistiques [10].

Or, du point de vue linguistique, la base ADM possède des lacunes à divers niveaux :

- le dictionnaire : il ne comporte aucune information syntaxique ni sémantique. Les différents contextes d'utilisation d'un mot ne sont pris en compte car tels que ceux-ci ont été définis, ils ne peuvent appartenir qu'à une seule famille à la fois (avoir qu'un seul sens). En outre, la définition des mots complexes est très large et conduit parfois à de mauvaises reconnaissances de termes.

- la nomenclature : quoique très fine et très précise dans la hiérarchisation, ne comporte qu'un seul type de relation, la relation générique **est_un**. La relation partitive (**composé_de**) qui est aussi importante que la relation **est_un** est complètement absente.

Ces lacunes ne mettent pas en cause les performances du système A.D.M. car celui-ci remplit pleinement les tâches pour lesquelles il a été construit.

En revanche, si la base doit être utilisée dans un système d'indexation automatique de textes en langage naturel, les lacunes énumérées doivent impérativement être comblées car leur présence est un frein énorme à l'élaboration de ce système. La base ADM-INDEX est le résultat de la restructuration de la base A.D.M.

3. Description de la base de connaissances

La base de connaissances de ADM-INDEX est constituée d'un dictionnaire, d'un ensemble de règles ainsi que d'un thésaurus. Cette façon de constituer la base nous permet de prendre en compte les caractéristiques essentielles des réseaux sémantiques [11, 12], les règles de production [13] et les schémas (frame) de Minsky [14].

3.1 Le dictionnaire

Le dictionnaire ADM-INDEX a essentiellement pour but de permettre la détection de concepts médicaux ainsi que d'expressions médicales quelque soit la forme sous laquelle ils apparaissent dans le texte. Cela suppose la résolution de nombreux problèmes linguistiques dont regorge le langage médical [15, 16].

Nous conservons la constitution du dictionnaire sous forme de famille ainsi que les différents types de termes (mot simple, mot composé et mot associé). Nous ajoutons des éléments de nature syntaxique et sémantique aux différents constituants du dictionnaire afin de le rendre plus qualitatif.

Définition des constituants

Nous distinguons dans le dictionnaire trois types de constituants à savoir : les concepts¹, les expressions (termes multimots) et les mots simples. Chaque constituant sera défini à partir de l'ensemble d'éléments suivant :

. CODE_SIG	: c'est le code associé au constituant
. CATEG	: indique si le constituant est un mot, un concept ou une expression
. TYPE	: indique si le constituant est un mot nul, ignorable ou non
. [MEDIC]*2	: indique si le constituant est un terme médical ou non
. [CATEGRAM]*	: catégorie grammaticale du constituant
. [CAT_SEM]*	: Catégories sémantiques du constituant
. [PERHIERAG]*	: Pères hiérarchiques selon le lien générique
. [PERHIERAP]*	: Pères hiérarchiques selon le lien partitif
. [MOT]	: libellé du constituant
. [LOC]*	: localisation du constituant par rapport aux autres si nécessaire
. [COD_COMP]*	: liste triée des CODE_SIG des mots d'une expression
. [CODE_RANG]*	: rang des mots d'une expression
. [PREC]*	: catégorie de concepts précisés
. [DEF]*	: définition du constituant si nécessaire
. [OPP]*	: concepts ou mots auxquels il est opposé
. [IMP]*	: sert à mentionner les concepts impliqués
. [CAUSE]*	: sert à mentionner les causes qui le génèrent

Nous utilisons également les opérateurs /// et // pour matérialiser les différents sens ou cas possibles. Le symbole /// permet de matérialiser les différents cas qui s'excluent mutuellement et le symbole // permet de matérialiser à l'intérieur de chaque cas, les différentes possibilités. Ils permettent de prendre en compte le cas des différents contextes

1. Un concept est un terme scientifique ou linguistique dont la définition est bien spécifiée et qui représente une classe d'objets ou d'idées.

2. * signifie que la rubrique est optionnelle

d'utilisation d'un concept, d'une expression ou simplement d'un mot. Ce qui donne la possibilité à un concept d'appartenir à plusieurs familles à la fois, bien qu'il soit enregistré dans une seule.

Leur utilisation permet d'éviter d'avoir plusieurs entrées pour un même concept dans le dictionnaire si celui-ci possède plusieurs sens. Prenons par exemple le concept *Sécrétaire* qui possède trois sens :

- *Sécrétaire 1* : assistant du patron,
- *Sécrétaire 2* : meuble à tiroir
- *Sécrétaire 3* : synonyme de serpenteaire : animal se nourrissant de serpents

Ce concept sera une seule entrée dans le dictionnaire ADM-INDEX. Il sera représenté de la façon suivante :

```
. CODE_SIG      : 27139000
. CATEG         : Concept
. TYPE          : non ignorable
. CATEGRAM      : Substantif
. CAT_SEM       : Etre_humain 1/// Animal 2/// Objet 3
. PERHIERA      : Etre_humain 1/// Reptil 2 /// Meuble 3
. MOT           : Secrétaire
. LOC           : Administration 1
. DEF           : assistant du patron 1/// voir: serpenteaire 2/// armoire à tiroir 3
```

La traduction de cette représentation est la suivante : si *secrétaire* est employé dans le cadre d'une administration alors il s'agit d'un être humain qui est l'assistant du patron; s'il est employé dans un contexte animal, alors il est synonyme (DEF:voir) de serpenteaire; s'il est employé dans un contexte de meuble, alors il s'agit d'une armoire à tiroir.

Nous utilisons un système de chiffres qui permet de bien lier les différentes caractéristiques au concept selon la catégorie sémantique considérée.

Les éléments du dictionnaire permettent d'attribuer à un terme donné du dictionnaire, des informations morphologiques, syntaxiques et sémantiques.

• **Informations Morphologiques** : ces informations sont prises en compte au sein des familles car nous savons qu'une famille contient entre autres un mot ainsi que toutes ses flexions. On aura donc toutes les formes du mot dans sa famille. C'est le **CODE_SIG** qui permet de lier un mot à une famille.

Exemple :

```
257490A0 PRANDIAL      257490AB PRANDIALES
257490AA PRANDIALE    257490AC PRANDIAUX
```

• **Informations syntaxiques** : ADM-INDEX permet d'indiquer (**CATEGRAM**) la catégorie grammaticale de chaque mot ou concept. Les différentes catégories grammaticales présentes dans le dictionnaire sont : Substantif, Verbe, Adverbe, Adjectif, Préposition, Préfixe, Déterminant.

Exemple : Gauche

```
. CODE_SIG      : 12660000      . MOT       : Gauche
. CATEG         : Mot simple     . PREC      : Organe // Région anatomique
. TYPE          : non ignorable  . OPP       : Droite
. CATEGRAM      : Adjectif
```

Gauche est un mot qui n'est pas employé seul. C'est un préciseur de localisation (organes ou régions anatomiques). Il est opposé à *Droite*. En médecine, le fait d'utiliser

Gauche comme un préciseur nous impose de le lier soit à un organe, soit à une région anatomique.

◦ **Informations sémantiques** : Ces informations sont prises en compte par l'utilisation des mots composés et associés, par les relations existant entre les différents concepts (CAT_SEM, CAUSE ...), par la définition explicite de certains concepts (DEF), par l'utilisation des opérateurs /// et //.

Exemple : *Coronaire*.

. CODE_SIG	: 06873000	. CATEGRAM	: Adjectif 1/// Substantif 2
. CATEG	: Mot simple	. MOT	: Coronaire
. TYPE	: non ignorable	. DEF	: Voir: Artère coronaire 2

Si *Coronaire* est employé en tant que adjectif, alors c'est tout simplement l'adjectif *Coronaire*. Si *Coronaire* est employé en tant que substantif, alors il s'agit de l'Artère coronaire. Il faudra donc dans ce cas prendre en compte les caractéristiques de *Artère coronaire* décrites dans le dictionnaire.

Cette façon de représenter les mots ainsi que les concepts permettra de mieux les reconnaître dans les textes. En plus, l'utilisation de syntagmes (mots composés et mots associés) permet de bien préciser l'idée ou la notion exprimée, donc de réduire les cas de polysémies. En outre, la structuration du dictionnaire en familles permet d'avoir toutes les formes fléchies d'un mot. Nous n'avons pas besoin de faire un traitement spécifique pour reconnaître les mots (traitements sur les terminaisons). Cela a pour avantage d'accélérer le processus de traitement.

3.2 Les règles de détection des concepts et/ou des expressions

C'est par le biais d'un ensemble de règles ainsi qu'une table d'acceptabilité des transformations que nous allons détecter les différents termes.

Notre base ne contient pas une multitude de règles ou de transformations. Cinq règles et trois transformations permettant d'effectuer les traitements adéquats afin de détecter les bons termes. Nous présentons ici la règle de détection des mots composés comme définie précédemment. Cette règle est la suivante :

$\forall M_1 M_2 \dots M_n \quad M_1 M_2 \dots M_n \in {}^*C \text{ si } M_1, M_2, \dots, M_n \in PH$
avec $M_1 \in F^*_{M_1}, M_2 \in F^*_{M_2}, \dots, M_n \in F^*_{M_n}$
et $\text{rang}_{M_n} - \text{rang}_{M_i} = \text{pos}_{M_n} - \text{pos}_{M_i} \quad i = 1, n-1$

avec

*C = ensemble des mots composés de la phrase, $F^*_{M_i} = \{ M_i \mid \text{flexions } M_i \}$,

PH = ensemble des mots de la phrase, rang_{M_i} = rang du mot M_i dans la phrase

L'application de cette règle permettra de reconnaître le mot composé *Fièvre jaune* dans la phrase *Virus de la Fièvre jaune* et non dans celle-ci *Hyperthermie jaune*. Bien qu'hyperthermie soit synonyme de Fièvre, on ne dit jamais "Hyperthermie jaune".

Nos règles peuvent subir certaines transformations. Nous en avons défini essentiellement trois, à savoir :

◦ La permutation qui consiste à échanger l'ordre des mots constituant le terme

Exemple: MOT1 MOT2 \Rightarrow MOT2 MOT1

- La réduction qui consiste à effectuer des transformations globales qui suppriment ou ajoutent des mots sans changer l'ordre.

Exemple: Atrophie musculaire => Atrophie du muscle (nominalisation)

- La substitution qui permet de prendre en compte essentiellement les synonymies afin de retrouver la même idée exprimée sous une autre forme.

La définition et l'utilisation de ces trois transformations se justifient par le fait que les constituants du dictionnaire se présentent essentiellement sous leur forme minimale. C'est uniquement soit une permutation, soit une réduction, soit une substitution, soit une de leurs combinaisons appliquée à la phrase analysée qui permettra de détecter les termes.

Il faut toutefois souligner que les règles n'acceptent pas toutes, les différentes transformations.

TRANSFORMATIONS RÈGLES	REDUCTION			PERMUTA- TION	SUBSTITU- TION
	Effacement		Autres formes		
	mots nuls	mots ignorables			
Règle des mots composés	+	-	-	-	-

Tableau n°2 : Table d'acceptabilité des transformations par la règle des mots composés

Ce tableau montre que la règle de reconnaissance des mots composés n'accepte que la réduction des mots nuls.

L'utilisation des règles et des transformations permet d'obtenir une même entrée dans le dictionnaire pour les termes suivants:

Atrophie du muscle Muscle atrophié Muscle très atrophié Atrophie musculaire
 Atrophie des muscles Muscles atrophiés Muscles très atrophiés etc

2.3 Le thésaurus

Les libellés du thésaurus ADM-INDEX sont organisés hiérarchiquement. Cette organisation s'est faite essentiellement sur la base des relations générique **Est_un** et partitive **Est_composé_de**.

La mise en place de cette hiérarchie de libellés s'est faite en se basant sur la définition des différents libellés. La définition est formée d'une partie *Genre* (Génus) et d'une partie *Différence* (Differentiæ). Le Genus est l'hyperonyme auquel sera rattaché le libellé dans la hiérarchie et la Differentiæ permet de différencier le libellé de son hyperonyme. La construction de taxonomies basée sur ce principe et déjà utilisée par d'autres chercheurs [17, 18] est fiable.

La hiérarchie de libellés est très importante dans un processus d'indexation dans la mesure où elle permet non seulement aux fils d'hériter des propriétés de leurs pères mais également de préférer un libellé plus fin (en l'occurrence un libellé feuille) à un libellé plus large (en l'occurrence un libellé racine) [19, 20]. En outre, au sein du thésaurus, la similarité entre termes est un élément nécessaire dans un système comme le notre car elle permet de

réduire les risques de silence¹ lors d'une recherche d'informations. Elle est prise en compte par le biais du signe //.

Exemple : M00209 Maladie du sommeil // Trypanosomiase africaine

Cet ensemble de termes similaires est appelé au sein de l'ADM-INDEX une **Entité**.

C'est donc sur cette base de connaissances que va s'appuyer le système pour analyser et indexer les différents textes médicaux.

4. Le processus d'indexation et d'interrogation

Indexation et Interrogation sont deux termes qui vont de pair. L'indexation recense les éléments sensés représenter le contenu des documents et l'interrogation se sert de ces éléments pour la recherche des documents dans la base.

4.1 L'indexation

Elle s'articule autour d'un certain nombre de modules. La présentation de ces modules se fera selon leur ordre d'enchaînement.

◦ Le module de **Découpage** et de **Reconnaissance des mots**. Il permet de découper le texte en phrases. Ceci a pour but de diviser le problème global (texte) en problèmes plus simples (phrases) à traiter. Ce module va utiliser essentiellement les signes de ponctuation ainsi que le dictionnaire ADM-INDEX. Nous considérons ici que le texte pris en entrée est bien ponctué. Chaque phrase extraite sera ensuite découpée en mots afin de procéder à la reconnaissance de chacun d'eux. Lors de cette phase, on procède éventuellement à des corrections orthographiques des mots inconnus par la base. Le système de corrections orthographiques prend en charge les cas suivants : un caractère manquant, un caractère supplémentaire, l'emploi erroné de certaines lettres pour d'autres tels que **F -> PH, T -> TH, Z -> CZ, K -> CK, R -> RH, CH -> SCH.**

A chacun des mots est attribué ses informations syntaxiques et sémantiques. Exemple de décomposition de la phrase : *L'échographie montre une absence congénitale du doigt.*

Phrase	Information syntaxique	Information sémantique
L'	Article	mot vide
échographie	Substantif	concept (échographie)
montre	verbe	montrer, marqueur fort
une	Article	mot vide
absence	Substantif	opp. présence
congénitale	Adjectif	def. présent à la naissance
du	Préposition	marqueur faible
doigt	Substantif	concept (doigt)

◦ Le **Découpage syntaxique** et l'**Analyse sémantique**. Ce module est composé de deux sous modules :

- le segmenteur syntaxique qui découpe la phrase en des séquences comparables aux entrées du dictionnaire.

Dans le cadre du traitement des textes médicaux, il est illusoire de vouloir utiliser des analyseurs syntaxiques complets du français car les textes sont souvent non conformes à la grammaire du langage naturel; d'où la nécessité d'utiliser un segmenteur syntaxique simple mais adapté.

¹ Le silence, c'est le fait de ne rien proposer ou de ne pas proposer assez de réponses pertinentes lors d'une consultation de la base

La segmentation syntaxique permet de limiter au maximum le nombre d'essais inutiles de comparaisons et d'éviter la détection de faux concepts. Elle se fait à l'aide de marqueurs de groupes nominaux et verbaux : conjonctions, prépositions, expressions prédicatives ainsi que des signes de ponctuation. Nous distinguons deux types de marqueurs, à savoir les marqueurs faibles et les marqueurs forts. Les marqueurs faibles (de, des, du, d', etc) séparent les mots susceptibles de former ensemble un groupe nominal correspondant à une entrée du dictionnaire. Quant aux marqueurs forts, ceux qui nous intéressent le plus, ils séparent des mots qui n'ont aucune chance de constituer ensemble un groupe nominal correspondant à une entrée du dictionnaire. Les paraphrases ne peuvent être générées qu'au niveau des séquences délimitées par des marqueurs forts. Ainsi, des expressions prépositionnelles comme "au cours de", les prépositions comme "avec", des expressions prédicatives telles que "dû à" sont des marqueurs forts qui indiquent pour la plupart des relations sémantiques entre les groupes de mots isolés par eux.

Les marqueurs forts sont subdivisés en deux groupes, les verbes et les autres types. La segmentation va se faire d'abord par rapport au verbe, ensuite, chaque groupe nominal obtenu sera segmenté par rapport aux autres types de marqueurs forts. Un exemple de segmentation est donné dans la figure n°1

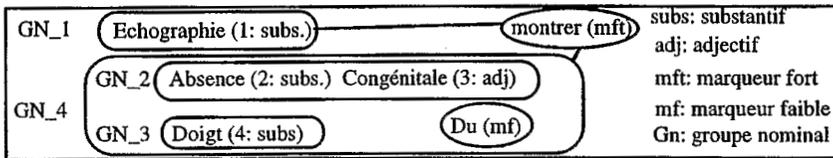


Figure n°1 : Exemple de segmentation d'une phrase

- l'analyseur sémantique, à partir de la segmentation de la phrase, va détecter les concepts et/ou expressions.

On commence d'abord par reconnaître les mots composés avant de reconnaître les mots associés. La reconnaissance des termes se fait dans cet ordre car les liens existant entre les constituants du mot composé sont plus forts que ceux du mot associé.

Cette phase commence par le choix, au sein de chaque groupe de mots, d'un mot (ou des mots) qu'on nommera *Principal*. Le principal nous servira d'entrée dans le dictionnaire afin de détecter une expression. Le principal est un mot de catégorie *substantif*, *préfixe* ou *adjectif* car ce sont ces trois catégories qui sont les plus susceptibles d'occuper la première place dans une expression. S'il existe plusieurs principaux dans un groupe, c'est l'ordre d'apparition des principaux dans la phrase qui fixe la priorité (plus le rang est faible, plus la priorité est grande). Pour la phrase *L'échographie montre une absence congénitale du doigt*, nous avons les principaux Echographie, Absence, Congénitale et Doigt.

Les Principaux obtenus servent essentiellement lors de la reconnaissance des mots composés. Ils permettent d'accélérer le processus de reconnaissance de ces mots tout en évitant les erreurs (mauvaise reconnaissance).

Sachant qu'un terme comporte au moins deux mots, tous les groupes de mots d'au plus un mot ne vont pas nous intéresser.

La reconnaissance des mots composés

Elle consiste d'abord en une consultation du dictionnaire avec les principaux. La consultation a pour but de fournir l'ensemble des entrées de mots composés commençant par le principal en considération ou une de ses flexions. Le principal *Absence* du GN_2 donne :

Absence altération Absence ossification Absence calcifications Absence vaccination

Nous allons appliquer ensuite à cette liste de prétendants obtenue, la règle des mots composés afin de ne retenir uniquement que les bons termes. Aucun des prétendants n'est

reconnu comme mot composé de GN_2. Le second principal *Congénital* ne donne rien. On arrête donc la recherche. Il faut néanmoins souligner que lors de cette phase, les marqueurs faibles sont très facilement transgressés afin de reconnaître des mots composés plus importants, et donc plus précis.

. La reconnaissance des mots associés

Cette reconnaissance consiste d'abord en une consultation du dictionnaire avec tous les mots non nuls et non ignorables ainsi que les mots composés (déjà détectés) de la phrase. Le résultat de la consultation est l'ensemble des mots associés contenant le mot considéré ou un mot de sa famille. Dans cette phase, nous n'allons plus tenir compte des marqueurs faibles. C'est le groupe GN_4 qui nous intéresse car il contient plus d'un mot.

En reprenant la phrase de la figure n°1, la consultation du dictionnaire avec *Absence*, *Congénitale* et *doigt* (ou un mot de leurs familles) donne l'ensemble suivant :

Anomalie congénitale	Déficit salivaire	Déficit intellectuel	Déficit moteur	Doigt petit
Absence congénitale	Cicatrice absence	Alymphocytose congénitale		Manque appétit

On applique ici les différentes règles de détection des mots associés. Ces règles subissent les différentes transformations nécessaires et possibles afin de reconnaître chacun des prétendants. C'est le terme *Absence congénitale* qui est retenu. On remplace dans la phrase les mots *Absence* et *congénitale* par le terme & *Absence congénitale*

La phrase, après détection des divers termes, devient *L'echographie montre une & Absence congénitale du doigt*

Après la reconnaissance des mots associés (reconnaissance qui se fait après celui des mots composés), s'il reste des concepts isolés, ceux-ci seront remplacés par leurs pères dans la phrase par le biais des liens partitifs et génériques du thésaurus afin de rechercher d'éventuels autres mots associés.

Exemple : Considérons que *Papule main* soit un terme du dictionnaire. Si nous avons dans la phrase les concepts isolés *Papule* et *Doigt*, on remplacera *Doigt* par *Main* par le biais de la relation partitive qui lie le *Doigt* à la *Main* (le *Doigt* faisant partie de la *Main*) dans le thésaurus. On pourra donc reconnaître le terme *Papule main*.

Il faut toutefois souligner que si l'on détecte par exemple deux mots composés ou associés et que l'un est contenu lexicalement dans l'autre, on retient le plus long car il est en principe le plus précis et reflète mieux ce qui a été exprimé.

Transgression des marqueurs

La transgression des marqueurs forts est nécessaire afin de détecter les concepts exprimés à travers deux phrases. Exemple de transgression de marqueur fort (*)¹ dans l'énoncé suivant : *Le patient est atteint d'un adénome du lobe droit de la thyroïde. Il a été déjà procédé à une lobectomie partielle gauche*

Gauche n'est pas lié à Lobectomie qui est un acte chirurgical

Gauche n'est pas employé seul car c'est un préciseur d'organe ou de région anatomique (cf définition des constituants du dictionnaire)

Lobe droit de la thyroïde a déjà été détecté

Lobe est une région anatomique

Toutes ces informations vont permettre de transgresser le marqueur fort (*) afin de détecter le concept lobe gauche de la thyroïde.

¹ Le point est un marqueur fort

Dans la version actuelle du système, la transgression des marqueurs forts n'est pas encore implémentée. Elle est encore à l'étude.

Les mots composés et associés ayant été détectés, on procèdera maintenant à la recherche de libellés de nomenclature.

• **Détection des libellés de la nomenclature ADM.** Ce module va se baser sur la phrase transformée (transformation obtenue après la détection des concepts et/ou expressions médicales) ainsi que la terminologie¹ A.D.M. Les libellés de la terminologie sont constitués par une combinaison des différents constituants du dictionnaire. Ce sont de véritables énoncés permettant d'exprimer toutes les notions voulues dans le domaine.

Cette détection va se faire de la même manière que celle des mots associés car nous considérons un libellé comme étant un mot associé.

• **Génération des index.** Il permettra de créer les tables index avec les libellés détectés, les concepts et termes isolés. La création des tables index se fera selon une méthode que nous avons définie. Notre méthode résulte de l'union de plusieurs méthodes existantes [21]. Elle consiste à ne retenir comme éléments pouvant faire partie de la table index que les concepts et/ou termes médicaux, les mots médicaux, les mots inconnus du dictionnaire ainsi que les libellés du thésaurus les plus précis possibles. En effet, par le biais de la hiérarchie du thésaurus, nous allons retenir tous les libellés fils et rejeter les libellés pères. Cela permet non seulement de compresser la table index mais également de la rendre beaucoup plus significative. Le fait d'inclure les mots inconnus dans la table index pourra générer du bruit² mais nous préférons générer du bruit que d'obtenir le silence.

Chaque index est lié à une liste. La liste comporte la référence du texte, le numéro de la phrase dans laquelle l'index apparaît ainsi que la nature de celui-ci.

Exemple : Hépatomégalie < doc 1, ph 3, niv 1> signifie que Hépatomégalie apparaît dans la phrase 3 du document 1 et que c'est un libellé du thésaurus (niv 1). La dernière information permet d'affecter des notes aux différents documents qui sont sélectionnés lors d'une interrogation afin de les afficher par ordre d'importance. Les libellés permettront aux documents d'obtenir la note la plus élevée (3), les concepts et/ou expressions, une note moins élevée (2) enfin les mots qui donneront une faible note (1).

Cette façon de générer les index va nous être très utile au moment de l'interrogation.

4.2 L'interrogation

Elle consiste en l'interprétation des requêtes ainsi qu'à l'extraction de documents. Elle s'articulera autour de deux modules.

• le module **Interprétation des requêtes.** Il permet de saisir la requête de l'utilisateur et d'en extraire les libellés du thésaurus, les concepts et/ou termes ainsi que les mots.

• le module de **Consultation du thésaurus.** Il permet de créer pour chaque libellé, un ensemble de consultation de la table index. En effet, pour chaque libellé, nous allons par le biais des liens de similarité qu'il a avec d'autres libellés du thésaurus, constituer un ensemble d'interrogation sémantiquement complet. C'est avec cet ensemble que nous allons consulter la table des index. Cet ensemble permet de sélectionner tous les documents qui sont sémantiquement proches à partir du libellé de consultation.

• le module d'**Extraction.** Il va se servir des ensembles de consultation créés à partir de chaque libellé de la requête, des concepts et/ou termes ainsi que des mots.

¹ Les libellés de la terminologie sont présents dans le thésaurus.

² Le bruit, c'est le fait de proposer trop de réponses non pertinentes comme solution lors d'une consultation de la base.

L'extraction consiste à sélectionner les documents à partir des ensembles de consultation, des concepts et/ou termes ainsi que des mots. Une fois que les documents ont été sélectionnés, il leur est attribué une note selon le nombre ainsi que la nature des index que chacun contient. Ils sont affichés par la suite par ordre décroissant de la note.

5. Implantation et évaluation

5.1 L'implantation

La représentation de la base de connaissances est fondée sur le modèle relationnel ORACLE. C'est donc un ensemble de tables reliées entre elles par des relations. Les informations y sont extraites grâce aux requêtes SQL.

Le système ADM-INDEX, quant à lui, a été développé en PRO C (langage C utilisant des requêtes SQL pour interroger la base). Les tables index sont également des tables ORACLE. Le système réalisé n'est qu'un prototype et n'intègre pas encore tous les éléments de la base de connaissances. La prochaine version sera plus complète.

Le système prend en entrée un fichier ou un compte rendu médical, y fait toutes les transformations ainsi que tous les traitements et affiche les index retenus. Afin de le tester, nous l'avons rendu accessible par le système d'information WWW (Word Wide Web) [22, 23]. Le système est donc interrogeable par le biais des logiciels clients de WWW (NETSCAPE, MOSAIC). Cette version est légèrement différente de la version initiale car elle prend en entrée une phrase et non un fichier.

La figure n°2 présente le résultat de l'indexation de la phrase Absence congénitale de la main.

The screenshot shows a web browser window titled "LOUIS-PROL". The address bar is empty. The browser's navigation toolbar includes buttons for Back, Forward, Home, Reload, Images, Open, Print, Find, and Stop. Below the toolbar are several menu items: Welcome, What's New?, What's Cool?, Questions, Net Search, and Net Directory. The main content area displays the title "RESULTATS DE L'INDEXATION" in large, bold, serif font. Below the title, the analyzed phrase is shown: "Phrase analysée: ABSENCE CONGENITALE DE LA MAIN". Underneath, the section "Concepts détectés" lists several terms with checked checkboxes: "AGENESIE DES MAINS //ACHEIRIE", "MAIN", "CARENCES", and "AGENESIE //APLASIE CONGENITALE //ABSENCE CONGENITALE". The section "Silences (ou concepts manquants)" contains three empty input fields, each preceded by a small circular bullet point. At the bottom, there is a "Commentaire:" label followed by a text input field. A "VALIDATION" button is located at the very bottom of the page.

Figure n°2 : Résultat de l'indexation de la phrase "Absence congénitale de la main"

L'écran affiche la phrase analysée, les libellés du thésaurus, les concepts et/ou expressions ainsi que les mots médicaux. Chacun des éléments est précédé par une case

précochée, cela donne la possibilité à l'utilisateur de pouvoir désélectionner ("décocher") un élément si celui-ci n'est pas pertinent et cela permet de pouvoir évaluer le bruit.

Il affiche également trois champs permettant de rajouter les codes des éléments éventuellement manquants, cela permet d'évaluer le silence. Enfin, il y a un champ nommé Commentaire qui permet de saisir toutes les remarques possibles concernant la phrase analysée (ex: pourquoi tel concept n'a pas été détecté, etc). Le bouton Validation permet d'enregistrer entre autres toutes les informations de l'écran dans un fichier. Le traitement de ce fichier issu des différents tests permettra d'améliorer notre base de connaissances.

5.2 L'évaluation

Le prototype réalisé a été testé pour le moment sur un corpus de 28 comptes rendus médicaux. Ces tests ont consisté à comparer l'ADM-INDEX à l'indexation effectuée manuellement par un codeur professionnel. A partir des résultats obtenus, nous avons pu calculer les valeurs liés aux indicateurs permettant de mesurer les qualités d'un système de recherches d'informations. Ces indicateurs sont :

- Le **Rappel** : mesure la proportion d'informations pertinentes retrouvées (IPR) par rapport au nombre total de réponses pertinentes (RP). Sa formule est IPR / RP (PR = ensemble de documents sélectionnés par la méthode manuelle et IPR = ensemble de documents sélectionnés par la méthode ADM-INDEX et présents dans PR)

- La **Précision** : mesure la proportion d'informations pertinentes retrouvées par rapport au nombre total de réponses données (RD). Sa formule est IPR / RD (RD = ensemble de documents sélectionnés par la méthode ADM-INDEX)

- Le **Silence** : c'est le complémentaire du rappel, Sa formule est $1 - \text{Rappel}$

- Le **Bruit** : c'est le complémentaire de la précision, Sa formule est $1 - \text{Précision}$

Codes	Libellés d'interrogation	Rappel	Précision	Silence	Bruit
786.5	Douleur thoracique	80 %	33 %	20 %	67 %
305.1	Tabagisme	100 %	50 %	0 %	50 %
Z88.5520	Coronaro	75 %	100 %	25 %	0 %
Z89.52	ECG	95 %	100 %	5 %	0 %
414.0	Lésion tritronculaire	15 %	100 %	85 %	0 %
414.0	Sténose d'artère coronaire	61 %	88 %	39 %	12 %
414.0	Lésion coronarienne	57 %	81 %	43 %	19 %
454.9	Varices des membres inférieurs	100 %	100 %	0 %	0 %
272.8	Hypercholestérolémie	57 %	100 %	43 %	0 %
278.0	Obésité	71 %	83 %	29 %	17 %
Z89.49	Radio pulmonaire	29 %	100 %	61 %	0 %
Z89.43	Epreuve d'effort	83 %	29 %	17 %	71 %
V17.3	Hérédité coronarienne	100 %	100 %	0 %	0 %
491.9	Bronchite chronique	100 %	100 %	0 %	0 %
427.8	Bradycardie	100 %	50 %	0 %	50 %
401.9	Hypertension artérielle	100 %	85 %	0 %	15 %
429.3	Hypertrophie ventriculaire gauche	100 %	100 %	0 %	0 %
285.9	Anémie	0 %	0 %	100 %	100 %
413	Angor	75 %	50 %	25 %	50 %
	Moyenne	69,9 %	72,5 %	30,1 %	27,5 %

Figure n°3 : Tableau récapitulatif des tests effectués sur l'ADM-INDEX

Les résultats obtenus sont prometteurs. On constate que certains libellés tels que Bradycardie permettent d'obtenir des valeurs très élevées. Cela s'explique simplement par le fait que ces libellés ne permettent de sélectionner manuellement qu'un ou deux documents aux maximum (ce que fait également le système ADM-INDEX). Même si on a l'impression qu'ils modifient anormalement les résultats du système, ils constituent de très bons tests car ils permettent de mettre en évidence la précision du système.

Nous remarquons également que les libellés *Lésion tritronculaire*, *Sténose d'artère coronaire* et *Lésion coronarienne* ne permettent pas d'obtenir les mêmes résultats pourtant ils ont tous le même code. Cela est dû à une absence de liens entre ces libellés dans notre base.

6. Discussion et perspective

Le système que nous avons développé est dépendant de son domaine d'application, c'est-à-dire sa base de connaissances. Mais cette base à l'avantage de couvrir pratiquement tous les domaines médicaux et d'être complète car elle prend en compte les particularités du langage médical, ce qui est très important.

En outre, le système n'est pas lié uniquement à la terminologie de L'A.D.M., il est possible de remplacer celle-ci par un autre thésaurus ou une autre terminologie médicale. Il suffit simplement de transcrire cette terminologie dans le formalisme ADM-INDEX par le biais d'une passerelle que nous avons développée. Nous avons donc pu l'adapter facilement à la CIM9 et CIM10 (Classification Internationale des Maladies, 9ème et 10ème révision), à SNOMED (Systematized Nomenclature of MEDecine), au CDAM (Catalogue Des Actes Médicaux, ancienne et nouvelle version) ainsi qu'au MESH (Medical Subject Headings).

Au niveau du module d'extraction, la méthode de sélection élimine au maximum le silence, ce qui est très important car le manque de document peut être nuisible si ces documents contiennent des informations pertinentes. En contre partie, elle augmente le bruit. Ce qui n'est pas trop gênant pour notre système mais il serait intéressant d'éliminer ce bruit inutile.

Pour éliminer celui-ci, nous allons comparer chaque document obtenu après la première sélection au contenu sémantique de la requête. Ceci est facile à faire dans la mesure où il nous est possible connaître la phrase dans laquelle un index se trouve. Le nouvel ensemble que nous allons obtenir sera l'ensemble de départ épuré des documents sélectionnés mais non pertinents. Cela suppose que nous devons, lors de l'extraction des libellés et des concepts de la requête, extraire également le sens de celle-ci. Le sens permettra entre autres de marquer d'une part positivement ou négativement les libellés auxquels elle est rattachée et d'autre part de repertorier les préciseurs (adjectifs).

Exemple : Si l'utilisateur rentre *Hépatomégalie chronique*. *Chronique* sera considéré comme un préciseur et *Hépatomégalie* comme un libellé d'indexation. Lors de la recherche de documents, on va d'abord sélectionner tous les documents contenant *Hépatomégalie*, évidemment, cette première sélection va générer du bruit, bruit que nous allons réduire en réalisant une deuxième sélection à partir de la première sélection en utilisant la sémantique de la question, c'est-à-dire en prenant en compte le préciseur *Chronique*. Sachant que notre index est structuré de la façon suivante < n° doc, n° phrase, niv >, on pourra accéder aux phrases et confronté leur contenu à celui de la requête et ne retenir finalement comme solution, les documents dont les phrases concordent le mieux avec le sens de la requête.

7. Références

- [1]. Grémy François: *Informatique médicale: introduction à la méthodologie en médecine et santé publique*. Paris, Décembre 1986.
- [2]. Ghazi Joseph: *Vocabulaire du discours médical, structure, fonctionnement, apprentissage*. Edition Didier Eudition, 1985

- [3]. Graitson M.: *Aspect du traitement computationnel des données médicales en langues naturelles*. Thèse de doctorat, Université de Liège, 1983
- [4]. Graitson M.: *Traitement automatique du français médicale*. cahiers de lexicologie, 30(1), pp. 47-74, 1977
- [5]. Miller R. A., M.D., Pople H. E., Jr., Ph.D., Myers J. D., M.D. : *INTERNIST-I, an experimental computer-based diagnostic consultant for general internal medicine*. The New England Journal of Medicine. 1982; 307: 468-477.
- [6]. Miller R. A., Masarie F. E., Myers J. D. : *Quick Medical Reference (Q.M.R.) for diagnostic assistance MD Comput*. 1986; 3: 34-48
- [7]. Nelson S. J., M.D., Blois M. S., Ph.D., M.D., Tuttle M. S., Eribaum M., M.D., Harrison P., M.D., Kim H., M.D., Winkelmann B., M.D., Yamashita D., M.D. : *Evaluating RECONSIDER - A computer program for diagnostic prompting J. Med. Syst.*, vol. 9, nos. 516. 1985
- [8]. Lenoir P., Riou C., Fresnel A.: *L'aide au diagnostic médical (ADM). Modalités et perspectives*. Médecine de l'homme N° 135.
- [9]. Lenoir P., Michel J. R., Frangeul C., Chales G.: *Réalisation, développement et maintenance de la base de données A.D.M.* Médecine informatique (1981), vol. 6, N° 1, 51-56
- [10]. Jayez Jacques: *L'inférence en langage naturel*. Ed. Hermès, Paris, 1988
- [11]. Sowa J. F.: *Conceptuel structures. Information processing in mind and machine*. Readings, Massachusetts: Addison-Wesley, 1984.
- [12]. Quillian M. R. : *Semantic memory*. In *Semantic information processing*. Minsky. MIT press. Cambridge Mass, 1968 : 227-270
- [13]. Charniak E. : *Towards a model of children's story comprehension*. T. R. n°266, MIT Artificial Intelligence Lab. 1972
- [14]. Minsky M.: *A framework for representing knowledge, in the psychology of the computer vision*, Ed. P.E. Winston, Mc Graw Hill, 1975
- [15]. Améli S.: *Construction d'un langage de dictionnaire conceptuel en vue du traitement du langage naturel: Application en langage médical*. Thèse CDS, UTC Compiègne, 1989.
- [16]. De Heulme M., Membrado M., Améli S., Vexler F.: *Ambiguïté et paraphrase dans le langage médical et leur traitement par TRANSLOG*. In *l'Ambiguïté et la paraphrase*. pp.287-291, Acte du colloque de Caen, 9-11 Avril 1987
- [17]. Amsler R. A.: *A taxonomy for english nouns and verbs*. ACL Conf. pp. 133-138. 1980
- [18]. Fargues J., Perrin A.: *Synthetising a large concept hierarchy from french hyperonyms*, Colins'90, vol. 2, 1990
- [19]. Burgun A., Botti G., Lukacs B., Mayeux D., Seka L. P., Delamarre D., Bremond M., Kohler F., Fieschi M., Le Beux P.: *A system that facilitates the orientation within nomenclatures through a semantic approach*. Med. Inform. (1994), Vol. 19, N°4, 297-310
- [20]. Landais P., Jais J.P., Frutiger P.: *Sémantique des classifications et nomenclatures*. Informatique et Santé, 1989, 1 : 211-222.
- [21]. Hersh W. R., Hickam D. H., Leone T. J.: *Words, concepts or both: optimal indexing units for automated information retrieval*. In *Proceedings SCAMC 93*, pp. 644-648, 1993
- [22]. Dagorn F., Gross C.: *Le projet World Wide Web*. Avril 1994
- [23]. Pouliquen B., Riou C., Denier P., Fresnel A., Delamarre D., Le Beux P.: *Using World Wide Web multimedia in medicine*. Medinfo 95 (1995), 1519-1523