

UNIVERSITE DE RENNES I
FACULTE DE MEDECINE

No attribué par la bibliothèque: _____

THESE

Pour obtenir le grade de
DOCTEUR DE L'UNIVERSITE RENNES I
Discipline : **Génie Biologique et Médical**

Présentée et soutenue publiquement

par

POULIQUEN Bruno

Le 7 juin 2002

Titre :

Indexation de textes médicaux par extraction de concepts, et ses utilisations

Directeur de thèse : Pr. Pierre Le Beux

JURY

Pr. Pierre Le Beux , Président
Pr. Régis Beuscart , Rapporteur
Pr. Pierre Zweigenbaum , Rapporteur
Mme Anita Burgun

Je tiens à remercier particulièrement le professeur Pierre Le Beux pour m'avoir offert la possibilité de rédiger cette thèse, et pour son dynamisme qui m'a toujours poussé à explorer de nouvelles technologies.

Je remercie sincèrement les professeurs Pierre Zweigenbaum et Régis Beuscart d'avoir accepté d'être rapporteurs de ma thèse.

Mes remerciements vont aussi à Anita Burgun pour toutes ses remarques pertinentes sur ma thèse, ses encouragements et son amitié.

Merci également au professeur Régis Duvauferrier qui nous a offert la possibilité d'appliquer ces techniques de traitement automatique du langage naturel. J'ai particulièrement apprécié ses encouragements et son enthousiasme.

J'exprime aussi ma gratitude à Pascale Sébillot qui a accepté de lire un premier brouillon de ma thèse. Ses conseils ont été extrêmement précieux.

Je tiens à remercier toute l'équipe du Laboratoire d'Informatique Médicale, et du Département d'Information Médicale de Rennes pour ces dix années passées en leur compagnie, pour leurs aides dans le travail, et, surtout, pour leur amitié. Je remercie Denis Delamarre, qui a eu l'une des idées à la base de ce travail, de m'avoir accompagné dans les divers développements.

Un grand merci à Michel Kerbaol et Jean-Yves Bansard pour m'avoir aidé à découvrir et expérimenter les travaux de Benzécri.

De même, je voudrais remercier l'équipe du CISMéF, et plus particulièrement Stefan Darmoni, qui a fait une évaluation détaillée de mes travaux. Leurs encouragements ont été très motivants au moment de la rédaction de cette thèse.

Une pensée particulière pour mon grand-père et ma mère, bretons dont la langue maternelle n'était pas le français, et, plus généralement, à tous ceux qui m'ont fait découvrir toute la richesse des langues.

Enfin, je ne remercierais jamais assez ma famille, Sylvie, Glen et la petite Anna, mes amis (Français, Italiens, Indiens et autres) d'avoir supporté mon travail, et, par-dessus tout, d'être là.

Résumé en français: Nous nous intéressons à l'accès à l'information médicale. Nous avons utilisé un lexique de flexions, dérivations et synonymes de mots spécifiquement créé pour le domaine médical, issu de la base de connaissances "Aide au Diagnostic Médical". Nous avons exploité les mots composés et les associations de mots de ce lexique pour optimiser l'indexation d'une phrase en mots de référence. Nous avons créé un outil d'indexation permettant de reconnaître un concept d'un thésaurus médical dans une phrase en langage naturel. Nous avons ainsi pu indexer des documents médicaux par un ensemble de concepts, ensuite nous avons démontré l'utilité d'une telle indexation en développant un système de recherche d'information et divers outils: extraction de mots-clés, similarité de documents et synthèse automatique de documents. Cette indexation diminue considérablement la complexité de la représentation des connaissances contenues dans les documents en langage naturel. Les résultats des évaluations montrent que cette indexation conserve néanmoins la majeure partie de l'information sémantique.

Titre en anglais: Medical texts indexation using concepts extraction, and its use

Résumé en anglais: The work presented specifically targets the accessibility to medical information. We used a French medical lexicon (specifically created for the medical domain), and built an index tool to particularly recognize a concept from a medical thesaurus that is present in a sentence written in natural language. First we indexed medical documents with a set of concepts and then demonstrated the utility of such indexing by developing a search engine and various tools which include: keyword identification, document similarity and automatic document synthesis. This indexing greatly aided in reducing the repository complexity of natural language documents. In addition, the evaluation results demonstrate that this indexing retains the main semantic information.

Discipline: Génie Biologique et Médical

Mots-clés : Traitement automatique des langues naturelles, Indexation, Médecine, Système de recherche d'information, Lexique, Thésaurus, Web.

Laboratoire d'Informatique Médicale – Faculté de Médecine - 35033 Rennes cedex – France

[Bruno.Pouliquen@univ-rennes1.fr](mailto: Bruno.Pouliquen@univ-rennes1.fr)

PLAN DE LA THESE

INTRODUCTION	1
MATÉRIEL ET MÉTHODES	5
I. ÉTAT DE L'ART.....	5
.I.1. Introduction.....	5
.I.2. Différentes indexations existantes.....	7
.I.3. Les descripteurs.....	7
.I.4. Simplification de documents.....	9
.I.5. Segmentation de textes.....	11
.I.6. Systèmes de Recherche d'Information (S.R.I.).....	12
.I.7. Similarité de documents.....	14
.I.8. Synthèse automatique de documents.....	15
.I.9. Visualisation et classification de documents.....	15
.I.10. Attribution de mots-clés.....	21
.I.11. Méthodes d'évaluation.....	22
.I.12. Ressources linguistiques.....	26
II. CONTEXTE.....	30
.II.1. La base de connaissances ADM.....	30
.II.2. UMLS.....	42
.II.3. Corpus de textes disponibles.....	45
.II.4. Outils existants.....	46
RÉALISATIONS	49
I. CHOIX LINGUISTIQUES.....	49
II. CHOIX TECHNIQUES.....	51
III. FONCTIONNEMENT DU SYSTÈME.....	54
IV. INDEXATION EN MOTS DE RÉFÉRENCE.....	56
.IV.1. Représentation du lexique.....	56
.IV.2. Fonctionnement du programme de reconnaissance de mots.....	61
.IV.3. Reconnaissance de mots multiples et performance.....	62
.IV.4. Préfixes et suffixes.....	63
.IV.5. Mots inconnus.....	63
.IV.6. Résumé du principe de fonctionnement.....	64
.IV.7. Indexation d'un thésaurus en mots de référence.....	65
V. INDEXATION DE DOCUMENTS EN CONCEPTS.....	70
.V.1. Principe de fonctionnement.....	71
.V.2. Algorithme.....	72
VI. INDEXATION DU CORPUS DE TEXTES AVEC NOMINDEX.....	73
VII. REPRÉSENTATION DE L'INDEXATION.....	76
VIII. INTÉRÊT DE LA HIÉRARCHIE.....	80
IX. SYSTÈME DE RECHERCHE D'INFORMATION.....	82
.IX.1. Pondération des poids des concepts.....	82
.IX.2. Similarité.....	84
X. SIMILARITÉ DE DOCUMENTS.....	87
XI. SYNTHÈSE AUTOMATIQUE DE DOCUMENT.....	89
.XI.1. Méthode.....	89
.XI.2. Développement.....	91
XII. ATTRIBUTION DE DOMAINE.....	92
.XII.1. introduction.....	92
.XII.2. Phase d'apprentissage.....	92
.XII.3. Extension des domaines au thésaurus MeSH.....	95
.XII.4. Attribution d'un domaine à un document.....	95
XIII. AUTRES UTILISATIONS DE L'INDEXATION.....	97
.XIII.1. Similarité de concepts.....	97
.XIII.2. Classification de documents.....	100
XIV. TRADUCTION AUTOMATIQUE POUR INDEXATION.....	101
.XIV.1. Introduction.....	101
.XIV.2. Constitution des lexiques.....	102

.XIV.3.	<i>Première étape</i>	102
.XIV.4.	<i>Enrichissement du lexique</i>	103
.XIV.5.	<i>Méthode de traduction</i>	104
.XIV.6.	<i>Quelques statistiques sur le lexique</i>	104
.XIV.7.	<i>Langues disponibles</i>	104
.XIV.8.	<i>Exemple de traduction</i>	105
.XIV.9.	<i>Reconnaissance de la langue d'un texte</i>	106
RÉSULTATS		108
.I.	EVALUATION.....	110
.I.1.	<i>Évaluation du lexique</i>	110
.I.2.	<i>Évaluation du système de recherche d'information</i>	113
.I.3.	<i>Extraction de mots-clés</i>	115
.I.4.	<i>Traduction pour l'indexation</i>	123
.II.	ANALYSE FACTORIELLE DES CORRESPONDANCES (APPLICATION AUX CONCEPTS).....	126
.II.1.	<i>AFC en texte intégral</i>	126
.II.2.	<i>AFC sur les concepts extraits</i>	126
.II.3.	<i>Perspectives</i>	129
DISCUSSION		130
.I.	EN AMONT.....	130
.II.	L'OUTIL NOMINDEX.....	132
.III.	EN AVAL.....	133
CONCLUSION		135
BIBLIOGRAPHIE		137
.I.	INDEX DES ILLUSTRATIONS.....	148
.II.	INDEX DES TABLES.....	148
.III.	INDEX DES ALGORITHMES.....	149
.IV.	INDEX DES ÉQUATIONS.....	149
GLOSSAIRE		150
ANNEXES		157

Introduction

L'informatisation des hôpitaux, des cabinets médicaux et le développement d'Internet entraînent une prolifération de l'information médicale. Cette information médicale est de plus en plus souvent écrite, et disponible sous forme informatique. Or cette information est souvent mal exploitée car hétérogène et difficile d'accès. On demande de plus en plus aux médecins de "coder" leurs informations (nomenclatures d'actes, de diagnostics, d'examens...), dans le but de pouvoir "indexer" de manière automatique ou semi-automatique ces informations.

L'avènement du réseau Internet et des sites web médicaux offrent une quantité jamais égalée de textes médicaux, et en constante augmentation. De plus en plus de professionnels de la santé ou d'étudiants sont maintenant connectés. Quand un simple menu suffisait auparavant pour retrouver un cours de médecine en France, ce menu aurait aujourd'hui plus de 10 000 propositions [Darmoni et al., 2000]. Les outils d'indexation et de recherche d'information sont très demandés à l'heure actuelle.

Les thésaurus médicaux constituent un ensemble de concepts médicaux sélectionnés par des experts du domaine (sélection dépendant bien sûr de l'objectif du thésaurus), ces thésaurus sont mis à jour régulièrement, ces systèmes constituent donc une représentation conceptuelle de la connaissance médicale. Cette représentation des connaissances peut être partielle et subjective, mais constitue néanmoins une base de travail très intéressante, "utilisant des concepts clairement bien définis" [Le Beux et al., 1995]. La notion de "concept" utilisée dans ce travail, n'est pas si loin de la définition de Platon et Aristote qui l'expliquaient par une entité devant avoir le même sens pour divers usagers, avec des relations parfaitement définies avec les autres concepts. Les thésaurus médicaux sont, de ce point de vue, un bon répertoire de concepts médicaux.

La compréhension complète d'un cours de médecine est un travail qui demande plus d'une dizaine d'années de formation pour un humain (qui lui-même a été "construit" depuis quelques millions d'années), pour une machine cela représente un travail titanesque (impossible à l'heure actuelle). On peut néanmoins essayer d'extraire de ce cours les informations contenant le plus de sémantique. Le but de ce travail est de reconnaître les entrées de thésaurus (que nous appellerons concepts) dans les textes médicaux, ce qui permet de réduire un texte à un ensemble de concepts. Nous acceptons donc d'emblée une perte d'information, mais cette nouvelle représentation a l'énorme avantage d'être facilement exploitable de manière informatique.

Une partie importante de ce travail consistera à résoudre les problèmes posés par le traitement automatique du langage naturel (afin de reconnaître des termes comme "Rachialgie" et "Dorsalgie" dans la phrase "Le patient s'est plaint de douleurs au niveau du rachis dorsal").

Des projets d'analyse du langage médical, comme MENELAS [Zweigenbaum et al., 1994], ont déjà vu le jour, mais ceux-ci s'appuient sur une architecture complexe et difficile à mettre en oeuvre, et requièrent une expertise complète du langage médical (qui peut se chiffrer en dizaines d'années-homme). Ajouter des connaissances médicales nécessite le recours à des experts de chaque domaine, et il n'est pas toujours facile d'obtenir un consensus, parfois même si on se limite à l'information lexicale (pour prendre l'exemple célèbre de Jules Romain, doit-on dire que "gratouiller" est un quasi-synonyme de "chatouiller"). Sans compter que la saisie d'une nouvelle information médicale, dans un système informatique, est rarement intuitive, et demande soit une formation de l'expert, soit la présence d'un opérateur avisé (qui "traduit" le langage naturel). Nous avons voulu construire un outil simple, robuste, qui se basera sur des connaissances existantes. Ces connaissances pourront être, dans le futur, des résultats d'acquisition d'informations à partir de corpus.

Au Laboratoire d'Informatique Médicale de la faculté de médecine de Rennes, nous disposons d'une base de connaissances ADM (Aide au Diagnostic Médical) [Lenoir et al., 1981] qui contient près de 200 000 concepts (essentiellement des symptômes, maladies et syndromes). Cette base de connaissances ADM contient également un lexique médical français qui couvre la plus grande partie du vocabulaire médical.

Le but initial, la base de la pyramide, sera de réaliser un "extracteur de mots de référence", un outil qui tiendra compte du dictionnaire ADM pour extraire d'une phrase en langage naturel tous les mots de référence.

Ensuite, nous créerons un "moteur d'indexation" pour extraire des textes médicaux (en langage naturel) des concepts appartenant à un thésaurus médical. Le résultat de l'indexation se résumant à l'ensemble des concepts détectés dans le document. S'il est clair que cette procédure entraîne une perte d'information importante, le résultat de l'indexation est un modèle beaucoup plus simple à traiter automatiquement qu'un texte en langage naturel, l'évaluation des résultats nous montre que cela permet de répondre à la plupart des attentes des utilisateurs.

Cette indexation devra être entièrement automatisée de manière à pouvoir être utilisée de manière intensive. Le domaine d'application reste vaste, aussi nous avons fait le choix de

tester notre système sur deux thésaurus médicaux: celui extrait de l'ADM, et le MeSH de la U.S. National Library of Medicine.

L'indexation n'étant pas un objectif en soi, nous avons orienté notre travail sur les diverses applications de cette indexation. La première application sera de pouvoir rechercher les documents d'un corpus de textes correspondant à une question en langage naturel. Les autres applications seront : la similarité de documents, la synthèse automatique et l'extraction de mots-clés. Nous n'avons que partiellement abordé la classification des documents, le sujet est très vaste et dépasse le cadre de ce travail, mais nous avons effectué des tests avec la méthode statistique d'analyse factorielle des correspondances (une des méthodes permettant de classer les documents) qui montrent que notre outil améliore très sensiblement les résultats (par rapport aux résultats obtenus en analysant les mots d'un document).

Après un aperçu des ressources existantes, nous présenterons le moteur d'indexation, que nous dénommons NOMINDEX, et son utilisation dans différents contextes. Nous finirons cet exposé par une présentation des résultats obtenus.

En résumé, les ressources utilisées seront : un lexique extrait de l'ADM (regroupement de mots en famille), un thésaurus médical (ADM, MeSH...), et un corpus de textes médicaux extrait du web.

Comme le dit Eric Laporte [Laporte, 2000] : "les documents électroniques accessibles dans les sites web constituent un champ de recherches documentaires et de veille technologique vaste et en pleine expansion". Mais ces documents sont, selon l'inventeur du web, "destinés aux humains plutôt que des données et informations qui peuvent être analysées automatiquement" [Berners-Lee et al., 2001], le défi est justement d'extraire automatiquement de l'information de ces documents écrits en langage naturel. "La puissance de la langue naturelle crée en même temps un obstacle à son utilisation pour le traitement de l'information" [Zweigenbaum, 1999].

Cette thèse s'articule en quatre points principaux :

Nous commencerons par faire un état des lieux des méthodes dont nous disposons pour améliorer l'indexation dans ce domaine.

Ensuite, nous présenterons les bases de connaissances et les outils existants.

Le second chapitre présentera les choix et les réalisations effectués : le moteur d'indexation et les différents outils qui l'utilisent.

Nous avons essayé, dans la mesure du possible, d'effectuer une évaluation des réalisations, ce sera l'objet du dernier chapitre.

Un glossaire se trouve en fin de document (p. 150)

Matériel et méthodes

.I. *État de l'art*

.I.1. Introduction

Le but de l'indexation est de créer une représentation permettant de repérer et retrouver facilement l'information dans un ensemble de documents. [Lancaster, 1998] donne cette définition : "Le but principal de l'indexation (et du résumé automatique) est de construire des représentations d'éléments publiés sous une forme adaptée pour le stockage dans tout type de base de données".

On utilise cette indexation, le plus souvent, pour les systèmes de recherche d'informations. Mais, nous le verrons dans les différentes applications, cette indexation peut également servir à comparer et classer des documents, proposer des mots-clés, faire une synthèse automatique de documents, calculer des co-occurrences de termes... Tout index de document perd une partie de l'information initiale. Un dictionnaire de la langue française, un index dont l'utilité n'est plus à démontrer, ne pourra jamais à lui seul représenter toute la complexité de la langue.

Le processus d'indexation peut être représenté par le schéma suivant :

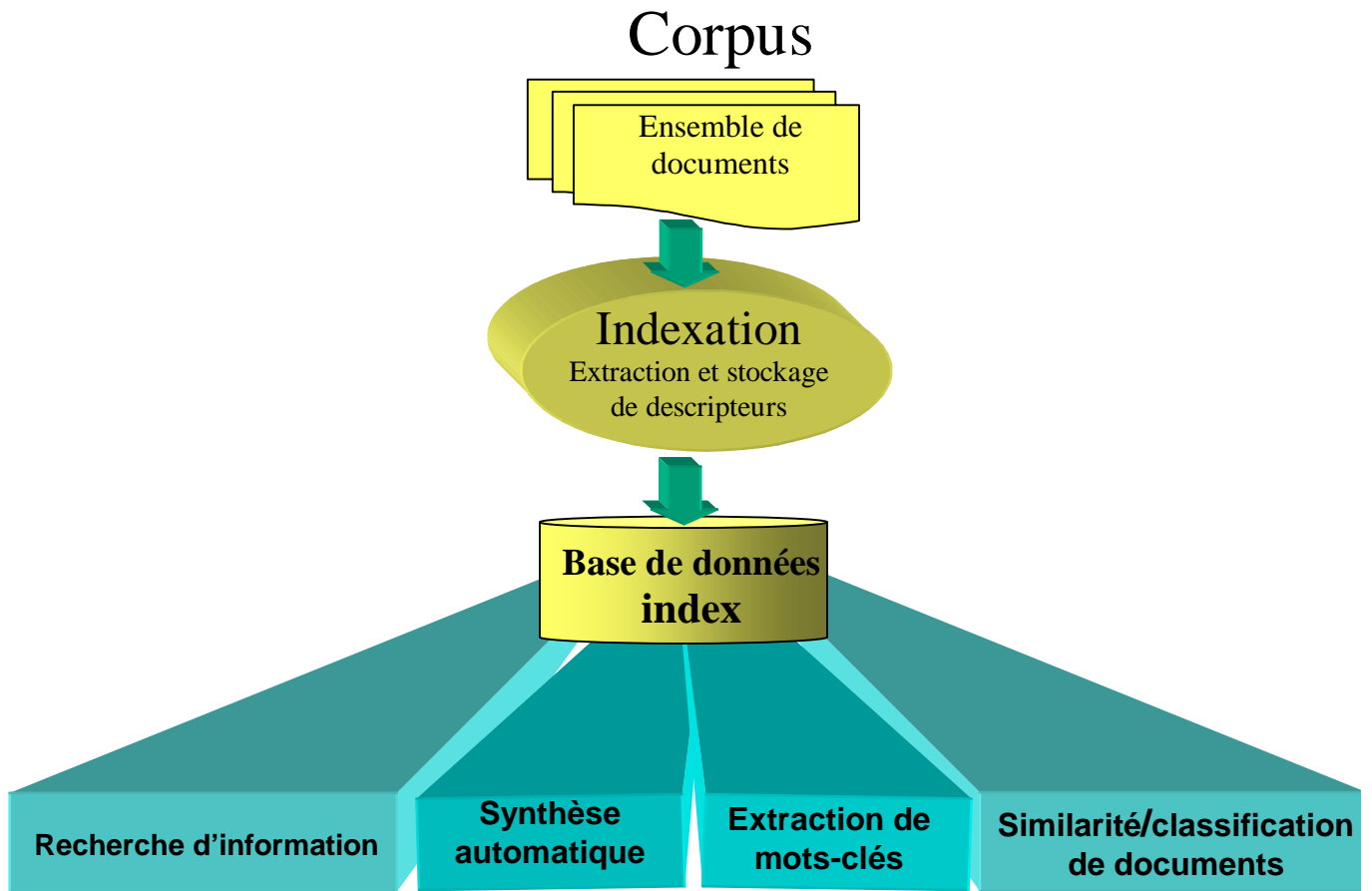


Figure 1 : Processus d'indexation

Après un rapide tour d'horizon des outils et des méthodes d'indexation existants, nous présenterons, dans ce chapitre, les différents usages qui en sont fait.

.1.2. Différentes indexations existantes

On distingue trois types d'indexation :

- L'indexation manuelle, une personne a préalablement désigné les termes d'indexation : les descripteurs (voir ci-dessous) associés à chaque texte
- L'indexation semi-automatique (ou supervisée), un programme détecte, pour chaque document, des descripteurs qui sont proposés à un utilisateur (qui peut valider, supprimer ou, parfois, ajouter des descripteurs)
- L'indexation automatique (par opposition appelée parfois "non-supervisée"), le programme fonctionne sans intervention humaine.

On distingue aussi deux types de langages d'indexation :

- Langage contrôlé : Utilise un lexique de descripteurs figé, l'indexation est le plus souvent manuelle (un professionnel choisit un ou plusieurs descripteurs pour représenter le document), parfois semi-automatique.
- Langage libre : Les descripteurs sont extraits automatiquement des documents, ou de la requête de l'utilisateur (le plus souvent un document est indexé par la liste des mots qui le composent)

.1.3. Les descripteurs

Ils représentent l'information atomique d'un index. Ils sont censés indiquer de quoi parle le document [Laporte, 2000]. On parle aussi d'unités élémentaires (en anglais "*tokens*") [Jacquemin et Zweigenbaum, 2000]. Le but étant de les choisir de manière à ce que l'index (qui réduit la représentation) perde le moins d'information sémantique possible.

Habituellement les descripteurs sont :

Les **mots** du document : toute chaîne de caractères compris entre deux séparateurs (espace, virgule...)

Les **lemmes** : un processus supplémentaire appelé lemmatiseur convertit les documents préalablement pour extraire des mots de référence (ainsi les mots "CŒURS" et "CŒUR" aboutiront au même descripteur). La racinisation ("stemming") quant à elle est un processus linguistiquement plus simple. Elle consiste, dans sa version la plus simple, à enlever des mots les derniers caractères (considérés comme décrivant les flexions de mots), par exemple :

enlever le "s" des mots au pluriel. Certaines racinisations utilisent des connaissances morphologiques plus complètes (suffixes, préfixes...). Au lieu d'indexer un texte par des mots on l'indexe alors par le lemme correspondant.

Les **concepts**, termes ou mots-clés : il s'agit d'expressions (pouvant contenir un ou plusieurs mots). Ces concepts sont le plus souvent entrés manuellement (cas de l'indexation manuelle, ou semi-automatique). Ces concepts peuvent être écrits de manière libre par un utilisateur, ou, ce qui est souvent le cas, doivent être choisis parmi une liste de concepts (on parle alors de vocabulaire contrôlé). Cette liste de concepts sera le plus souvent décrite dans un thésaurus (dans le cas des termes, on parlera de terminologie).¹

Plus rarement :

Les **N-grammes** : Il s'agit d'une représentation originale d'un texte en séquences de *N* caractères consécutifs. On trouve des utilisations de bigrammes et trigrammes dans la recherche documentaire (ils permettent de reconnaître des mots de manière approximative et ainsi de corriger des flexions de mots ou même des fautes de frappe ou d'orthographe). Ils sont aussi fréquemment utilisés dans la reconnaissance de la langue d'un texte (par exemple: [Harbeck et Ohler, 1999], [Dunning, 1994]).

Les **contextes** : dans le cas du "Latent Semantic Indexing" ([Deerwester et al., 1990]) les documents et leurs mots sont représentés sur d'autres dimensions où les mots apparaissant dans un même contexte sont proches. Cette indexation est le résultat d'une analyse des co-occurrences des mots dans un corpus (tout comme l'Analyse Factorielle des Correspondances, cf. p. 16).

Voici quelques exemples de résultats d'indexation sur un document ne contenant que la phrase "Les accidents vasculaires cérébraux" :

Par **mots** : "Les", "accidents", "vasculaires", "cérébraux"

Par **lemmes** : "le", "accident", "vasculaire", "cérébral"

Par **racines** : "l", "accident", "vascul", "cérébr"

Par **concepts** : "A.V.C."

Par **bigrammes** : "_l", "le", "es", "s_", "_a", "ac", "cc", "ci", "id", "de", "en", "nt", "ts", "s_", "_v", "va", "as" ... "au", "ux", "x_"

¹ Ces concepts peuvent éventuellement être tous les groupes nominaux, repérés par des outils syntaxiques (ou mixtes) tels que définis dans [Jacquemin et Tzoukermann, 1999].

Le plus souvent, un système d'indexation avec vocabulaire contrôlé travaille à partir d'une base de connaissances. Cette base de connaissances (la plupart du temps contextuelle) lui permet de "choisir" les descripteurs les plus appropriés en fonction du document analysé.

Le choix du type de descripteurs utilisé est primordial, et sera déterminant pour les performances de l'indexation. La plupart des moteurs d'indexation fonctionnent sur les mots (notamment les moteurs de recherche sur le Web). La littérature anglophone fait le plus souvent l'éloge de l'indexation par mots, une évaluation a même montré que la racinisation ("Stemming", la plus simple des méthodes linguistiques) n'améliorait pas de manière significative les performances [Frakes, 1992, p. 141]. Ce qui a même conduit des auteurs à déclarer: "S'il est difficile de démontrer l'utilité d'un processus aussi simple que la racinisation, comment pouvons-nous justifier notre intérêt dans des procédés plus ambitieux du traitement du langage naturel" [Church, 1995, p. 310]. Le danger serait de ne se référer qu'à la littérature anglophone, car de multiples autres études ont démontré que pour les autres langues, les processus supplémentaires augmentaient de manière significative les performances des systèmes (pour la racinisation par exemple citons [Sheridan et Ballerini, 1996], [Choueka et Zampoli, 1992]).

.1.4. Simplification de documents

Pour améliorer l'indexation quelques processus supplémentaires sont souvent mis en œuvre (le but étant de simplifier les documents) :

- **L'élimination des mots-outils** (mots stop, stopword) : il s'agit d'éliminer les mots du langage courant qui ne contiennent pas beaucoup d'information sémantique. (Exemple : "a", "le", "la", "de" ...). Certains indexeurs retirent systématiquement tous les petits mots (d'un, deux voire trois caractères).
- **La conversion de caractères** :
 1. Minuscules/majuscules : beaucoup de moteurs d'indexation sont insensibles à la casse et traitent de la même manière les mots en majuscules ou en minuscules. Une fonction transformera par exemple tous les mots en majuscules.
 2. Caractères diacritiques : il existe encore d'anciens textes écrits en ascii sur 7 bits (c'est-à-dire sans accents), qu'il faut comparer avec des textes accentués, encore aujourd'hui, certains titres de documents, en majuscules, ne sont pas accentués, une

solution consiste donc à convertir chaque caractère diacritique en ascii sur 7 bits dans le processus d'indexation.²

3. Lorsque l'on veut indexer des textes multilingues, on convertit les caractères dans un format commun (ex: iso-latin [ISO-latin, 1987], ou unicode [Unicode, 1997])
- **Les corrections orthographiques** : les fautes d'orthographe ou fautes de frappe peuvent être corrigées automatiquement avant la phase d'indexation. Si ces fautes sont relativement rares dans les documents, elles sont nombreuses dans les requêtes d'interrogation des utilisateurs (Comme nous le verrons sur un exemple dans le chapitre évaluation p. 110), une correction automatique est très souvent associée au moteur d'indexation afin de ne pas créer deux descripteurs différents pour un mot mal orthographié et sa forme correcte³. Le danger étant de "corriger" abusivement certains mots corrects mais inconnus du lexique (notamment les noms propres⁴).
 - **La reconnaissance de mots composés** : Source de problèmes, certains indexeurs utilisent une table de mots composés du langage pour les identifier comme ne formant qu'un seul mot ([Gross, 1986] ou [Habert et Jacquemin, 1993]). En effet, il paraît indispensable de considérer "bec de lièvre" comme un mot à part entière (et ne pas l'indexer par "bec" et "lièvre").
 - **L'étiquetage lexical** : il s'agit d'un outil qui associe à chaque mot (voire à plusieurs mots) des informations d'ordre morphologique, grammatical, syntaxique voire sémantique (ex: "les poules du couvent couvent"). Certains systèmes d'étiquetage lexical intègrent une reconnaissance des mots composés (qui seront traités comme un mot simple par la suite).

² Notons qu'il existe des outils d'accentuation automatique, qui peuvent utiliser un étiquetage lexical des mots, ce qui permet, dans la plupart des cas de lever les ambiguïtés [Simard, 1996]. Certains outils utilisent un corpus d'apprentissage (comme [El-Beze, 1995]) ce qui permet de lever les ambiguïtés selon l'usage le plus fréquent. Exemple : la phrase "FRACTURE DE COTE" peut tout à fait être accentuée sous les deux formes : "Fracture de côté" ou "Fracture de côte", mais dans le domaine médical la seconde forme sera, de loin, la plus probable.

³ Lire à ce propos le chapitre "détection et correction d'erreurs" de [Laporte, 2000, p. 39].

⁴ Lire à ce propos l'article de [Bodenreider et Zweigenbaum, 2000] sur la détection de noms propres dans l'UMLS

.1.5. Segmentation de textes

Un programme d'indexation, s'il travaille sur de grands documents, pourra tenir compte des différentes unités d'indexation (unité linguistique) que sont la phrase, le paragraphe, ou le document dans son ensemble. Ce qui implique l'utilisation d'un programme de segmentation. Le plus simple (mais le moins performant) étant de reconnaître une phrase comme étant une suite de mots suivie d'un point. Lorsque l'on travaille sur des textes ayant un format défini, on peut parfois extraire la notion de phrase ou de paragraphe en analysant le format (SGML, XML, dans une moindre mesure HTML).

Une autre solution consiste à utiliser des étiqueteurs syntaxiques qui permettent, notamment, d'identifier les paragraphes et les phrases d'un texte.

L'indexation de textes est, le plus souvent, une étape préalable d'un système plus complet comme la recherche d'informations, l'attribution de mots-clés, la similarité de documents ou la synthèse automatique.

.I.6. Systèmes de Recherche d'Information (S.R.I.)

Un S.R.I. doit permettre d'extraire des documents correspondant à une question exprimée dans un langage plus ou moins contrôlé. L'utilisation habituelle est de permettre à un utilisateur d'écrire sa requête en langage libre et de lui proposer une liste de documents correspondants.

Les plus simples des S.R.I. se contentent de rechercher dans l'ensemble du corpus tous les mots contenus dans la phrase de l'utilisateur. Ceci peut même se faire sans indexation préalable (utilisation par exemple de la commande "grep" d'Unix). Cependant, les performances d'un tel système se heurtent très vite à la barrière linguistique et à la limite de la puissance de calcul de la machine.

Un véritable S.R.I. ne peut se passer d'une étape préalable d'indexation des documents. Outre l'indexation du corpus, un S.R.I. offre les fonctionnalités suivantes :

- 1) une interface d'interrogation
- 2) un processus de comparaison de la requête avec les documents indexés ("matching")
- 3) présentation des documents résultats (avec un score de similarité)

On pourra lire à ce propos le livre de [Baeza et Ribeiro, 1999].

Parmi les méthodes utilisées couramment dans la recherche d'information, citons les modèles booléens, vectoriels, probabilistes, ou encore des modèles spécifiques au langage naturel.

Fonctionnement d'un S.R.I. :

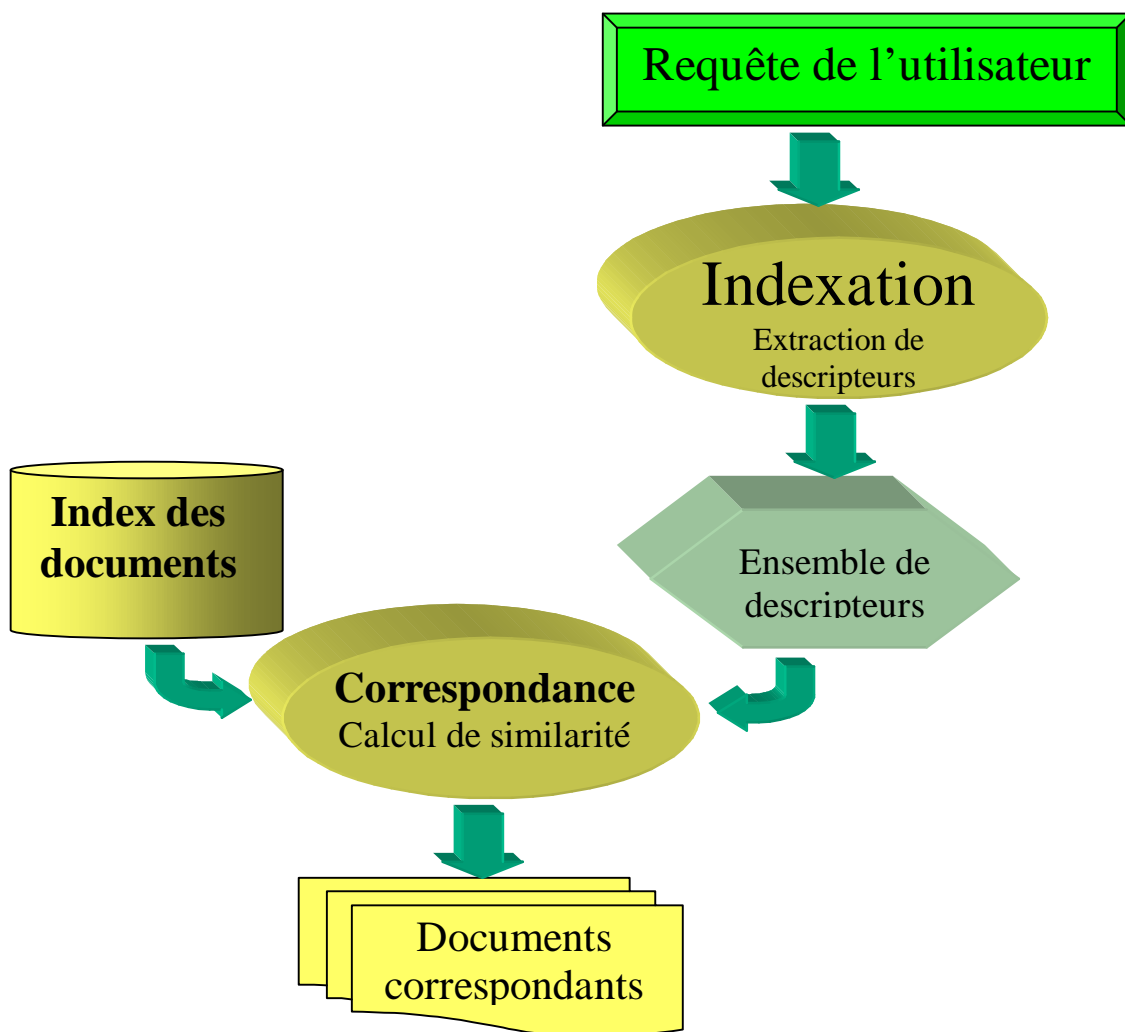


Figure 2 : Fonctionnement général d'un Système de Recherche d'Information

.I.7. Similarité de documents

Notons que la notion même de similarité est ambiguë, un document de 20 pages décrivant le cancer du foie est-il similaire à un document contenant simplement les trois mots "cancer du foie"?

Dans la logique booléenne, un document est similaire à une requête s'il contient tous les mots de la requête (opérateur "ET") ou s'il contient l'un des mots de la requête (opérateur "OU"). Cette logique booléenne ne peut plus être utilisée sur des documents entiers. Il ne s'agit pas de savoir si deux documents sont similaires ou non, mais plutôt de savoir quel est le score de similarité entre ces deux documents.

Pour des raisons de commodité, ce score de similarité sera le plus souvent exprimé en pourcentage (ou en probabilité, entre 0 et 1). Une propriété habituelle de la similarité peut être la symétrie (" a est similaire à b " alors " b est identiquement similaire à a "). Ce qui est le cas quand la similarité est le résultat d'un calcul mathématique de "distance", par exemple le nombre de mots communs, la distance euclidienne, la distance du CHI2, l'angle de deux vecteurs (espace vectoriel).

Quand un S.R.I. est capable de calculer des scores de correspondance entre une requête d'un utilisateur (une phrase) et un ensemble de documents, il est tentant de remplacer la phrase de recherche par un document dans son entier. Le résultat sera alors une liste de documents correspondant au premier avec, pour chacun d'eux, un score de similarité. La similarité de deux documents pourra être pondérée par la taille de chaque document (afin de favoriser les documents ayant des tailles similaires).

.1.8. Synthèse automatique de documents

La synthèse automatique de documents (ou résumé automatique) consiste, à partir d'un texte, à produire un texte plus court qui donne les informations principales contenues dans le document d'origine, ou permet, tout au moins, de s'en faire une idée [Jacquemin et Zweigenbaum, 2000]. On distingue principalement deux méthodes : Celles qui sont fondées sur la compréhension automatique de textes, utilisant les méthodes linguistiques et d'intelligence artificielle. Le principe consiste à "représenter" le texte sous forme de graphe, de réduire ce graphe en ne gardant que les nœuds les plus importants, et de régénérer un texte correspondant au graphe réduit. Les méthodes par extraction utilisent quant à elles des ressources linguistiques légères, elles consistent à sélectionner les unités textuelles (phrases, voire paragraphes) par calcul d'un score de similarité par rapport au document ou par rapport aux autres unités, et n'extraire que les unités les plus importantes.

La première méthode (appelée "méthode du bon élève") est plus proche du raisonnement humain, en théorie, elle produit de meilleurs résumés. Mais elle est confrontée aux problèmes habituels du TALN (Traitement Automatique du Langage Naturel) : elle requiert de telles connaissances linguistiques qu'elle est difficilement généralisable. La seconde ("méthode du mauvais élève") produit des résultats acceptables, mais sans plus. Le texte reconstruit peut être parfois incohérent (la cohésion des phrases est perdue). Mais elle ne requiert pas (ou peu) de connaissances linguistiques.

.1.9. Visualisation et classification de documents

Lorsque nous avons une fonction de similarité de documents, nous pouvons utiliser des outils de visualisation (cartographie) de documents, dont le but est de représenter sur un graphique synthétique (habituellement en deux dimensions) le contenu de notre corpus de textes. Ces outils permettent également de classer les documents, c'est à dire, en fonction de leurs distances respectives, essayer de regrouper des textes identiques dans des agrégats appelés "classes". Pour la cartographie, deux méthodes sont classiquement utilisées : les réseaux neuronaux, et les méthodes d'analyse factorielles de correspondances. Les résultats de ces deux méthodes sont classiquement relativement comparables aussi nous avons choisi de nous focaliser sur l'analyse factorielle des correspondances. Nous présentons ci-dessous le fonctionnement de cette méthode.

.I.9.1. Analyse Factorielle des Correspondances

C'est une méthode statistique permettant de comparer des données qualitatives. Cette analyse est utilisée, dans la plupart des cas, pour schématiser (sur plusieurs graphiques en deux dimensions) un tableau lexical, c'est-à-dire une matrice représentant, pour chaque mot, sa fréquence d'apparition dans chaque document (appelé aussi table de contingence, tableau croisé). Une autre utilisation consiste à classer les documents.

Nous avons choisi cette méthode pour l'appliquer non plus aux mots, mais aux concepts.

.I.9.1.a. Historique :

Benzécri a publié dès 1973 un ouvrage présentant cette technique d'analyse exploratoire de données multidimensionnelles [Benzécri et al., 1973]. Avant lui, on peut citer [Guttman, 1941] et [Hayashi, 1956] qui ont, tous les deux, publié des travaux relatifs à ce qu'ils appelaient les "méthodes de quantification".

.I.9.1.b. Principe général :

Dans un tableau de données, les méthodes factorielles permettent de calculer des distances entre chaque point-ligne (exemple: un mot ou un concept) et chaque point-colonne (exemple: un document).

Dans le cas d'un tableau lexical, nous pouvons exprimer un document comme étant un point dans un espace à n dimensions (chaque mot étant une dimension). De même, un mot peut être exprimé comme un point dans un espace à p dimensions (chaque document étant une dimension). L'interprétation d'un espace à plus de trois dimensions est difficile, voire impossible pour un humain, qui est plus habitué à lire des données dans un espace à deux dimensions (un graphique sur une feuille de papier par exemple).

Le principe de l'analyse factorielle des correspondances (AFC) est justement de faire une représentation de ces deux espaces (n et p dimensions) sur plusieurs graphiques à deux dimensions. Il va de soi, que beaucoup d'informations sont perdues quand on projette tous les points sur un espace plus réduit, mais la méthode permet justement de chercher les meilleurs plans de projection (ceux pour lesquels la perte est minimum). Comme un photographe cherche le meilleur angle de projection pour présenter en deux dimensions des objets du monde réel (en trois dimensions, voire quatre dimensions si l'on tient compte de l'axe temporel).

Pour ce faire, l'AFC cherche le premier axe (axe de covariance maximale) appelé axe 1, puis l'axe 2 (orthogonal au premier), etc... Jusqu'à arriver à une perte minimale d'information. Ainsi l'axe 1 et l'axe 2 forment le plan qui contient le plus d'information. Mais cela ne signifie pas que les autres plans (axe 1 et axe 3, axe 2 et 3 ...) ne doivent pas être pris en compte,. Ceci fait partie du travail d'interprétation ultérieur à l'AFC.

L'originalité de l'AFC est de "laisser parler les chiffres", sans établir a priori de classes bien établies. Comme le disent [Lebart et al., 1995] dans leur préface, grâce aux graphes d'analyse factorielle, "Benzécri a rendu les individus à la statistique; longtemps ignorés à force d'être confondus dans de vastes agrégats (...) les individus effectuent leur rentrée sur la scène statistique sous la forme de points dans un nuage". Dans l'utilisation que nous allons en faire, nous verrons que nos individus (les concepts et documents médicaux) apparaîtront chacun sur le graphique. Nous disposerons alors d'un outil extraordinairement synthétique pour visualiser notre corpus de documents.

.I.9.1.c. Comparaison avec le Latent Semantic Indexing

Remarquons que le "Latent Semantic Indexing" (LSI) fonctionne de la même manière. On projette les documents et les mots sur chaque axe, et, au lieu de représenter un document par les mots qu'il contient, on le représente par ses coordonnées sur chacun des axes.

.I.9.1.d. Exemple sur un espace à 3x3 dimensions :

L'exemple, fictif, représente le tableau de contingence des trois mots "foie", "rein" et "pancréas" dans trois documents, l'un parlant d'"hépatomégalie", l'autre de "diabète" et le troisième de "gastroentérologie". Par exemple, le document "hépatomégalie" contient quatre fois le mot "foie", deux fois "pancréas" et une fois "rein".

	Hépatomégalie	Diabète	Gastroentérologie
Foie	9	1	7
Rein	1	4	6
Pancréas	2	8	7

Table 1 : Tableau lexical d'un exemple simple

Chaque document s'exprime dans l'espace des mots, chaque mot est un point dans l'espace des documents.

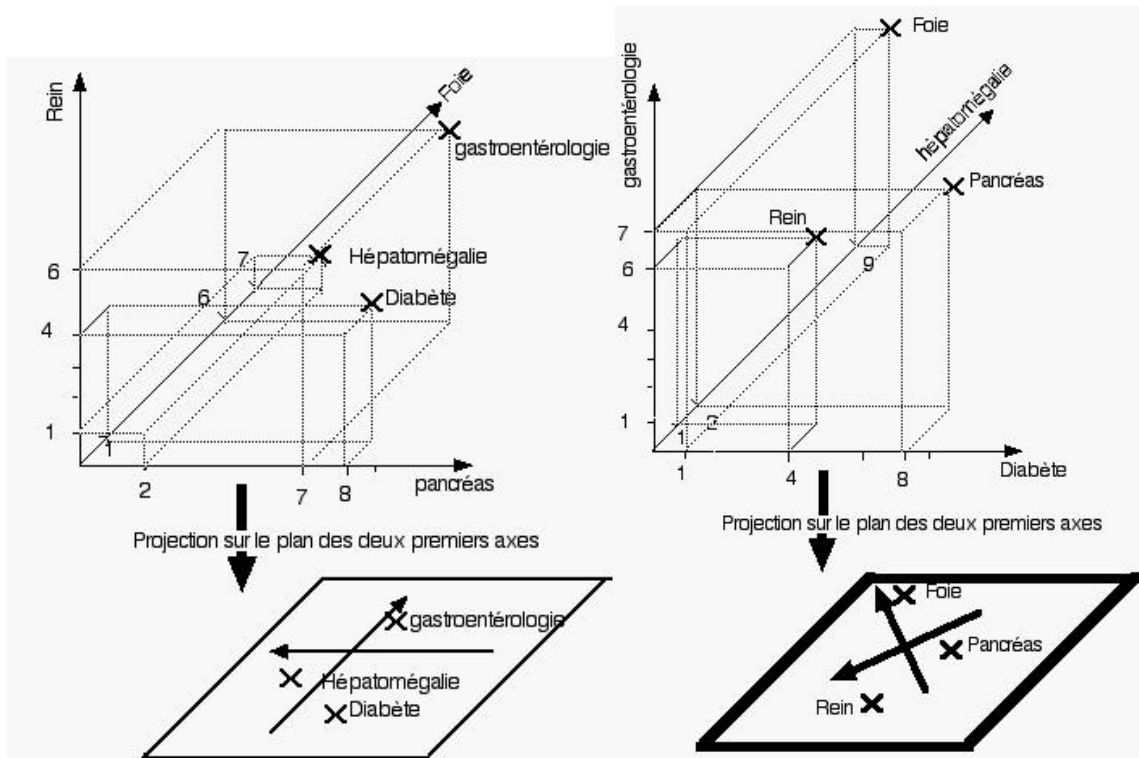


Figure 3 : AFC: exemple de projection sur un plan

Les deux projections (à un coefficient près) ont les mêmes barycentres (quasi-barycentriques), on prend l'habitude de représenter simultanément les deux projections :

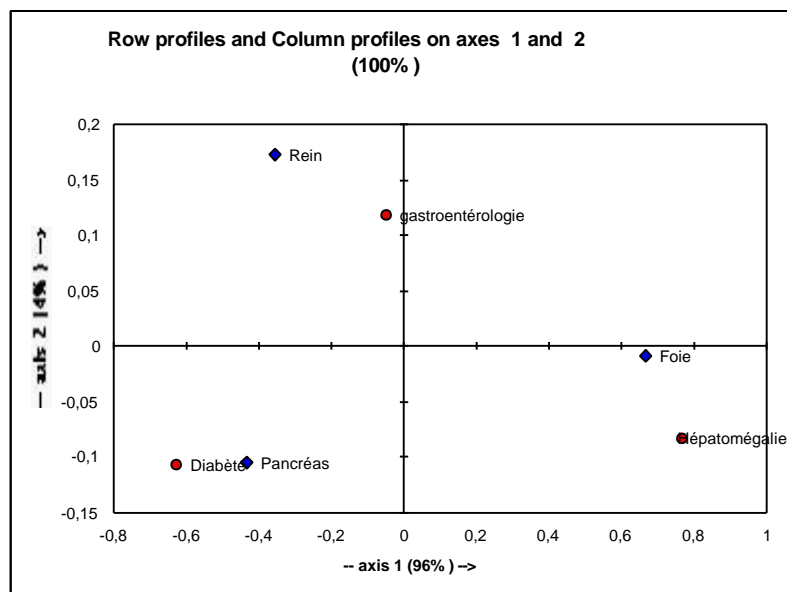


Figure 4 : AFC: Graphe résultant de la projection

.I.9.1.e. Interprétation

C'est la partie la plus délicate de l'analyse des correspondances. Le but est de présenter de manière synthétique l'information. Mais cela ne dispense pas l'utilisateur d'interprétation, ce qui nécessite une bonne compréhension théorique et pratique de la méthode et du domaine. Si, dans notre exemple, seuls deux axes sont nécessaires pour représenter toute l'information, il est très fréquent que les deux premiers axes n'en représentent qu'une partie infime. Il faut alors visualiser et essayer de "comprendre" les autres axes, ce qui ne peut être fait automatiquement.

Ce dernier graphique permet bien d'avoir une vue synthétique d'un espace très complexe. On peut interpréter, les relations existant entre les mots, entre les documents, et les relations mots-documents. Ainsi le document Hépatomégalie est "attiré" par le mot "Foie", "Gastroentérologie" est plus partagé (plus attiré par le mot "Rein" car, dans notre exemple, aucun autre document ne parle particulièrement de rein...).

L'analyse des correspondances ne se justifie pas pour ce genre de données (peu de documents, peu de mots). Mais son intérêt est justement de pouvoir traiter un très grand nombre de données. Inutile de "démontrer l'efficacité d'un filet de pêche dans un aquarium de salon" [Lebart et al., 1995].

La plupart du temps, l'analyse des correspondances est faite sur un découpage lexical des textes en mots, ce qui fait que l'on ignore délibérément de nombreuses informations de type sémantique ou syntaxique. On ne tient alors pas compte des synonymies ni des homonymies [Lebart et al., 1995, p 145]. Une solution envisageable est de recourir aux lemmes (toutes les flexions d'un mot sont ignorées et remplacées par le terme de référence). Ceci est un sujet de polémique: En effet, si deux formes fléchies sont réellement utilisées indifféremment, elles apparaîtront proches sur le graphique final, si, par contre, elles sont sémantiquement différentes, elles seront dissociées. Le mot "Cancer" au singulier ne signifie pas tout à fait la même chose que le mot "cancers" au pluriel... Nous verrons dans le chapitre "Analyse factorielle des correspondances (application aux concepts)(p. 126) une utilisation de cette méthode sur les concepts des documents, plutôt que les mots, ce qui permet d'enlever la barrière de la synonymie des termes.

.I.9.1.f. Outils utilisés

Trois outils ont été utilisés :

- QNOMIS II : [Kerbaol et al., 1997] Outil robuste, permet de visualiser de manière interactive le graphique et de sélectionner les axes.
- XLSTAT : un *shareware* qui est un ensemble de macros ajoutées au logiciel Excel©, et qui permet, de manière très simple, de faire une analyse des correspondances (et bien d'autres choses !) à partir d'un tableau croisé saisi sur ce tableur. Ne permet pas d'entrer beaucoup de données, ni de filtrer les résultats graphiques (par exemple : n'afficher que les points ayant une certaines contribution à la valeur propre)⁵.
- Citons également l'outil gratuit Rstat et son module CoCoAn [Ihaka et Gentleman, 1996]⁶.

⁵ URL : <http://www.xlstat.com/>

⁶ URL : <http://cran.r-project.org>

.I.10. Attribution de mots-clés

L'attribution de mots-clés consiste à "coder" des documents en leur assignant différents mots-clés choisis dans un thésaurus. En médecine, par exemple, chaque patient faisant un séjour hospitalier, aura son dossier codé par des actes et des diagnostics. L'attribution des mots-clés est le plus souvent manuelle (cas des séjours hospitaliers).

Parmi les méthodes automatiques d'attribution de mots-clés, on distingue essentiellement deux méthodes : par apprentissage ou par extraction.

La première méthode (statistique) calcule, à partir d'un corpus indexé manuellement, une matrice de co-occurrence de chaque mot avec chaque mot-clé. L'attribution de mots-clés à un nouveau texte consistera à calculer la probabilité d'apparition des mots-clés en fonction des mots du texte. Ces méthodes donnent de relativement bons résultats si le corpus est suffisamment important (qualitativement et quantitativement). On peut les utiliser sur différentes langues (par exemple : [Steinberger, 2001], [Steinberger et al., 2002]). Mais ces méthodes fonctionnent très mal sur des petits textes, encore moins sur des phrases (une phrase contient trop peu de mots pour que les co-occurrences aient une information sémantique pertinente).

La seconde méthode (linguistique) consiste à décrire chaque mot-clé par des termes (les différentes variantes linguistiques pouvant le représenter), et à essayer de reconnaître ces termes dans le texte. Cette méthode a l'énorme avantage de fonctionner sur de petites unités textuelles. La principale difficulté étant de définir la liste exhaustive des termes associés à chaque mot-clé. De plus, il faut recourir aux outils linguistiques pour reconnaître un terme quelle que soit sa forme syntaxique (par exemple: [Jacquemin et Tzoukermann, 1999]).

Les outils présentés fonctionnent plus ou moins bien selon les contextes. Il est indispensable alors de pouvoir les évaluer. C'est pourquoi nous présentons maintenant quelques méthodes d'évaluation classiquement utilisées en TALN.

.I.11. Méthodes d'évaluation

On distingue habituellement deux critères d'évaluation différents :

- Les critères quantitatifs : combien de documents peuvent être indexés, quel est le temps de réponse maximum à une requête ?
- Les critères qualitatifs : quelle est la pertinence des réponses ?

Les critères quantitatifs ne posent pas de problème, ils sont directement quantifiables. Nombre de Kilo-octets maximum autorisé, nombre de secondes ou millisecondes pour la réponse à une question de "taille" moyenne, ou le rapport temps de réponse en fonction du nombre de documents...

L'évaluation de la qualité d'un système de recherche d'information, peut se faire selon des critères subjectifs (réponses pertinentes, peu pertinentes, inadaptées, aberrantes...), mais, si l'on veut être le plus objectif possible, il faut trouver des métriques pour quantifier la pertinence des réponses [Fluhr, 2000] [Mizzaro, 1997].

Le principe consiste à répartir les documents en trois ensembles pour une requête donnée :

- Les documents correspondant réellement à la requête (pertinents)
- Les documents ne correspondant réellement pas à la requête (non pertinents) (ces deux premiers ensembles étant disjoints)
- Les documents retournés par le S.R.I. (retournés)

Ces trois ensembles peuvent être schématisés ainsi :

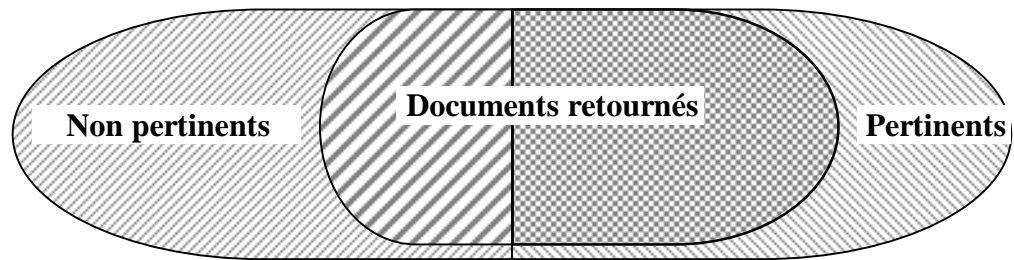


Figure 5 : Evaluation, classification des documents

On peut ensuite quantifier les résultats en termes de bruit et de silence :

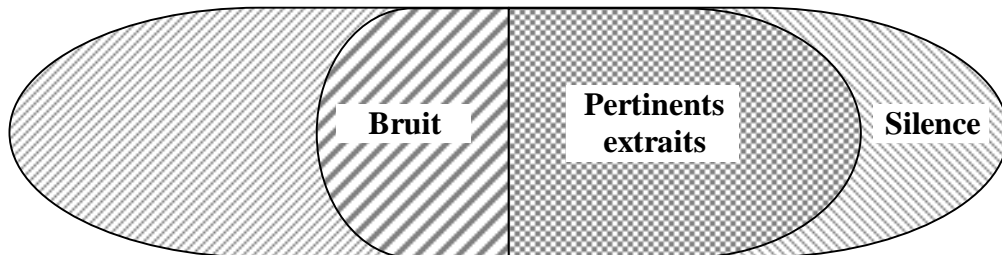


Figure 6 : Evaluation, bruit et silence

- Le bruit représente les documents extraits mais non pertinents

$$\text{bruit} = \frac{\text{Nombre de documents retournés et non pertinents}}{\text{Nombre de documents extraits}}$$

- Le silence représente les documents pertinents non extraits

$$\text{silence} = \frac{\text{Nombre de documents non retournés et pertinents}}{\text{Nombre de documents pertinents}}$$

Un S.R.I. sera d'autant meilleur que le bruit et le silence seront faibles. On pourra représenter le bruit et le silence en proportion du nombre de documents extraits.

Deux autres mesures sont souvent utilisées, qui sont en fait les compléments des deux précédentes :

- La précision qui représente le nombre de documents pertinents extraits par rapport au nombre de documents extraits

$$\text{précision} = \frac{\text{Nombre de documents retournés et pertinents}}{\text{Nombre de documents extraits}}$$

- Le rappel, qui représente le nombre de documents pertinents extraits par rapport au nombre de documents pertinents

$$\text{rappel} = \frac{\text{Nombre de documents retournés et pertinents}}{\text{Nombre de documents pertinents}}$$

La "F-measure" proposée par Van Rijsbergen [Van Rijsbergen, 1979] permet de conjuguer la précision et le rappel dans une formule :

$$F - measure = \frac{(rappel \cdot précision)}{a \cdot (rappel + précision)}$$

Si l'on donne la valeur 0,5 à **a** cela signifie que l'on donne la même importance à la précision qu'au rappel.

Un S.R.I. sera d'autant meilleur que la précision et le rappel seront forts (et par conséquent la "F-measure" sera forte).

Un schéma idéal d'une évaluation ressemblera à la figure suivante :

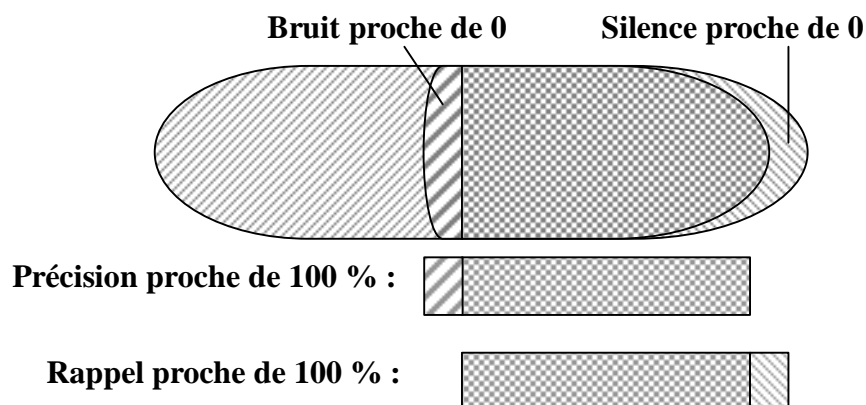


Figure 7 : Évaluation, schéma idéal

Par contre, l'évaluation d'une synthèse automatique de document ne peut être évaluée avec ces critères. Une méthode intéressante a été proposée dans le projet Tipster-Summac, l'évaluation des systèmes de résumés automatiques sont plus qualitatifs que quantitatifs, on cherche à mesurer les gains en temps et en qualité apportés par de tels systèmes [Mani et al., 1999].

Les différents outils abordés dans ce chapitre (indexation, SRI, ...) utilisent des ressources linguistiques. Les plus couramment utilisées sont les lexiques, les thésaurus et les corpus de textes.

.I.12. Ressources linguistiques

.I.12.1. Lexiques

Un lexique est un ensemble de mots répertoriés qui sera utilisé dans la phase d'indexation. Celui-ci pourra, par exemple, contenir toutes les flexions des mots afin de reconnaître un mot quelle que soit son orthographe.

Un lexique contiendra, au minimum, l'orthographe de chaque mot. On lui associera souvent des relations de flexions, de synonymies, de dérivations... Chaque entrée du lexique peut avoir une étiquette grammaticale (mot singulier, pluriel, nom propre, masculin, féminin, adverbe, ...). Il pourra également contenir une liste de mots composés, des relations sémantiques entre les mots. Si le lexique contient la définition de chaque mot, on parlera alors de dictionnaire.

Quelques exemples d'entrées de lexique :

Sclérodermie, nom féminin singulier, définition : "Dermatose caractérisée par l'épaississement avec induration de la peau et du tissu cellulaire sous-cutané et parfois des tissus profonds"

Flexion: Sclérodermies, Dérivation: Sclérodermique

Notons qu'il existe des lexiques informatiques que peuvent utiliser les outils linguistiques (exemple: Wordnet [Miller, 1995] et Eurowordnet [Bloksma et al., 1996])

.I.12.2. Thésaurus

Un thésaurus peut être défini comme un répertoire de termes normalisés. Le plus souvent le thésaurus contient aussi un réseau sémantique reliant ces termes (synonymes, termes reliés, termes spécifiques, termes génériques).

La définition de "terme" est ambiguë, Felber le définit comme "représentant linguistique d'un concept dans un domaine de connaissance" [Felber, 1987]. L'utilisation de "terme" en indexation est parfois abusive, et se confond avec la notion de "descripteur" (unité d'indexation - mot ou unité lexicale). Bourigault et Jacquemin le définissent comme le "résultat d'un processus d'analyse terminologique", qui permet de définir un mot ou une unité

lexicale comme terme par décision humaine, selon son utilisation dans le domaine (par extraction semi-automatique à partir d'un corpus de textes par exemple) [Bourigault et Jacquemin, 2000].

Un terme se situe entre les concepts et les mots (le concept de "rachialgie", peut s'exprimer par les termes "Rachialgie" ou "Douleur rachis", eux-mêmes s'exprimant par des mots différents). Il est le représentant linguistique d'un concept, un concept peut s'exprimer par des termes synonymes dans une même langue. Dans un thésaurus multilingue, un concept sera représenté par des termes dans chaque langue.

Un thésaurus médical est une représentation d'un ensemble de concepts médicaux. Chaque concept est décrit par un ou plusieurs termes synonymes. Ces concepts sont, dans la plupart des cas, organisés de manière hiérarchique (mono-axiale ou multi-axiale...) et parfois de manière sémantique. La synonymie des termes est en générale relative ("gratouiller" pourra être synonyme de "chatouiller", parfois les deux termes désigneront des concepts différents), "des dénominations différentes se distinguent souvent par une nuance que l'on peut négliger dans un contexte fixé, mais qui pourrait avoir son importance dans un contexte particulier" [Zweigenbaum, 1999].

Dans la suite de notre étude, nous utiliserons des thésaurus existants (MeSH et ADM), mais des outils de constitution de thésaurus existent. Ils se basent sur un corpus de textes existant, et, par des méthodes statistiques, linguistiques ou mixtes extraient des "candidats termes" qui sont ensuite validés par un opérateur afin de constituer un thésaurus. On pourra se reporter au chapitre "construction de ressources terminologiques" de [Bourigault et Jacquemin, 2000], ou aux travaux de [Sébillot et Pichon, 1997]. Parmi les logiciels les plus importants d'acquisition de ressources terminologiques citons :

- Termino [David et Plante, 1990] : Considéré comme le premier logiciel d'acquisition terminologique, développé à l'Université de Québec (Montréal), il effectue une analyse syntaxique du corpus et soumet chaque candidat terme à l'utilisateur via une interface de validation.
- "Ana" de Chantal Enguehard [Enguehard et Pantera, 1995] procède par analyse statistique à partir d'un corpus de textes et d'un ensemble de termes existant. Il consiste à essayer de trouver, dans le corpus, des associations récurrentes de deux termes, ou d'un terme et d'un mot.
- "Acabit" de Béatrice Daille [Daille, 1994], pour la construction de lexiques multilingues. Qui procède par une analyse linguistique du corpus (extraction de

séquences nominales, regroupement en termes binaires), puis par un filtrage statistique (pour ne proposer que les candidats termes les plus fréquents).

- "Lexter" de Didier Bourigault [Bourigault, 1996] écrit dans le but de mettre à jour un thésaurus utilisé dans un système d'indexation automatique. Il extrait du corpus les syntagmes nominaux maximaux, puis découpe récursivement ces syntagmes maximaux, pour construire un réseau de candidats termes.
- "Xtract" de Smadja [Smadja, 1993] est un extracteur de collocations, par des méthodes statistiques, il extrait les couples de mots anormalement fréquents, (avec réitération sur les nouvelles unités trouvées), et classe les collocations résultantes selon différentes formes linguistiques.

.I.12.3. Corpus de textes

L'indexation se fait en général sur un "ensemble" de textes, appelé corpus de documents. Ce corpus peut être local (ensemble de rapports, de courriers, ...), ou extérieur (CDROM, documents téléchargés sur Internet...).

Plus le corpus de texte est homogène, plus le vocabulaire sera homogène, et plus le processus d'indexation sera performant (car il y aura moins d'ambiguïtés linguistiques à résoudre). Notons qu'un autre usage du corpus de texte est d'en extraire automatiquement de l'information, par exemple, extraire l'information lexicale [Zweigenbaum et Grabar, 2000], [Sébillot et Pichon, 1997], la sémantique [Habert et al., 1997], [Sébillot et Pichon, 1999] ou pour construire une ressource terminologique [Bourigault et Jacquemin, 2000].

De plus en plus d'outils de TALN utilisent les textes disponibles sur Internet. Le web est, à ce jour, la base de connaissances la plus étendue au monde. L'inconvénient étant de devoir intégrer un "aspirateur web" ("crawler", "web agent"), qui, à la demande, télécharge automatiquement un document. Ces documents sont hétérogènes (au niveau du format et du contenu), ce qui ne facilite pas la tâche des outils de traitement automatique du langage. Mais le groupe de travail "W3C Semantic Web"⁷ commence à définir des standards permettant justement à des outils automatiques d'extraire l'information sémantique des documents disponibles sur le web [Berners-Lee et al., 2001].

⁷ <http://www.w3.org/2001/sw>

Tous ces outils de TALN ont été testés dans notre domaine, nous présentons maintenant le contexte de notre travail (et notamment les ressources dont nous disposons).

.II. Contexte

Nous présentons ici la base de connaissances ADM et, plus spécifiquement, le lexique et le thésaurus qui en ont été extraits. Puis nous présenterons le meta-thésaurus UMLS, et plus spécifiquement le MeSH. Enfin nous présenterons quelques outils qui utilisent déjà ces connaissances.

.II.1. La base de connaissances ADM

Le projet "Aide au Diagnostic Médical" (ADM) est né au Centre Hospitalier Universitaire de Rennes à l'instigation du professeur Lenoir [Lenoir et al., 1981]. Il avait deux objectifs principaux : Aider les médecins à porter des diagnostics et leur offrir cette information via les réseaux télématiques. L'ADM a été le premier service télématique français à destination des médecins. Cette base de connaissances a maintenant plus de vingt ans.

Ce projet repose sur une base de données (maintenant représentée dans un modèle relationnel sur le SGBDR Oracle©) constituée principalement de pathologies et de symptômes. Les pathologies décrites (plus de 10 000 maladies, syndromes et formes cliniques) couvrent tous les domaines de la médecine. Ces descriptions utilisent plus de 100 000 signes ou symptômes. Les concepts ADM (signes et pathologies) sont liés dans une hiérarchie (ou on ne différencie pas la méronymie et la taxonomie), qui représente la relation hyperonyme - hyponyme ("terme générique" - "terme spécifique"), cette hiérarchie possède environ 40 000 entrées. Du point de vue de l'utilisateur, l'ADM offre principalement deux fonctionnalités :

- Consulter la description d'une pathologie (liste des signes et symptômes apparaissant dans une maladie ou syndrome).
- Connaître toutes les pathologies contenant un signe donné

Une troisième fonctionnalité expérimentale est de proposer des diagnostics en fonction d'un ensemble de symptômes.

La Figure 8 montre un exemple d'interrogation sur la pathologie "fente du palais". En premier lieu le système propose (via le dictionnaire ADM) toutes les pathologies correspondant à la question. L'utilisateur choisit l'une d'elles (ici "division vélo-palatine isolée") pour accéder à sa description (les symptômes apparaissant dans cette pathologie). Cette maladie est ici décrite par la maladie "toxoplasmose congénitale" comme antécédent, l'utilisateur peut consulter cette autre pathologie.

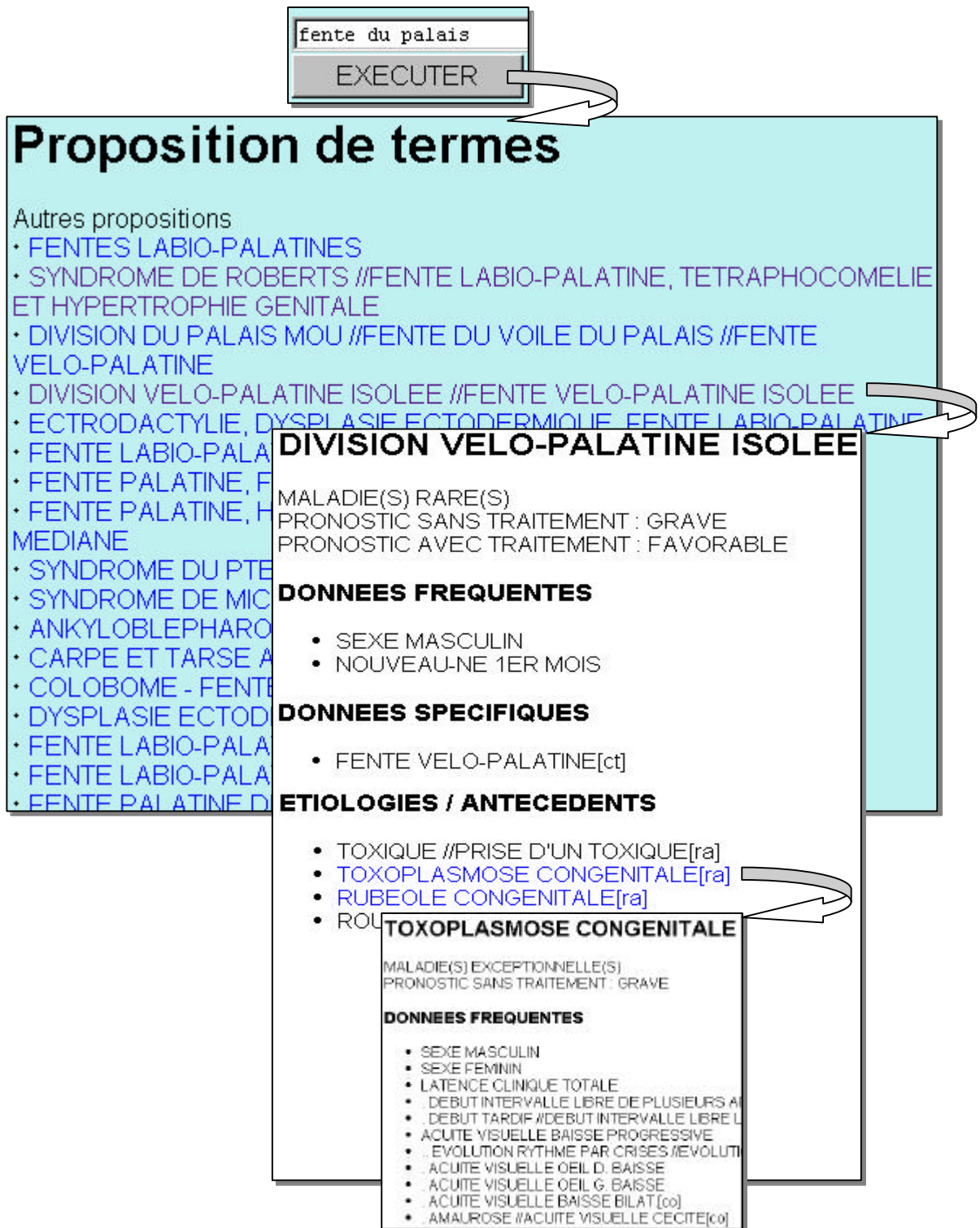


Figure 8 : Exemple d'interrogation de l'ADM

Cette base de données est, quantitativement, une des plus grandes bases généralistes du monde (les autres bases : DXPLAIN [London, 1998], ILIAD [Murphy et al., 1996], QMR [Arene et al., 1998] ne contenant pas plus de 2000 pathologies décrites [Cleret et al, 2001]). Les informations contenues dans cette base de données ont été décrites par des étudiants de médecine au cours de leur thèse, chaque pathologie a ensuite été validée par un expert du domaine.

Initialement développée pour le Minitel, l'interface a ensuite été adaptée pour le Web, en utilisant la méthode CGI (dès 1994, historiquement, la première base de données médicale accessible sur le Web) [Pouliquen et al., 1995]. La technologie hypertexte du Web est particulièrement intéressante dans ce domaine car une pathologie peut être décrite par un syndrome, qui, lui-même, possède sa propre description.

Actuellement, une version hypermédia est disponible gratuitement sur Internet ⁸ [Séka et al., 1995], [Fresnel et al., 1996]. Il s'agit d'un service de type documentaire, l'utilisateur pouvant obtenir des réponses à des questions type (exprimées en langage naturel). Un module d'évocation diagnostique est disponible (permet de lister, par ordre de pertinence, les pathologies pouvant engendrer une liste de symptômes donnés). Ce dernier module est accessible exclusivement en Intranet, en effet, des améliorations sont indispensables avant de pouvoir le diffuser.

Une partie de ce serveur est réservée aux professionnels de santé, mais la majeure partie reste accessible au grand public. À l'heure actuelle, près de 10 professionnels nouveaux par jour s'inscrivent sur l'ADM, et près de 3000 requêtes sont faites sur ce serveur chaque jour⁹.

⁸ A l'adresse <http://www.med.univ-rennes1.fr/htbin/adm/reponse.pl>

⁹ Les statistiques sont disponibles à l'adresse http://www.med.univ-rennes1.fr/stat/med/stat_actuelle/projet/index.html (pour le projet "adm")

L'ADM a entraîné trois autres sous-projets :

- ATM : Afin de couvrir la médecine dans son ensemble, les effets indésirables de médicaments ont été insérés dans la base de données (ainsi que les intoxications), ce qui a généré un projet annexe : l'Aide à la Thérapeutique Médicale (ATM) qui décrit les médicaments : formes commerciales, substances (DCI), classements ... [Riou et al., 1999].
- AEDM : Utilisant cette base de données, un autre projet annexe a vu le jour : L'Aide à l'Enseignement du Diagnostic Médical (AEDM) [Meadeb et al., 1986], [Riou et al., 1990], [Riou et al., 1994], qui propose une simulation de cas cliniques basée sur les connaissances de l'ADM. L'originalité de cet EAO est de travailler en texte libre grâce au dictionnaire ADM. Une version multimédia de ce système a été développée pour le Web [Pouliquen et al., 1995]
- Dictionnaire ADM : Le nombre de pathologies, de signes, de symptômes (que nous appelons "concepts ADM") est tel qu'il n'est pas envisageable de sélectionner un concept dans un menu. Il s'est donc avéré indispensable de pouvoir accéder aux concepts par un moteur de recherche. Les mots composés, les synonymies, dérivations et flexions de mots ont donc été stockés dans un lexique, que nous appelons : "le dictionnaire ADM" (mal nommé, car il ne contient pas de définitions de mots).

Ce dernier travail est particulièrement intéressant pour l'indexation, car il permet de décomposer un terme ADM en mots. Le langage de description des termes ADM est dit semi-naturel ou laconique car sa syntaxe est pauvre et l'usage du groupe verbal est très rare.

De cette base de connaissances ADM, nous avons extrait le "dictionnaire ADM", et le "thésaurus ADM".

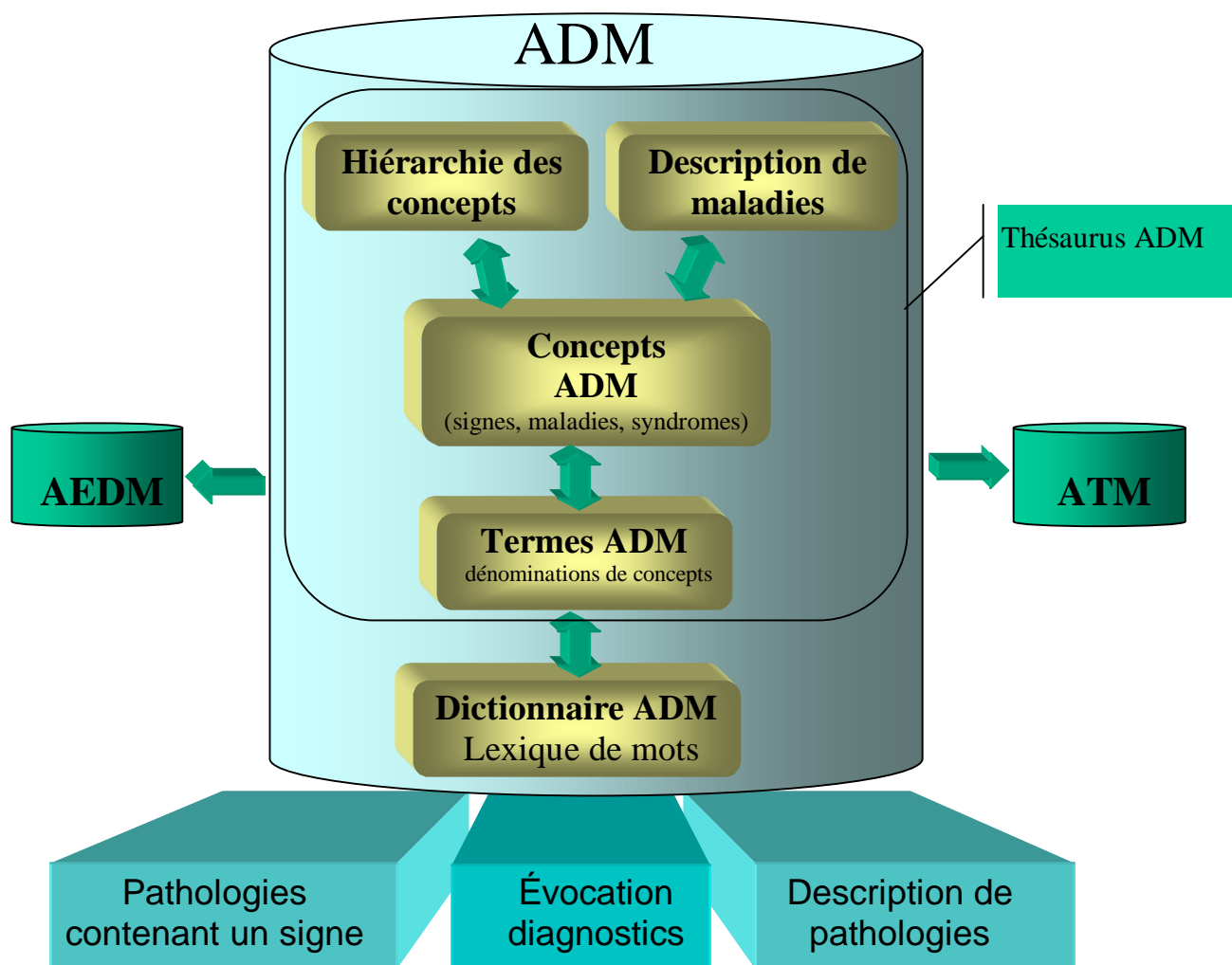


Figure 9 : Fonctionnalités essentielles de l'ADM

.II.1.1. Dictionnaire ADM :

Ce lexique, créé manuellement dans un premier temps d'après plusieurs dictionnaires médicaux, a, par la suite été mis à jour après alimentation de nouvelles entrées (les mots apparus dans les nouvelles descriptions de pathologies ADM). Notons que nous le dénommons historiquement "dictionnaire" même s'il s'agit plutôt d'un "lexique".

La base de données ADM date du temps où l'informatique ne traitait pas les caractères diacritiques. Toutes les chaînes de caractère ont été écrites en majuscules, sans accents. Les entrées du dictionnaire ADM sont donc écrites en majuscules (non accentuées) exclusivement. Ce qui n'est pas sans poser quelques problèmes d'homographie ("côte" et

"côté", ou "aine" et "âiné" par exemple). Mais les deux thésaurus sur lesquels porte notre étude (ADM et MeSH français) sont tous deux non accentués.

Afin de limiter le "silence" dans les requêtes d'interrogation, les mots ont été regroupés en "familles". Chaque famille est décrite par son "mot de référence" (terme vedette), puis sont décrits, au premier niveau, tous les synonymes, quasi-synonymes ou dérivations de ce mot de référence. Au second niveau sont décrites toutes les flexions des mots.

Ce qui permet de regrouper les flexions de mots "COEUR" et "COEURS", ainsi que les synonymes ou dérivations "COEUR", "COEURS", "CARDIAQUE" et "CARDIAQUES". La distinction est faite entre les flexions, et les synonymes¹⁰.

Exemple de famille de mots :

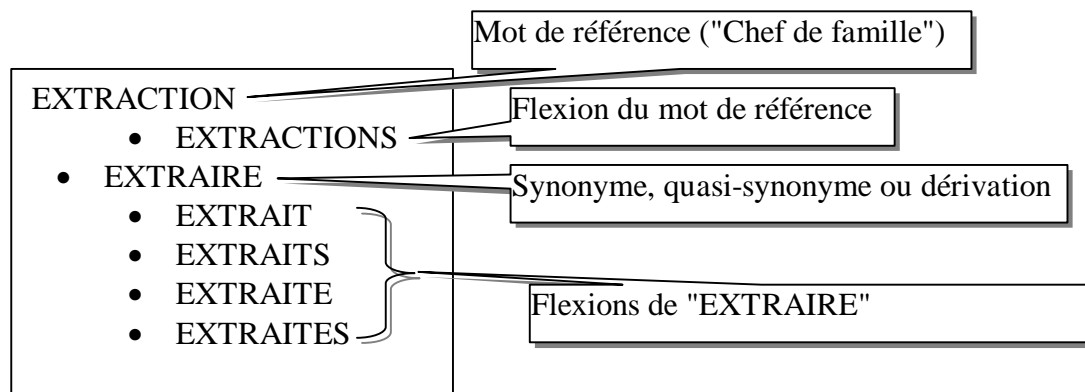


Figure 10 : Exemple de famille de mots dans le dictionnaire ADM

Par contre, le lexique ne possède aucune étiquette grammaticale. Nous n'avons aucune information sur le genre d'un mot, ce qui est une limite linguistique très importante.

.II.1.1.a. Mots nuls :

Appelés aussi "stop-word", ce sont des mots n'ayant pas beaucoup d'importance sémantique dans une phrase (comme les articles, ou les conjonctions de coordination, "LE", "A", "ET", "OU"...). Ils sont retirés automatiquement de la liste des mots d'une phrase. Le dictionnaire ADM, dans son état actuel, n'en comporte que quatorze : de, a, en, aux, la, l, des, au, le, du, les.

Les portions de phrase suivantes seront codées de la même manière :
"fièvre", "de la fièvre", "la fièvre", "à la fièvre", "une fièvre"...

.II.1.1.b. Mots multiples :

Nous appelons mot multiple une entrée du lexique qui contient plusieurs mots (par opposition à mot simple, une entrée ne comportant qu'un seul mot), on parle aussi d'unités multimots ou d'expressions figées. Dans le dictionnaire ADM, la distinction est faite entre deux types de mots multiples, les *mots composés* et les *mots associés*.

.II.1.1.b.a. Mots composés :

Il s'agit d'un ensemble de mots dont la signification est différente (voire éloignée) de l'union des significations de chacun des composants.

Des mots comme "pomme de terre", "fer à cheval", "chemin de fer", "angine de poitrine", "bec de lièvre" ou "fièvre jaune" sont des mots composés typiques.

Il apparaît important de les traiter comme un seul mot pour deux raisons :

1. L'interrogation sur un seul des composants ne doit pas proposer les phrases comportant ce mot composé (il est amusant, mais souvent agaçant, de se voir proposer des documents parlant de pommes de terre quand on interroge sur "pomme" dans les moteurs de recherche habituels).
2. Il faut pouvoir gérer des synonymies avec ces mots composés ("angor" est synonyme de "angine de poitrine", ou "fièvre jaune" synonyme de "amarile", "bec de lièvre" quasi-synonyme de "fente palatine").

Un module de reconnaissance de ces mots composés est nécessaire. Inutile de s'aventurer dans la reconnaissance d'un mot composé s'il contient des tirets entre chacun de ses composants, d'une part, cette écriture n'est pas toujours respectée, d'autre part certains des mots composés ne sont jamais écrits avec des tirets ("fièvre jaune" par exemple).

On reconnaît un mot composé dans une phrase si :

1. Tous les composants (ou l'une de leurs flexions) sont présents dans la phrase
2. Ils apparaissent dans le même ordre
3. Ils sont contigus (à l'exception de mots nuls)

¹⁰ Dans tout le document nous utilisons "*synonyme*" pour désigner un synonyme, quasi-synonyme ou dérivation d'un terme

Ainsi le mot composé "POMME DE TERRE" ne sera pas reconnu dans les phrases suivantes :

- La pomme de l'arbre sur la terre. (*Non contigu*)
- La terre des pommes. (*Ordre des composants non respecté*)
- La pomme de terrain. (*On n'accepte que les flexions des composants, pas les synonymes ou dérivations*)
- Mettre la pomme sur la terre. (*Tous les composants ne sont pas présents*)

Mais ce mot composé sera reconnu dans les phrases suivantes :

- La pomme de terre
- Les pommes de terre

Il est important de noter qu'un mot composé peut contenir un mot nul (voire plusieurs), dans ce cas, le mot nul est aussi important que les autres composants (exemples: "Vitamine A", "Hépatite A"...).

Usage des mots composés :

Aucun programme automatique de reconnaissance automatique de nouveaux mots composés n'a été utilisé (au mieux, seul un programme de proposition de mots composés pourrait être utilisé ou développé, par exemple: [Habert et Jacquemin, 1993]). Ils ont tous été entrés manuellement. Ces mots composés ont été créés pour quatre usages différents :

1. Sémantique manifestement différente : "queue-de-cheval", "angine de poitrine"...
2. Synonymie d'un mot simple avec un mot multiple : "fièvre jaune" (synonyme de "amarile")
3. Permettre l'exclusion d'un des composants : la création du mot composé "herbe à fièvre" (désignation d'une plante) permet d'exclure le mot "fièvre" de la phrase. Ou, pour gérer certaines négations (exemple : "NON STEROIDIEN"¹¹).
4. Différenciation des homographes : "AINE FAMILLE" (différenciation entre "Aîné" et "Aine"), "A COTE DE", "DES DEUX COTES" "POINT DE COTE" permettent d'éviter de reconnaître le mot "côte" dans des phrases où l'on parle de "côté" (dans notre lexique, le mot "COTE" désigne l'os).

¹¹ La création de ce mot composé évite de reconnaître le concept "anti-inflammatoires stéroïdiens" dans la phrase "anti-inflammatoires non stéroïdiens"

.II.1.1.b.b. Mots associés :

Il s'agit d'une extension de la définition des mots composés. Les mots associés sont des mots multiples dont la signification est l'union des significations des composants. Contrairement aux quatre usages des mots composés, les mots associés n'existent que pour créer une synonymie entre un mot simple et un mot multiple (voire entre deux mots multiples).

Nous voulions, par exemple, pouvoir gérer la synonymie entre le mot " Odontalgie" et "douleur dents", or la définition du mot composé est trop restrictive, car le mot composé "douleur dents" n'aurait pas été reconnu dans les phrases suivantes :

"Douleur dentaire", "Mal de dents", "Maux dentaires" (Synonyme)

"Douleur intense aux dents" (mots intermédiaires)

"(...) au niveau des *dents*, une *douleur* intense (...)" (intersion des composants)

Or nous voudrions reconnaître ce mot multiple dans des phrases du type :

"Il se plaint que sa dent lui fait mal" (Synonyme, mots intermédiaires et intersion).

Des "mots associés" ont donc été créés avec la définition suivante :

"Le mot associé sera reconnu dans une phrase si chacun de ses composants (ou l'un des synonymes) est présent".

Cette définition simple (voire simpliste) permet effectivement de reconnaître le mot associé dans la plupart des cas, mais entraîne aussi beaucoup de "bruit" (mot reconnu de manière erronée), par exemple, le mot associé "avant repas" (synonyme de préprandial) sera reconnu dans les phrases suivantes:

Il mange avant de dormir (ordre non respecté)

Avant d'aller dormir, il mange (beaucoup de mots intermédiaires)

Mais, par contre, le mot associé "PRE PRANDIAL" (synonyme, lui aussi, du mot simple "PREPRANDIAL") sera reconnu dans la phrase suivante :

"Hypoglycémie pré ou post prandiale"

.II.1.1.b.c. Préfixe, suffixes et dérivation de mots :

Le langage médical a la faculté originale d'agglutiner les préfixes et les suffixes (et créer ainsi beaucoup plus de mots dérivés que le français courant), par exemple "ANTIRETROVIRALES". Les mots composés savants sont plus nombreux également, il n'est pas rare de rencontrer des mots comme : "Oesogastroduodéal" ou "oesogastroduodénoscopie".

(Ainsi, à titre anecdotique, le mot le plus long de notre lexique bat le record de "anticonstitutionnellement" : "ELECTROSTEREOENCEPHALOGRAPHIQUES" qui a 32 caractères ! Signalons également, dans le MeSH des noms de molécules comme "DICHLORORIBOFURANOSYLBENZIMIDAZOLE" qui fait 34 caractères).

A l'heure actuelle, le lexique décrit les compositions de mots les plus fréquentes par un mot associé équivalent : "Hypercalcémie" est synonyme du mot associé "Calcémie élevée", de même, "Hypocalcémie" est synonyme du mot associé "Calcémie basse".

.II.1.2. Thésaurus ADM

De la base de connaissances ADM, nous avons pu extraire un véritable thésaurus contenant les concepts (pathologie, symptôme), les termes (l'ADM contient les diverses terminologies utilisées pour chaque maladie ou signe), et même un réseau sémantique (comprenant essentiellement une taxonomie, que nous appelons "hiérarchies de concepts", et la relation exprimant qu'une pathologie est décrite par un symptôme).

La Figure 11, montre un exemple de contenu du thésaurus ADM.

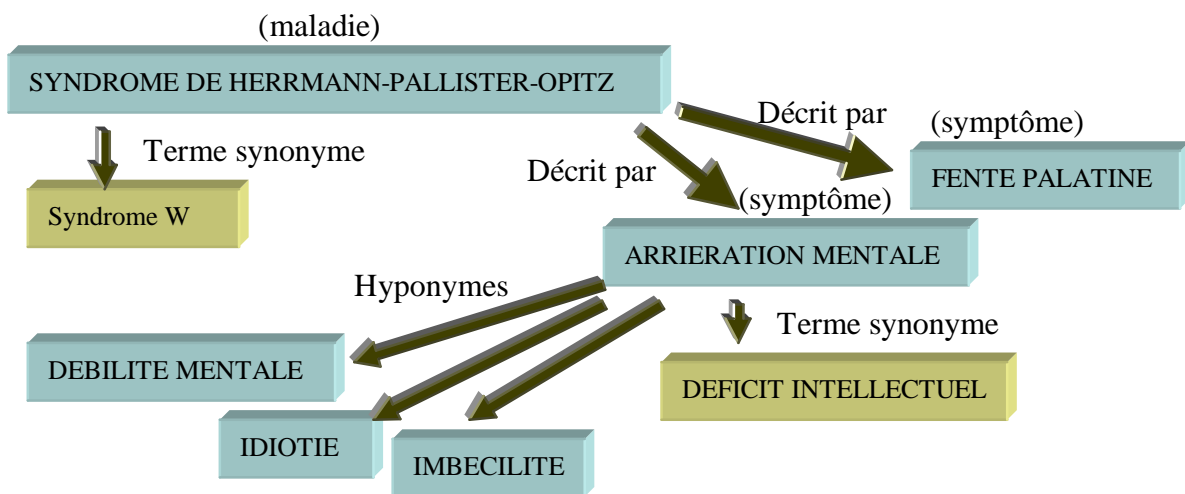


Figure 11 : Exemple de contenu du thésaurus ADM (concepts, termes, hiérarchie et relation sémantique)

.II.2. UMLS

Le projet UMLS (Unified Medical Language System) a été lancé en 1986 par la U.S. National Library of Medicine (NLM). Le but était de créer un thésaurus des termes utilisés dans tous les domaines de la médecine, en utilisant le plus grand nombre possible de terminologies ou de classifications médicales. Ces termes sont ensuite regroupés en concepts (un concept est un "cluster" de termes) et organisés dans un seul système : le meta-thésaurus UMLS [Lindberg et al., 1993]

Une équipe de la NLM est chargée de superviser et de mettre à jour cette base de connaissances, et le lexique qui y est associé ("Specialist lexicon"). Une nouvelle version de l'UMLS est diffusée chaque année via le Web ou par CD-ROM. Précisons que le but de la NLM est la mise au point de cette terminologie et non son exploitation. Exception faite de la recherche dans la base de données MEDLINE sur le site de recherche bibliographique PUBMED¹².

Ce meta-thésaurus est constitué d'un ensemble de concepts biomédicaux, ainsi que de leurs informations sémantiques. Ces informations sémantiques sont elles-mêmes hiérarchisées dans un réseau sémantique, qui permet d'organiser les types et les catégories sémantiques des concepts. De plus, l'UMLS contient un lexique spécialisé qui décrit les informations syntaxiques des termes utilisés.

On parle de meta-thésaurus, car il s'agit bien d'une base de connaissances regroupant plusieurs thésaurus médicaux, parmi lesquels :

- Le MeSH ("Medical Subject Headings"), lui aussi édité par la NLM, dont le but premier était d'indexer les références bibliographiques biomédicales [MeSH, 1986].
- L'ICD-10-CM ("International Classification of Diseases", 10e édition), CIM-10 en français [CIM 1977]
- La CPT ("Current Procedural Terminology") de l'association Médicale Américaine (A.M.A.) [CPT, 1996]
- La DSM III R ("Diagnostic and Statistical Manuel of Mental Disorders") de l'association psychiatrique Américaine,
- Le SNOMED ("Systemised Nomenclature of Medicine") [Cote, 1995]

- L'index de radiologie ACR ("American Society of Radiology")

On peut donc considérer l'UMLS comme une représentation conceptuelle du langage médical, chaque concept extrait des différents thésaurus a un code unique.

D'un point de vue technique, le meta-thésaurus UMLS peut être représenté dans une base de données ayant la structure suivante :

- Le concept : identifié par un code unique appelé "code UMLS" et un terme appelé "terme préférentiel" (par opposition aux termes synonymes)
- Les types sémantiques de chaque concept
- Les variantes lexicales

Son utilisation :

- Transcodage automatique [Cimino et al., 1993] [Burgun et al., 1992]
- Accès aux bases de données [Fieschi et Joubert, 1994]
- Traitement du langage [Mc Cray et al., 1993] et indexation [Aronson et al., 2000].
- Constitution de treillis de concepts [Volot et al., 1993]
- Base de connaissances pour décrire des données médicales [Burgun et al., 1996] [Bodenreider et al., 1998]
- Enrichissement de thésaurus [Le Duff et al., 2000], enrichissement de connaissances linguistiques [Zweigenbaum et Grabar, 2000]
- Etiquetage sémantique de textes [Ruch et al., 1999]
- Classification de documents [Bodenreider, 2000]

La Figure 12 représente un extrait du meta-thésaurus UMLS, extrait de [Le Beux et al., 1995].

¹² <http://pubmed.gov>

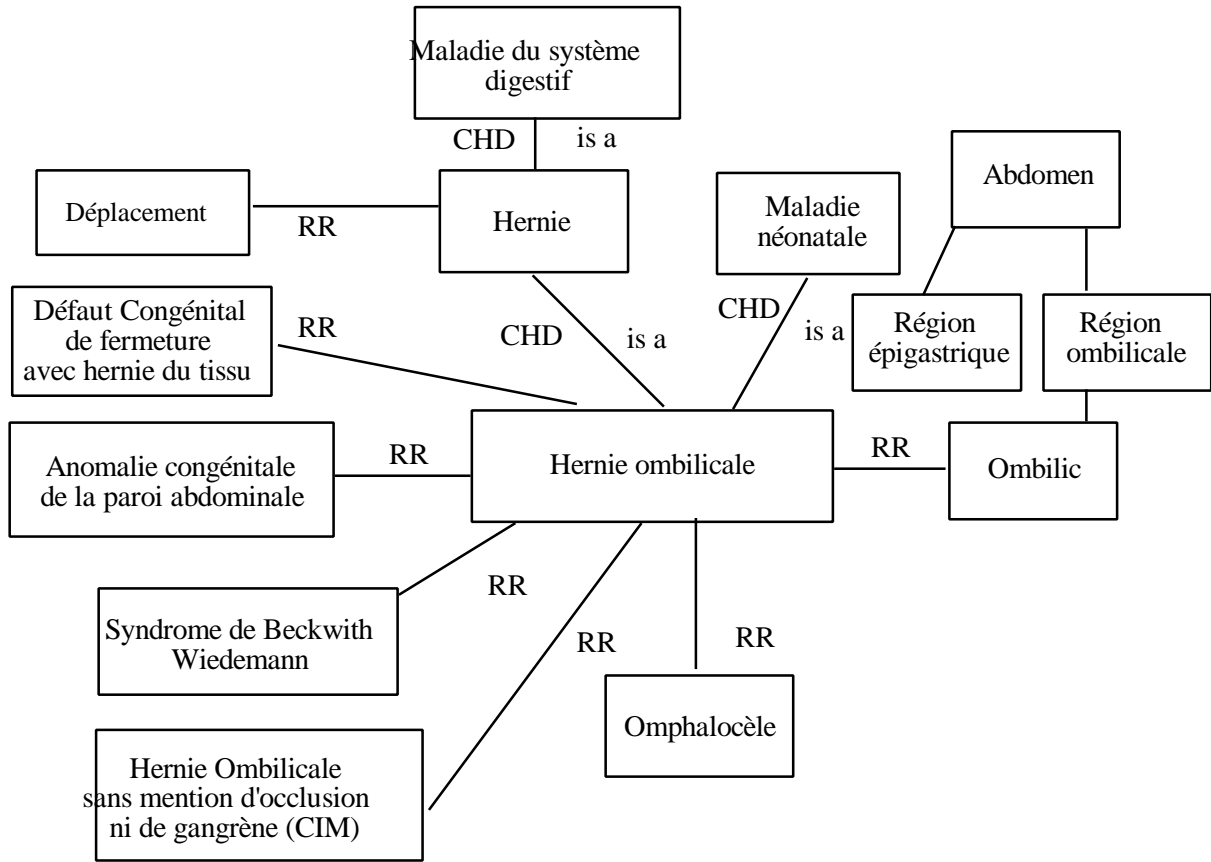


Figure 12 : Un réseau sémantique basé sur UMLS

Multilinguisme :

Le MeSH a été traduit en français par l'INSERM. Le meta-thésaurus possède donc ces traductions. Nous avons un sous-ensemble (non négligeable) du meta-thésaurus qui est disponible en version française. Le meta-thésaurus étant en constante évolution, peut-être disposerons-nous un jour de la traduction française de la plupart des concepts de l'UMLS.

Notre but étant d'indexer des textes en français, notre système d'indexation se basera uniquement sur le MeSH traduit en français.

"Le MeSH a été construit et utilisé essentiellement dans un but documentaire et ne prétend pas décrire toute la richesse et la complexité du langage médical. Il a cependant suivi, de manière continue, l'évolution des sciences et techniques biomédicales et constitue, de ce fait, un corpus incontournable, bien maintenu et utilisant des concepts clairement définis." [Le Beux et al., 1995].

Par la suite nous parlerons de NOMINDEX-MeSH, un outil permettant d'indexer des textes par des entrées de ce thésaurus. Néanmoins, nous gardons à l'esprit que cet outil pourra être utilisé sur le meta-thésaurus UMLS complet, lorsque tous les concepts auront été traduits en français. D'autre part, nous utiliserons la taxonomie extraite du réseau sémantique du meta-thésaurus UMLS (plus détaillé que la taxonomie extraite du MeSH).

.II.3. Corpus de textes disponibles

Notre corpus de textes (à indexer) doit être le plus étendu possible, nous avons pris le parti d'indexer la plupart des textes médicaux disponibles sur le web, sans nous focaliser sur une spécialité médicale particulière.

Les premiers essais de l'outil d'indexation ont été faits sur des cours et des cas cliniques de radiologie disponibles sur le serveur de la faculté de médecine de Rennes¹³[Duvaufferrier et al., 1995]. Ceci nous a permis ensuite, d'indexer les cours de médecine disponibles sur le réseau pédagogique de la faculté de médecine¹⁴ [Fresnel et al., 1998]. La dernière utilisation a consisté à indexer tous les documents répertoriés sur le CISMef¹⁵ [Darmoni et al., 2000]. Une utilisation de cet outil pourra ensuite être mise en œuvre pour l'Université Médicale Virtuelle Francophone¹⁶ [Le Beux et al., 2000].

Après cette présentation des bases de connaissances dont nous disposons, voici maintenant les outils existants exploitant ces données.

¹³ URL: <http://www.med.univ-rennes1.fr/cerf>

¹⁴ URL: <http://www.med.univ-rennes1.fr/resped>

¹⁵ URL: <http://www.chu-rouen.fr/cismef>

¹⁶ URL: <http://www.umvf.org>

.II.4. Outils existants

Nous disposons de plusieurs outils de traitement du langage exploitant les thésaurus précédemment cités.

.II.4.1. Système actuel de recherche de concepts ADM

La première application du dictionnaire ADM a été de retrouver des symptômes ou des signes dans la base de données ADM (par extension, nous parlons de concept ADM). La recherche de concepts ADM se fait en langage naturel, l'utilisateur entre une phrase de recherche, et l'outil lui propose une liste de concepts ADM qui "contiennent" tous les mots de la phrase¹⁷. Deux types de recherche étaient alors mis en œuvre :

- 1) Recherche exacte : les concepts ADM dont un des termes est codé exactement de la même façon que la phrase. Pour des raisons d'optimisation, la recherche ne se fait que sur les termes contenant 5 mots au plus.
- 2) Recherche inexacte : les concepts dont un des termes contient tous les mots de la phrase.

Un exemple de recherche est disponible en annexe 1 ("Aide au Diagnostic Médical, exemple d'interrogation").

.II.4.2. Système de recherche en texte intégral

Dans l'attente d'un outil d'indexation plus sophistiqué, nous avons développé un système de recherche d'information en texte intégral. Le fonctionnement est relativement simple: une phase d'indexation segmente chaque document en phrase, chaque phrase est ensuite segmentée en mots, ces mots constituent l'index (ce sont les descripteurs). La recherche d'information consiste à segmenter la phrase de l'utilisateur en mots, et à proposer les documents ayant une phrase qui contient tous les mots de la requête de l'utilisateur.¹⁸

Ce moteur de recherche est actuellement disponible à l'adresse :

<http://www.med.univ-rennes1.fr/cgi-bin/idx/rechidx.pl> , un exemple de recherche est disponible en annexe 2 ("Exemple de recherche en texte intégral").

¹⁷ Plus exactement les entrées du lexique détectées dans la phrase (pouvant être un mot simple ou composé)

¹⁸ Nous avons choisi de développer ce moteur afin de l'intégrer plus facilement aux autres applications par la suite. L'autre solution aurait consisté à utiliser un moteur de recherche classique et libre (HT Dig, Harvest ...)

Il est très utilisé (actuellement, plus de 15 recherches par jour) bien que possédant tous les défauts inhérents à la recherche en texte intégral. La liste historique des interrogations est très conséquente et constitue une bonne référence sur les requêtes formulées par les utilisateurs.

.II.4.3. UMLS-index

Un premier travail a été accompli au Laboratoire d'Informatique Médicale par Philippe Rouault [Rouault, 1997]. Il consistait déjà à constituer un index de documents textuels à partir des informations linguistiques extraites de l'UMLS.

L'outil extrait les concepts, termes et unités lexicales du meta-thésaurus UMLS pour créer une indexation de documents médicaux multilingues. L'avantage de cet outil est de ne nécessiter aucune connaissance linguistique préalable, toutes les connaissances sont extraites de l'UMLS, il est par conséquent directement multilingue. L'inconvénient majeur étant que les informations linguistiques contenues dans l'UMLS sont parfois très pauvres. Cet outil fonctionne mieux pour la langue anglaise que pour le français, les flexions, dérivations et synonymes de mots ne sont pas répertoriés dans l'UMLS pour les autres langues que l'anglais (hors synonymes de termes).

.II.4.4. Outils de TALN de la NLM

Le moteur de recherche pubmed¹⁹ mis au point par la NLM permet de rechercher des articles publiés dans des revues médicales (et stockés dans la base de données medline). Celui-ci est bien connu des médecins. L'un des inconvénients majeurs est de ne travailler qu'en langue anglaise.

Un autre outil, lui aussi développé à la NLM, a pour but d'extraire des concepts UMLS (restreints au MeSH) d'un texte en langage naturel (ceci afin d'aider les indexeurs à trouver les concepts MeSH qu'il faut associer aux articles de la base de données Medline). Ce projet, appelé "IND" (NLM Indexing Initiative) est décrit dans [Aronson et al., 2000]. Le principe est relativement simple, pour chaque phrase du texte il génère tous les groupes nominaux, pour chaque groupe nominal, il génère tous les synonymes, flexions, sigles, possibles. Ensuite il recherche dans le thésaurus UMLS tous les concepts ayant l'une des chaînes de caractères ainsi générées.

¹⁹ URL: <http://pubmed.org>

Si la méthode paraît robuste, elle entraîne un nombre d'itérations assez conséquent. Elle n'est pas applicable telle quelle pour le français pour deux raisons :

- Les flexions et synonymes des mots ne sont présents que pour l'anglais dans l'UMLS
- Il faudrait développer – ou adapter – les deux modules d'extraction des groupes nominaux du texte et de génération des synonymes, flexions et sigles.

Après cette introduction, nous présentons maintenant notre travail, après avoir précisé les méthodes et les ressources que nous avons choisies.

Réalisations

Notre objectif est de créer un moteur d'indexation entièrement automatique (pour ses performances quantitatives), nous utiliserons certains des outils linguistiques (ceux qui nécessitent le moins de travail préalable) pour extraire des textes les descripteurs. Les applications de cette indexation utiliseront, elles aussi, les méthodes statistiques qui sont habituellement utilisées sur les mots.

.I. Choix linguistiques

Nous avons volontairement écarté les outils d'analyse grammaticale ou syntaxique et cela pour plusieurs raisons :

- Les thésaurus (dont l'indexation en mots est le premier objectif de notre système) ont un langage grammaticalement très pauvre (absence de verbes, très peu d'articles...)
- Il est bien connu qu'il est impossible de définir une grammaire couvrant la totalité d'une langue [Abeillé et Blache, 2000], et le langage médical a son propre vocabulaire, sa description grammaticale et syntaxique requiert un travail d'experts dont nous ne disposons pas.
- Nous disposons déjà du dictionnaire ADM qui contient suffisamment de synonymes et de flexions pour se passer de lemmatiseur.
- Les thésaurus ADM et MeSH sont suffisamment précis pour représenter la plus grande partie de l'information conceptuelle des documents.

[Abeillé et Blache, 2000] reconnaissent que l'on peut se passer de syntaxe dans le cas où les études "portent sur des domaines extrêmement limités et, d'autre part, utilisent des bases de connaissances très détaillées dans lesquelles les structures sémantiques associées aux objets contiennent implicitement les informations syntaxiques". Si le domaine médical n'est sûrement pas "extrêmement limité" les thésaurus médicaux contiennent effectivement suffisamment d'informations implicites pour lever beaucoup d'ambiguïtés.

Par exemple, le mot "PORTE" est extrêmement ambigu, s'agit t'il de la porte ("d'entrée"), de la veine porte, d'une conjugaison du verbe "porter" (d'autant plus ambigu quand on ne tient pas compte des accents)... Mais ici, l'ambiguïté est aussitôt levée car, dans le thésaurus, ce mot n'est employé que dans les concepts : "veine porte", "thrombose veine porte", ... Par conséquent, il n'est pas vraiment utile d'essayer d'attribuer une étiquette syntaxique au mot

"porte" dans une phrase car il n'aura d'importance, dans notre contexte, que s'il est accompagné du mot "veine"²⁰.

Comme le précise Pierre Zweigenbaum : "l'emploi des termes normalisés d'une terminologie de référence résout la plupart des difficultés d'échange d'information (...) l'ambiguïté est énormément réduite, si ce n'est supprimée." [Zweigenbaum, 1999].

²⁰ Notons que le fait de travailler uniquement sur le domaine médical diminue considérablement les effets de bords. A titre anecdotique, notons que le concept "Veine porte" pourrait être reconnu dans la phrase "La porte d'entrée est faite en bois exotique veiné de lignes noires" (mais la probabilité d'apparition d'une telle phrase dans notre domaine est extrêmement faible).

.II. Choix techniques

Tous les développements sont basés sur le système habituel en trois couches : les données, les traitements et l'interface. Nous avons la volonté de construire, dans la mesure du possible, un système indépendant de tout produit commercial. En ce qui concerne l'interface, nous avons tout naturellement choisi le web, qui permet d'accéder à notre outil depuis n'importe quel poste connecté à Internet, qu'il s'agisse d'un poste Unix, d'un PC/windows ou d'un Macintosh (particulièrement répandu dans le domaine médical). Pour les données, nous avons choisi l'architecture la plus souple à l'heure actuelle, c'est-à-dire une base de données relationnelle, en effet, ce modèle permet d'interroger notre base de connaissances à partir de n'importe quelle entité, sans aucune limite théorique. Les traitements consistent essentiellement à interroger la base de données, à analyser des textes, et à produire des résultats sur l'interface choisie (HTML). Nous avons naturellement choisi Perl [Wall et al., 1996] comme langage de programmation.

Le choix aurait pu se porter sur les langages Prolog ou Lisp utilisés en Intelligence Artificielle, mais l'interface avec les bases de données (commerciales ou du domaine public) est beaucoup moins souple qu'avec Perl. D'autre part, ces langages se révèlent beaucoup moins pratiques pour la manipulation de textes et de fichiers. Or une grande partie de notre travail est justement d'exploiter des textes, des fichiers et des bases de données existantes. L'expérience de [Brandt et Nadkarni, 1999] est intéressante, ils ont testé deux méthodes de manipulation des concepts UMLS, l'une avec Lisp, l'autre avec une base de données et un langage de script (ASP), leur conclusion est que "les développeurs utilisant les langages traditionnels d'Intelligence Artificielle et qui développent pour le web devraient sérieusement étudier si les langages de scripts ne satisfont pas mieux leurs besoins". Ici, le but principal n'est pas de développer des applications web, mais nous pensons que la manipulation de textes et de fichiers justifient l'usage d'un langage tel que Perl.

Pour citer [Piotrowski, 2000] : "Perl est un langage particulièrement adapté à la manipulation de textes, s'il offre la possibilité d'écrire rapidement de "petits" scripts pour des tâches particulières (grâce à sa nature interprétée), il offre également des concepts orientés objet de très haut niveau. L'intégration aisée de routines C++ enlève toute limite sur les possibilités d'extension. Perl est probablement le seul langage offrant une si grande diversité de modules (bibliothèques) du domaine public²¹. Les programmes Perl sont adaptables sur tout

²¹ Avec le "*Comprehensive Perl Archive Network*" – CPAN (<http://www.cpan.org>)

type de machine, sur tout système d'exploitation, ce que les autres langages (comme C++ ou Java) ont encore du mal à prouver".

Ajoutons que Perl offre une interface (DBI) aux bases de données relationnelles particulièrement intéressante : une simple ligne décrit le SGBDR utilisé à la connexion, toutes les autres requêtes à la base de données étant standards [Descartes et Bunce, 2000]. Ici, le système de gestion de base de données (SGBD) choisi est Oracle, pour sa rapidité essentiellement. Mais, le système est adaptable, sans grande difficulté, sur un SGBD relationnel du domaine public comme postgres ou Mysql.

Le serveur web utilisé est Apache, dont la robustesse n'est plus à prouver. Il a par ailleurs le mérite d'être gratuit et indépendant de la plateforme. Un module Perl "CGI" permet l'interfaçage avec le serveur [Gundavaram, 1996].

Nous avons travaillé sur un serveur Sun, sur le système d'exploitation Solaris. Mais les choix techniques font que le système pourrait être utilisé sur d'autres systèmes Unix (Linux, MacOS 10...) ou même sous Windows/NT.

Notons que les performances des systèmes de recherche d'informations sont parfois meilleures avec une architecture spécifique pour l'accès aux données (exemple de SMART par [Buckley, 1985]). Mais, dans notre contexte, les multiples possibilités des bases de données relationnelles compensent largement les temps de réponses qui, nous le verrons dans le chapitre "Evaluation" (p. 110), restent largement compatibles avec une utilisation intensive.

Voici en résumé l'architecture choisie pour les développements :

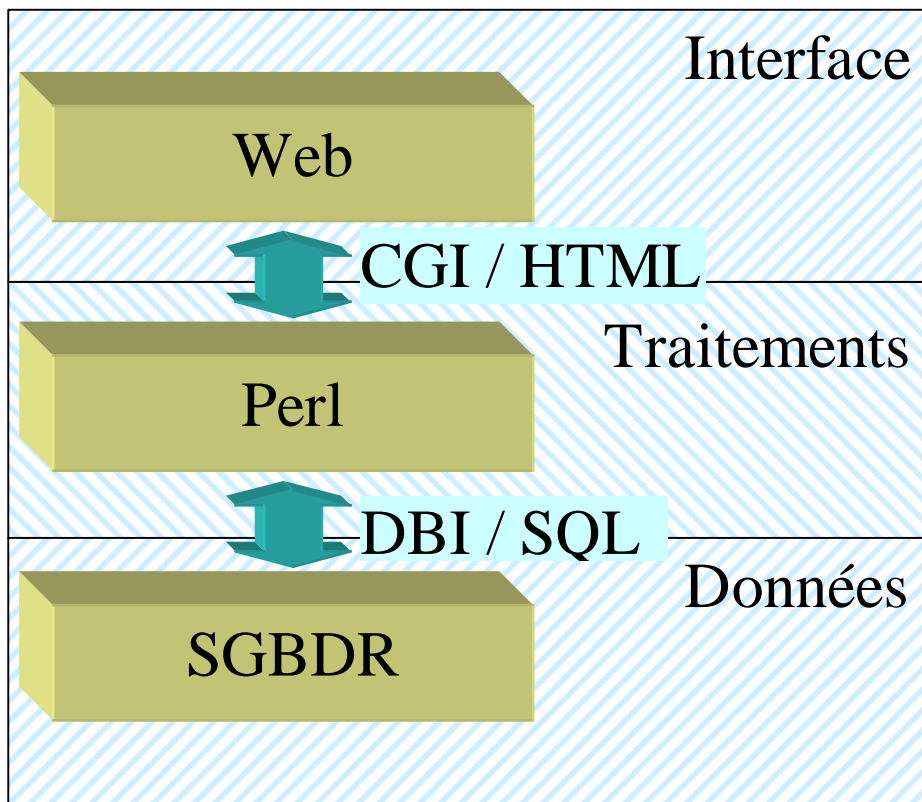


Figure 13 : Schéma en trois couches

Ces choix linguistiques et techniques ayant été faits, nous abordons maintenant le fonctionnement du système d'indexation, qui constitue le cœur de notre travail.

.III. Fonctionnement du système

Nous souhaitons décrire chaque terme du thésaurus (ADM ou MeSH) par les mots de référence qu'il contient, afin de permettre la reconnaissance du terme quels que soient les synonymes ou flexions utilisés. Ensuite nous essayerons de reconnaître les termes utilisés dans une phrase à l'aide de l'indexation précédente. Des termes nous extrayons ensuite les concepts. Une étape de "généralisation" produira les concepts implicites de la phrase (concepts hyperonymes).

La Figure 14 représente le schéma général du système NOMINDEX. Nous détaillerons ensuite chacune des fonctionnalités.

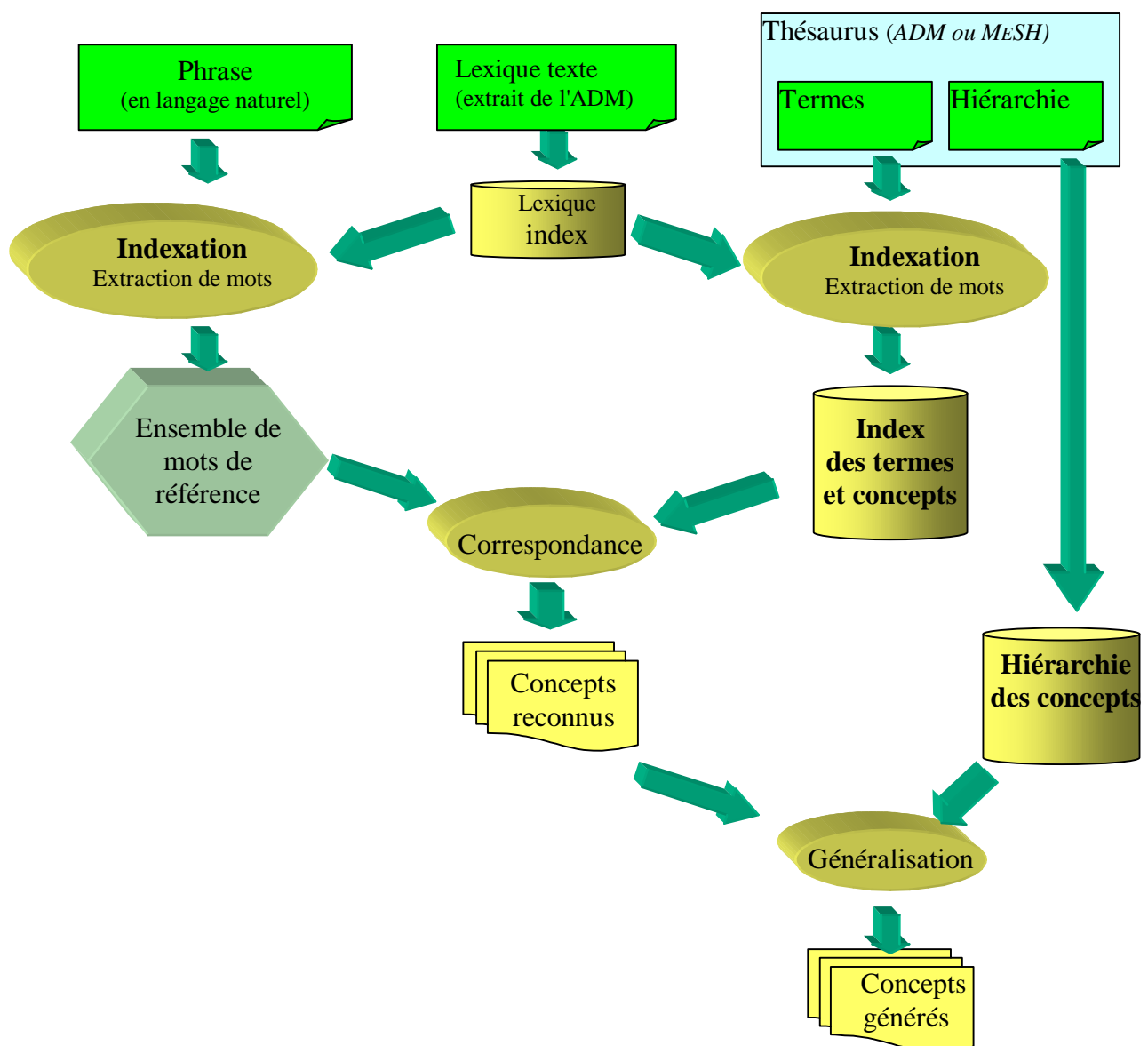


Figure 14 : Fonctionnement du système d'indexation NOMINDEX

Sur un exemple particulier :

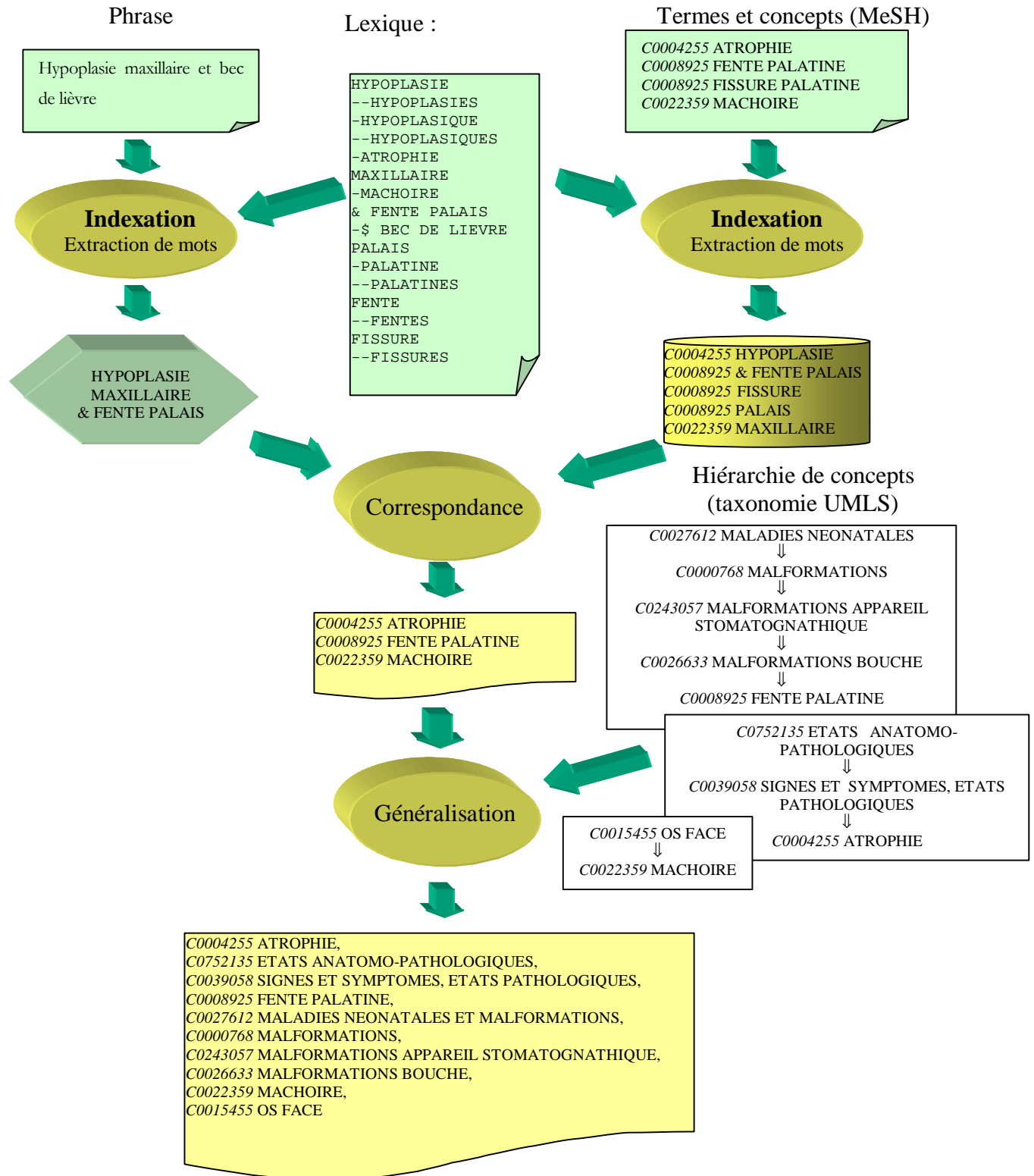


Figure 15 : Exemple de fonctionnement du système NOMINDEX

.IV. **Indexation en mots de référence**

La première réalisation est d'exploiter le dictionnaire ADM pour résumer chaque phrase par un ensemble de mots de référence. Ainsi la phrase "Néphrite glomérulaire lupique" sera indexée par les deux mots "Lupus" et "Glomérulonéphrite" (car on a reconnu le mot composé "Néphrite glomérulaire" synonyme de "Glomérulonéphrite"). Un peu à la manière d'un lemmatiseur, à la différence qu'un lemmatiseur se limite aux flexions sans gérer les dérivations ni la synonymie.

Notre système d'indexation sera alimenté par un lexique des mots médicaux français, ce lexique sera initialement créé à partir du dictionnaire ADM, mais pourra évoluer séparément par la suite.

Le dictionnaire ADM contenant d'emblée les flexions, dérivations, synonymes et quasi-synonymes des mots du vocabulaire, nous n'avons pas non plus eu besoin d'utiliser un lemmatiseur.²²

Cette première indexation permet de diminuer le silence lors de la recherche d'information, ainsi, quels que soient les flexions, dérivations ou synonymes des mots utilisés dans une recherche, on retrouvera les mêmes informations. Notamment grâce aux mots composés et associés.

Comme nous l'avons vu lors de la présentation du dictionnaire ADM, le système prend en compte deux types de mots multiples : les mots composés ("BEC DE LIEVRE") et les mots associés ("FENTE PALATINE").

.IV.1. Représentation du lexique

Le lexique est stocké dans une base de données relationnelle, avec la structure suivante :

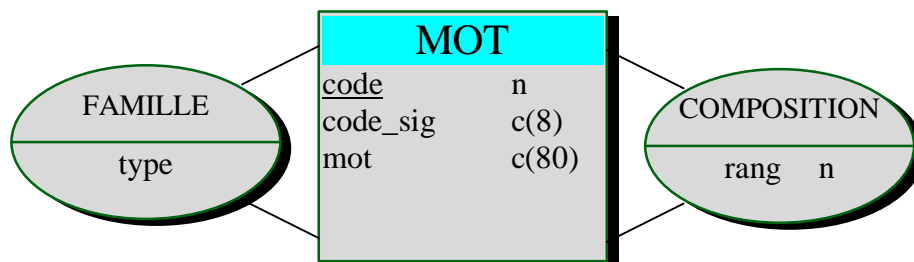


Figure 16 : Schéma de base de données des mots

²² Il aurait cependant été utile d'en disposer pour gérer les ajouts de mots dans le lexique, mais le vocabulaire médical étant assez spécifique, le travail de configuration aurait été très fastidieux, pour un résultat incertain. Des outils comme [Zweigenbaum et Grabar, 2000] pourraient permettre d'enrichir le lexique par de nouvelles formes

Qui, en résumé, contient une entrée par mot, chaque mot étant identifié par un code unique (attribut "code"), un code de signature (attribut "code_sig") qui est construit à partir du code du "mot de référence" et de trois caractères indiquant s'il s'agit d'une flexion ou d'un synonyme (qui permet, à lui seul, de représenter physiquement la relation "famille"). Notons qu'un même mot ne peut appartenir à deux familles différentes²³. Une entrée du lexique contient l'orthographe du mot (attribut "mot"). Les mots multiples seront représentés par le caractère "\$" (pour les mots composés) ou le caractère "&" (pour les mots associés) suivi des composants séparés par un espace. Dans le cas d'un mot multiple, la relation "composition" indiquera quels en sont les composants (le rang indiquant l'ordre de la composition, qui est pris en compte dans la reconnaissance des mots composés).

Rappelons que la notion de "synonymie" est assez large dans notre lexique, ainsi "radiologie" et "radiologue" sont actuellement étiquetés comme synonymes, même si le sens des deux mots est clairement différent, ceci permet de diminuer le silence à l'indexation (au détriment parfois du bruit). Ainsi le concept "radio thorax" pourra être reconnu dans une phrase comme : "Le radiologue a examiné le thorax".

Originellement, la mise à jour du dictionnaire ADM se faisait directement sur la base de données (un formulaire de saisie faisait la mise à jour immédiate). Ceci entraînait des erreurs, car le thésaurus avait été mis à jour avec une version préalable du dictionnaire... Ce qui imposait parfois de mettre à jour une grande quantité d'enregistrements après la modification d'un mot très utilisé. La complexité de la prise en compte des changements dans les mots multiples imposait de recoder tout le thésaurus.

à partir d'un thésaurus, ou d'aider à créer un nouveau lexique (pour une autre langue que le français par exemple).

²³ Ce qui pose parfois des problèmes d'ambiguïté (le mot "PORTE" ne pourra appartenir à deux familles de mots), les mots composés sont utilisés pour résoudre une partie des ambiguïtés (cf. p. 37) .

Dans la version actuelle, le lexique est maintenant importé depuis un fichier texte. Toute modification du lexique se fait maintenant dans un éditeur de texte habituel. La prise en compte d'une nouvelle version ne pose plus de problème (Sur un serveur SUN Ultra Enterprise 450, l'importation dure environ une demi-heure, mais surtout, le recodage du thésaurus nécessite deux heures de plus). Cette méthode permet de mettre à jour aisément le lexique, et ainsi de l'adapter à cette nouvelle utilisation (rappelons que le dictionnaire ADM n'avait pas été créé dans ce but, il faut donc prévoir la possibilité de l'adapter en fonction des besoins).

La structure du fichier texte est volontairement très simple :

Chaque mot de référence est représenté par le mot avec un espace en début de ligne.

Chaque synonyme est représenté avec deux espaces, chaque flexion avec trois espaces.

La figure ci-dessous est un extrait du lexique :

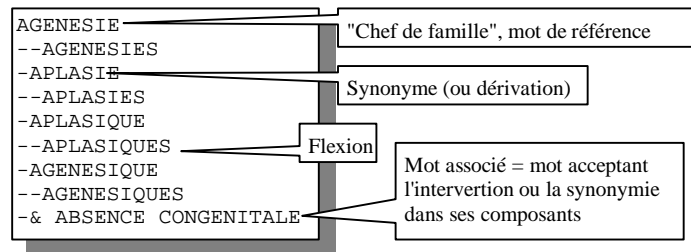


Figure 17 : Un exemple d'entrée du dictionnaire ADM

Le lexique peut être exporté au format XML ("eXtensible Markup Language"), et ainsi être exploité, ou mis à jour, par d'autres applications.

```
<?xml version="1.0" standalone="yes"?>
<?xml:stylesheet type="text/xsl" href="dico.xsl"?>
<!DOCTYPE dico SYSTEM "dico.dtd">
<dico>
<famille code="00972" terme="AGNESIE">
<chef><mot code="972">AGNESIE</mot>
  <flexion><mot code="973">AGNESIES</mot>
</flexion>
  <synonyme><mot code="974">APLASIE</mot>
    <flexion><mot code="975">APLASIES</mot>
  </flexion>
</synonyme>
  <synonyme><mot code="976">APLASIQUE</mot>
    <flexion><mot code="977">APLASIQUES</mot>
  </flexion>
</synonyme>
  <synonyme><mot code="978">AGNESIQUE</mot>
    <flexion><mot code="979">AGNESIQUES</mot>
  </flexion>
</synonyme>
  <synonyme>
    <mot code="980">& ABSENCE CONGENITALE</mot>
  </synonyme>
</chef>
</famille>
</dico>
```

A l'aide d'une feuille de style appropriée (XSL), cet exemple pourrait apparaître sous la forme suivante :

Lexique

- **AGENESIE**
Mot de référence : **AGENESIE**
 - flexion : **AGENESIES**
 - Synonyme :
APLASIE
 - flexion : **APLASIES**
 - Synonyme :
APLASIQUE
 - flexion : **APLASIQUES**
 - Synonyme :
AGENESIQUE
 - flexion : **AGENESIQUES**
 - Synonyme :
& ABSENCE CONGENTALE

Figure 18 : Fichier XML d'export du dictionnaire ADM, et visualisation

Ce lexique, une fois constitué, nous permet de reconnaître les mots de références contenus dans une phrase en langage naturel, nous présentons maintenant le fonctionnement de cette fonctionnalité.

.IV.2. Fonctionnement du programme de reconnaissance de mots

Tmots := découpe_phrase_en_mots(phrase);	(1)
Tmots_composés := recherche_mots_composés(Tmots);	(2)
Tmots_associés := recherche_mots_associés(Tmots);	
Si (Tmots contient des mots nuls) alors	
Enleve_les_mots_nuls(Tmots);	
Tmots_composés := recherche_mots_composés(Tmots);	(3)
Fsi;	
Tmots_associés := recherche_mots_associés(Tmots);	(4)
Tmots := enleve_composants(Tmots);	(5)

Algorithme 1 : Reconnaissance de mots dans une phrase

- (1) : La phrase est convertie en majuscules non accentuées. Un petit traitement (par substitution d'expression régulière) traite les sigles écrits avec des points ("A.D.N." sera remplacé par "ADN", ces deux orthographes étant souvent utilisées). Un mot sera ensuite reconnu comme une suite contiguë de caractères alphanumériques. Une option permet de supprimer les parties de phrases entre parenthèses (ou entre crochets).
- (2) La recherche de mots composés d'après un tableau de mots est détaillée plus loin. Le tableau Tmot est mis à jour automatiquement, chaque mot composé est ajouté au tableau.
- (3) Ceci permet de rechercher les mots composés après suppression des mots nuls (par exemple, dans la phrase "Angine de poitrine", ou même "Angine de la poitrine", le mot composé "ANGINE POITRINE" ne sera pas reconnu à la première recherche, mais le sera à ce niveau car les mots nuls auront disparu).
- (4) Cette nouvelle recherche paraît inutile (elle l'est dans la plupart des cas !), mais elle permet de reconnaître les mots associés dont les composants sont eux-mêmes des mots associés (ou composés).
- (5) Noter une petite ambiguïté : que doit-on faire quand un mot simple apparaît dans plusieurs mots multiples ? Pour les mots associés, cela ne pose pas de problèmes, on reconnaîtra les deux mots associés "PRE PRANDIAL" et "POST PRANDIAL" dans la phrase "pré ou post prandial" (le composant "PRANDIAL" participant à deux mots associés). Pour les mots composés, cela est plus ambigu, mais on se rend facilement compte que, même un humain ne saura pas ce qu'il doit comprendre dans des phrases (bien improbables) comme "Herbe à fièvre jaune" ou "chemin de fer à cheval"²⁴.

²⁴ En fait, notre programme reconnaîtra les deux mots composés ("Fièvre jaune" et "Herbe à fièvre" dans le premier cas, "Chemin de fer" et "Fer à cheval" dans le second)

.IV.3. Reconnaissance de mots multiples et performance

La reconnaissance de mots multiples dans une phrase pourrait être très coûteuse en temps si nous ne l'avons pas optimisée. En effet, il serait catastrophique de parcourir tous les mots multiples du lexique pour voir si tous leurs composants sont présents dans la phrase. Et sélectionner les mots multiples contenant un des mots de la phrase n'est pas non plus suffisant (une très longue phrase engendrera un nombre exponentiel de recherches).

Nous avons donc opté pour l'optimisation suivante :

Chaque mot multiple est codé par un tableau contenant l'identifiant de chaque mot de référence de ses composants. Ce tableau est ensuite trié par ordre d'identifiant croissant. Ce tableau ne peut contenir plus de 5 éléments (Les mots associés de plus de 5 éléments sont anecdotiques).

Au moment du codage de la phrase, nous appliquons le même traitement, en ne prenant en compte que les mots apparaissant au moins une fois dans un mot multiple.

```

Tmot := mots de la phrase;
Tmp := mots_apparaissant_dans_un_mot_multiple(Tmot);
/* n_tmp. taille de Tmp */
Pour i (1.. n_tmp - 1) faire      => le dernier mot (Tmp[n_tmp]) n'a pas d'importance
    Chercher les mots multiples dont l'élément 1 est égal à Tmp[i]
    Vérifier si tous les autres composants appartiennent bien à Tmp[i+1.. n_tmp]
Fait

```

Algorithme 2 : Fonction de recherche de mots multiples

Ainsi le nombre d'itérations est grandement réduit. Sachant que l'on alloue les codes d'identification des mots de référence par ordre inverse d'apparition d'un des mots de la famille dans un mot multiple (ainsi, si l'on avait plusieurs mots multiples comme "maladie de Parkinson", "maladie de Caroli", "maladie de Kaposi", on les reconnaîtrait par le troisième mot, et surtout pas par "DE" ou "MALADIE").

.IV.4. Préfixes et suffixes

Un traitement particulier a été fait pour que, lorsque le mot n'est pas présent tel quel dans notre lexique, on essaie de le segmenter en deux composants si le premier composant est un préfixe identifié et que le second est un mot connu de notre lexique. Cette liste est ensuite proposée à l'utilisateur pour qu'il identifie les segmentations abusives (notre programme, ne connaissant pas les mots "POSTAUX" et "TRANSPARENTS" proposait de les segmenter en "POST-AUX" et "TRANS-PARENTS" !).²⁵

.IV.5. Mots inconnus

Une fonction de correction automatique de fautes d'orthographe est intégrée à l'algorithme. Son fonctionnement est simple : quand un mot est inconnu du lexique, on commence par regarder si une flexion possible de ce mot est présente (par substitution d'expression régulières, si le mot "Endocrines" est absent, on essaiera de trouver le mot "Endocrine"). Sinon, on regarde si une segmentation du mot en deux composants est possible (par exemple "ACTINOBACILLOSE" pourrait être décomposé en "Actino" et "Bacillose"). Sinon, on cherche les mots s'orthographiant de la même manière à un caractère près.

Dès qu'un mot inconnu est rencontré, un message est automatiquement généré dans un fichier de trace (fichier des mots inconnus), ce message indique si une flexion du mot a été trouvée, s'il y a eu segmentation de mot, si un mot "voisin" a été trouvé ou si aucune correction n'a pu être mise en œuvre. Ce fichier de trace est extrêmement utile pour la mise à jour (a posteriori) du lexique d'indexation.

Cette fonction pourrait être beaucoup améliorée par d'autres méthodes plus ambitieuses de correction automatique.

²⁵ Erreurs corrigées par la suite en insérant ces deux mots "postaux" et "transparents" dans le dictionnaire ADM

.IV.6. Résumé du principe de fonctionnement

Synthétiquement, voici le fonctionnement complet de la reconnaissance de mots de référence dans une phrase:

- 1) A partir d'une phrase (en langage naturel), détecter tous les mots simples appartenant à notre lexique (en incluant le traitement des mots inconnus)
- 2) Détecter les mots multiples dans cet ensemble de mots simples
- 3) Retirer les mots nuls
- 4) Représenter chaque mot par son "mot de référence", ordonner le tableau.

Le système est ainsi capable d'indexer une phrase en langage naturel, mais aussi tous les termes d'un thésaurus.

Cette fonctionnalité est ensuite exploitée pour indexer le thésaurus cible, on crée ainsi une base d'indexation des termes du thésaurus, et, par extension, des concepts du thésaurus.

.IV.7. Indexation d'un thésaurus en mots de référence

Le programme de pré-codage du thésaurus est relativement simple. Il prend, en entrée, un fichier texte contenant un terme du thésaurus par ligne, sous la forme :

code<tabulation>terme

Chaque ligne de ce fichier contient un terme, le code est un identifiant de concept.

Pour chaque terme, le programme décrit au chapitre précédent extrait un ensemble de mots de référence (comme un lemmatiseur incluant les synonymes et dérivations).

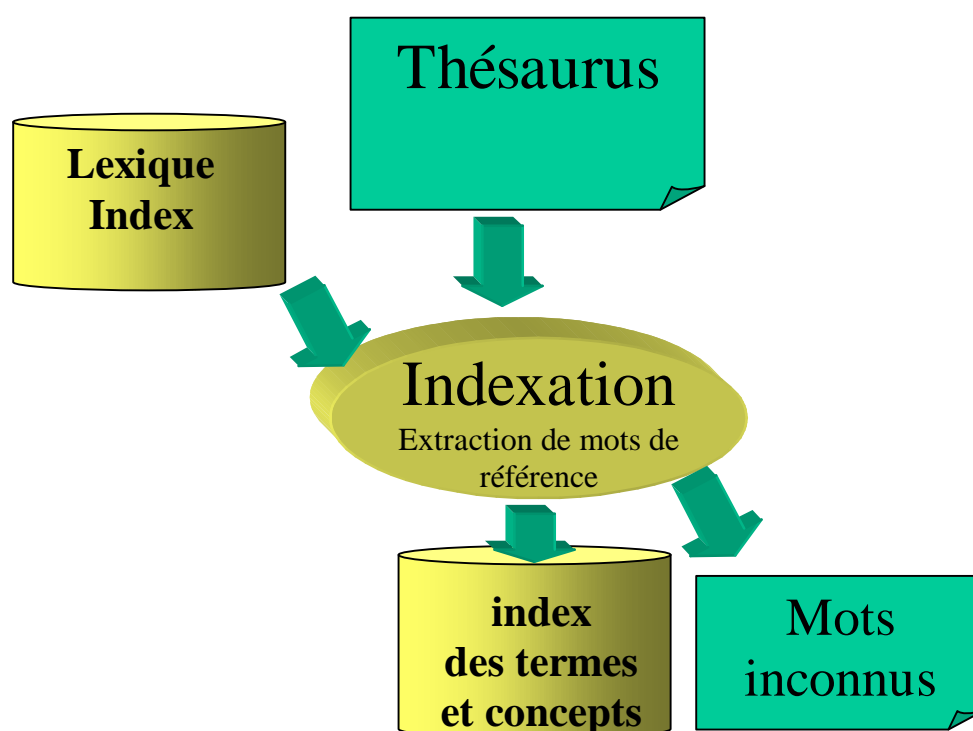


Figure 19 : Principe d'indexation d'un thésaurus

Comme l'indique le schéma, l'indexation d'un thésaurus produit aussi un fichier des mots inconnus. Ce fichier est extrêmement utile pour mettre à jour le lexique, les termes contenant des mots inconnus ne sont pas indexés.

Par exemple :

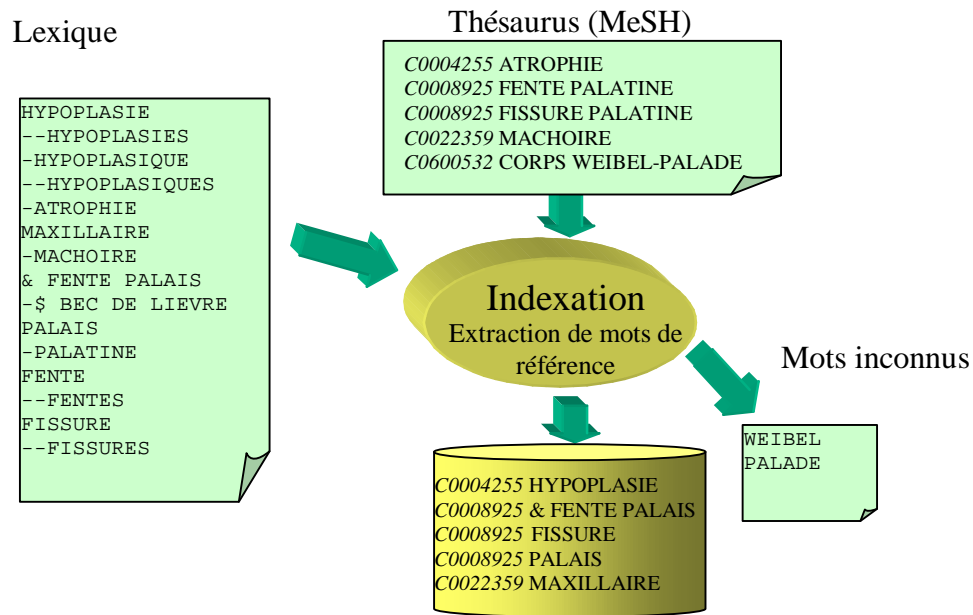


Figure 20 : Exemple d'indexation de thésaurus

La plupart des thésaurus ont des termes synonymes (l'entrée "Maladie du légionnaire" a comme terme synonyme "Légionellose"). Le plus souvent, l'un des termes est choisi comme terme préférentiel, celui le plus couramment utilisé pour dénommer le concept correspondant. Pour ce faire, dans le fichier d'entrée, nous définirons des termes synonymes comme étant des lignes qui possèdent le même code, le terme préférentiel sera le premier par ordre d'apparition (dans l'exemple ci-dessus "FENTE PALATINE" sera le terme préférentiel, et "FISSURE PALATINE" un terme synonyme).

Toujours pour des problèmes d'optimisation, nous avons décidé, de manière empirique, que seuls les termes contenant au plus 5 mots seront codés. Il est très peu probable que des termes contenant plus de cinq mots comme "PHYSIOLOGIE MUSCULAIRE ET APPAREIL LOCOMOTEUR, NEUROLOGIQUE ET OCULAIRE" soient détectés dans une phrase. De plus, il s'agit de mots au sens ADM, c'est-à-dire que les mots multiples sont automatiquement remplacés. Ainsi, si nous voulons absolument que certains termes de thésaurus, contenant plus de 5 mots, soient pris en compte, nous avons toujours la possibilité de créer des mots associés appropriés. Par exemple, dans le thésaurus ADM, un concept est exprimé par "HISTOLOGIE VESICULE BILIAIRE TUMEUR MESENCHYMATEUSE MALIGNNE LEIOMYOSARCOME" (qui fait donc 7 mots), mais, une fois codé en mots

ADM, ce concepts n'en comportera plus que 5 : "Histologie", "Vésicule biliaire", "tumeur maligne", "mésenchymateuse", et "léiomyosarcome".

Le résultat de ce codage de thésaurus est stocké dans une base de données, sous la forme de 2 tables :

La table CON qui contient une entrée par terme (qui est en quelque sorte le miroir du fichier texte d'entrée), chaque terme étant désigné par son code de concept, son texte et son numéro de terme ("nolib"), la clé primaire de cette table étant le code et le numéro de terme.

La table NOM qui contient, pour chaque terme, son codage en mots du dictionnaire ADM. Une relation représentée par la table HIER exprime la hiérarchie (taxonomie) entre un concept hyperonyme et son hyponyme.

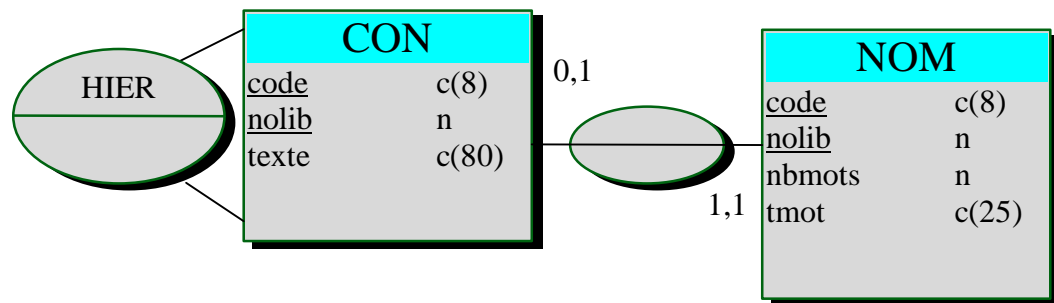


Figure 21 : Schéma de la base de données des concepts et de leur indexation en mots

Notons que la table CON contient les termes, pour en extraire les concepts on prendra les entrées telles que "nolib" est égal à 0 (terme préférentiel pour désigner le concept).

.IV.7.1. Problème des termes "équivalents"

Chaque terme étant codé en mots, nous pouvons ensuite épurer la table NOM en enlevant tous les doubles codages. C'est-à-dire ne garder qu'une seule entrée quand deux termes d'un même concept sont codés de la même manière (ce qui, dans toutes les applications de cet outil, est inutile). Par exemple, le concept C0018681 ("Céphalée") a cinq termes différents ("Céphalée", "Céphalalgie", "Céphalées", "Mal de tête", "Gêne dans la tête"), dans notre dictionnaire ADM, les mots "Céphalée", "Céphalées", et "Mal de tête" appartiennent à la même famille, seul le premier terme sera codé par NOMINDEX.

Par contre, il existe des termes (ADM ou MeSH), appartenant à des concepts différents, qui sont équivalents (dont tous les mots de référence sont identiques), par exemple les termes "SENTIMENT D'IRREEL" et "SENTIMENT D'ETRANGETE" sont équivalents dans notre codage (le mot "irréel" étant synonyme d'"étrangeté" dans le lexique).

Une option de notre programme permet de ne garder qu'un seul concept parmi ces concepts équivalents. Cette option n'est pas utilisée à l'heure actuelle, mais elle est explicitée en annexe 3 ("Épuration de concepts synonymes").

.IV.7.2. Fichier des mots inconnus

Ce programme engendre également un fichier de trace indiquant les problèmes qu'il a rencontrés avec des mots inconnus du lexique (quatre types d'erreurs : mot inconnu mais le programme a trouvé une segmentation automatique, mot inconnu mais le programme a trouvé une flexion automatique, mot inconnu mais le programme a trouvé une correction orthographique possible, mot inconnu introuvable). L'utilisateur peut ainsi analyser ce fichier afin de mettre à jour le lexique.

Ce qui lui permettra, par exemple :

- D'ajouter les mots inconnus dans le lexique (exemple: "Cérémonial", extrait du terme MeSH "Comportement cérémonial", qui pourrait constituer une nouvelle famille de mots avec "Cérémonials", "Cérémoniels", "Protocolaire", "Cérémonieux"...).
- D'ajouter des flexions non présentes dans le lexique ("Affirmations").

Notons que nous n'ajoutons aucun mot automatiquement dans le lexique, seul un opérateur avisé est autorisé à insérer un nouveau mot dans une famille existante. Par exemple, le mot "Proctite" est inconnu de notre lexique, mais il ne faut pas pour autant le créer tel quel car il est un quasi-synonyme de "Rectite".

Cette fonctionnalité a d'ailleurs été utilisée pour ajouter dans notre lexique la plupart des mots inconnus issus du MeSH français, avec un certain succès (voir chapitre "Évaluation du lexique", p. 110) car cela a permis de diminuer de manière très significative les mots non reconnus. Une telle mise à jour pourra être mise en œuvre à chaque nouvelle version du MeSH (et, à plus forte raison, si l'on teste l'outil avec un nouveau thésaurus).

Ce processus d'indexation de thésaurus en mots est ensuite utilisé par NOMINDEX qui fait la correspondance entre les mots d'une phrase et les termes du thésaurus pour reconnaître les concepts dans une phrase. Par extension, permet d'indexer des documents par les concepts extraits.

.V. Indexation de documents en concepts

Un document médical est segmenté en phrases, chaque phrase est ensuite analysée par le processus précédent pour en extraire les mots de référence.

La segmentation d'un document en phrases est un processus assez simple, étant donné que les documents sont au format HTML (voire SGML ou XML), nous avons défini des balises HTML qui sont des marqueurs de fin de paragraphes (<P>, , <H1> ...). Puis, à l'intérieur de ces paragraphes, nous segmentons les phrases selon des caractères de ponctuation (".", "?", ...) ²⁶. Nous sommes conscients qu'une segmentation plus évoluée serait plus adaptée, mais cela dépasse le cadre de notre travail.

On effectue une recherche des termes du thésaurus qui sont inclus dans la phrase (qui s'apparente à une recherche sur un modèle booléen). C'est-à-dire que l'on extrait les termes dont tous les mots sont présents dans la phrase. On retiendra ensuite les concepts représentés par ces termes. Et cela constituera l'indexation de la phrase. Par exemple, dans la phrase "Néphrite glomérulaire lupique", on reconnaîtra les concepts "Glomérulonéphrite lupique" et " Glomérulonéphrite" dans le thésaurus MeSH. Dans le thésaurus ADM, on reconnaîtra les concepts "Néphropathies glomérulaires" et "Lupus". Remarquons au passage les différences d'interprétation selon le thésaurus utilisé.

Une option du processus d'indexation permet de générer automatiquement les concepts hyperonymes des concepts extraits (en utilisant l'information de la hiérarchie). Ce qui permettra, à partir du concept "Glomérulonéphrite lupique" de générer les concepts "Glomérulonéphrite", "Néphrite", "Rein, maladies", "Appareil urinaire, maladies" et "Maladies urologiques et appareil génital male" (selon la relation de taxonomie "est-un" - "is_a" - du réseau sémantique de l'UMLS). Ce processus, par "pré-expansion", est équivalent à une "explosion" de la requête de l'utilisateur.

Nous avons fait le choix d'attribuer un poids à chaque phrase, afin de privilégier les concepts apparaissant dans le titre du document (ou un entête de paragraphe, ou une phrase en gras...). Ce poids est pré-établi, de manière empirique, en fonction de la balise HTML dans laquelle se trouve la phrase (un poids de 6 pour une balise <TITLE>, un poids de 4 pour une balise <H1>, ...).

Le résultat de cette seconde indexation sera ensuite stocké dans une base de données.

²⁶ En évitant, à l'aide d'expression régulière, de segmenter un sigle ("I.R.M.") ou un chiffre décimal ("37.2")

L'ancienne recherche de concepts ADM (cf. Système actuel de recherche de concepts ADM p. 46) recherchait tous les concepts qui contenaient une partie des mots de la question de l'utilisateur, ainsi, une interrogation sur "cardiopathie" proposait différents concepts : "cardiopathies congénitales", "cardiopathie du nouveau-né"... On recherchait tous les termes contenant au minimum tous les mots de la phrase. L'idée du nouveau moteur de recherche est au contraire de rechercher tous les termes dont tous les mots sont inclus dans la phrase.

.V.1.Principe de fonctionnement

Dans une phrase contenant n mots (m_1, m_2, \dots, m_n) , on crée son *index* de mots :

$I(i_1, i_2, \dots, i_m)$ à partir du lexique.

On détecte les entrées de notre thésaurus telles que :

Pour chaque ensemble T , index de l'entrée du thésaurus, l'entrée sera reconnue si :

$$\forall t \in T, t \in I$$

Algorithme 3 : Reconnaissance de concepts dans une phrase

En utilisant la même méthode d'optimisation de la recherche de mots associés (cf. Reconnaissance de mots multiples et performance, p. 62), les temps de calculs deviennent compatibles avec une utilisation en ligne (le temps nécessaire à la détection de termes de thésaurus dans une phrase doit être inférieur à une seconde si l'on veut pouvoir interroger notre base de données interactivement).

Nous allons donc considérer chaque terme de thésaurus comme une association de mots. Une phrase sera codée de la même manière. Ensuite nous regarderons quels sont les termes du thésaurus dont tous les mots sont présents dans la phrase.

Le résultat de l'indexation d'une phrase sera donc un ensemble de termes de thésaurus (ensemble qui peut bien sûr être vide). Connaissant les termes, on en déduit les concepts. On ajoute éventuellement les concepts hyperonymes.

Une entrée de thésaurus étant un terme correspondant à un concept médical, nous avons, pour des raisons de commodité de langage, appelé cette indexation : indexation par concepts médicaux.

.V.2.Algorithme...

Cet algorithme suppose que tout le thésaurus a été, au préalable, codé en mots de référence (cf. chapitre précédent).

```
Tmot := mots de la phrase;
Trier(Tmot); /* tri identique à celui utilisé lors de la constitution de l'index */
/* n_mots. taille de Tmp */
Pour i (1.. n_mots) faire
    Chercher les termes dont l'élément 1 est égal à Tmot[i]
    Vérifier si tous les autres mots du terme appartiennent bien à Tmot[i+1.. n_mot]
Fait
```

Algorithme 4 : Fonction de correspondance de NOMINDEX

Disposant de cet outil d'indexation d'un texte, l'étape suivante consiste à indexer un corpus de textes.

.VI. Indexation du corpus de textes avec NOMINDEX

Nous rappelons que l'indexation de textes a pour but de représenter chaque texte d'une manière commode afin de permettre de :

- Voir l'indexation qui a été faite d'un texte (si possible phrase par phrase)
- Rechercher les textes correspondant à une phrase de l'utilisateur
- Faire une synthèse automatique de texte (voire de plusieurs textes)
- Proposer automatiquement des mots-clés typiques du texte
- Classer automatiquement des textes, proposer une cartographie du corpus
- Calculer une proximité de textes entre eux (et ainsi proposer des documents sémantiquement proches)

L'utilisation principale restant la recherche documentaire.

À l'aide du module NOMINDEX, nous sommes maintenant capables d'indexer une phrase. Un texte étant constitué de plusieurs phrases, il était logique de poursuivre avec l'indexation de textes.

Le dictionnaire ADM étant écrit en français, nous n'indexerons que des textes écrits dans la langue de Molière. Nous verrons cependant dans le chapitre "Traduction automatique pour indexation" (p. 101) que notre système conserve une possibilité d'indexer des textes non francophones.

Le but premier est de détecter les concepts présents dans un texte, et, accessoirement, stocker le résultat de cette indexation. La méthode choisie est d'"étiqueter" chaque phrase d'un document par les concepts détectés.

Nous avons choisi de travailler sur des textes écrits soit en format texte standard (ISO-LATIN-1), HTML (voire SGML ou XML). Les autres formats de documents (Word, RTF, PDF, ...) seront codés après traduction en texte, HTML ou XML, bien que cette fonctionnalité ne soit pas à l'heure actuelle encore intégrée. Des outils de conversion, tels que "pdftotext"

(<http://www.foolabs.com/xpdf>) ou "wvhtml" (<http://www.wvware.com>) pourraient aisément être intégrés à NOMINDEX.

Le codage est simple : On segmente le texte en phrases, on appelle NOMINDEX avec chacune des phrases. Le résultat est retranscrit sous la forme d'une balise de type SGML (compatible HTML et XML). Sous la forme :

```
<A NAME="THESnuméro" INDEX="Code1,Code2,...">phrase</A>
```

Avec

- THES : un code représentant le nom du thésaurus utilisé ("MESH", "ADM")
- *Numéro* : un numéro d'ordre de la phrase
- *Coden* : un code de concept détecté

Nous avons choisi ce type de balise, car c'est une balise HTML de positionnement d'ancre. Il sera possible, dans une utilisation ultérieure, de se positionner directement à l'endroit où se trouve cette ancre dans le texte.

Ce codage sera ensuite très facilement accessible par un analyseur HTML (ou XML) ("HTML parser").

Exemple :

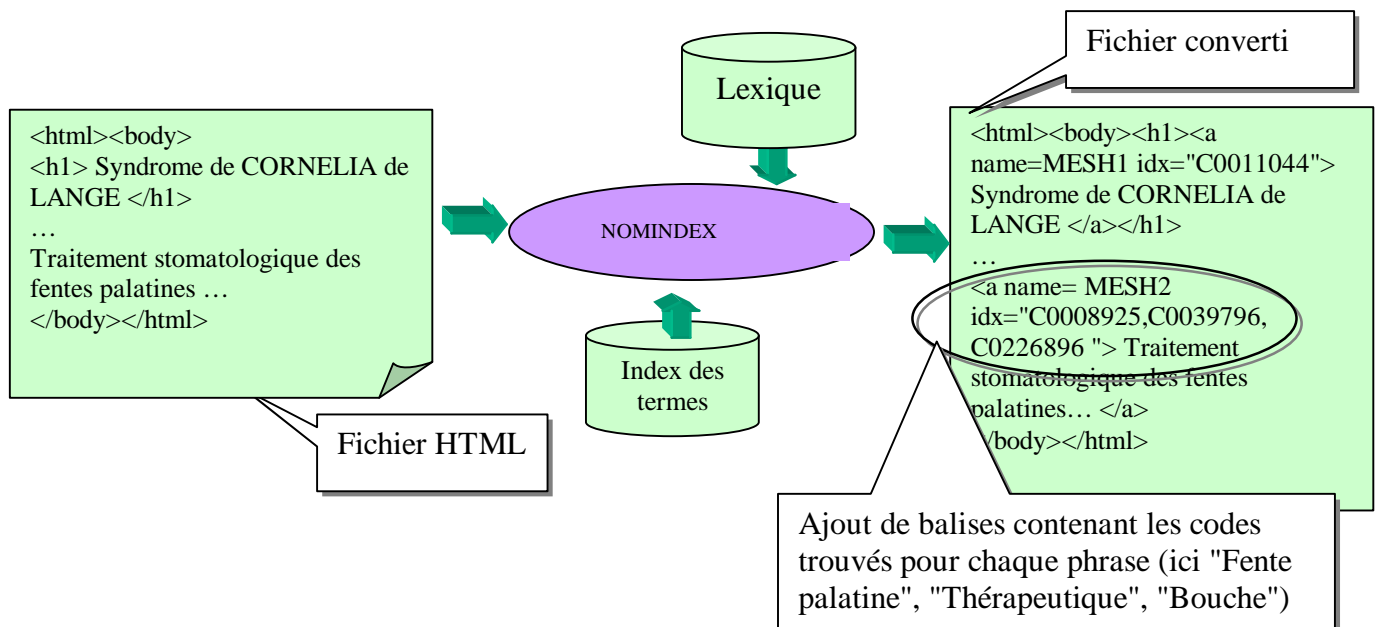


Figure 22 : Exemple d'étiquetage de document HTML

Le fichier converti est, par la suite, analysé pour produire notre base d'index. Notons cependant que cet "étiquetage conceptuel" pourrait tout à fait constituer un fichier d'entrée à diverses autres applications.

.VII. Représentation de l'indexation

Nous écartons d'emblée l'approche consistant à tout calculer "à la volée", c'est-à-dire, recalculer l'indexation dynamiquement à chaque requête. Une recherche en texte libre pourrait durer quelques heures dans ce cas ! Nous choisissons plutôt de stocker le résultat de l'indexation d'une manière commode en vue des utilisations qui en seront faites.

La solution adoptée est la plus souple : nous avons stocké l'indexation des textes dans une base de données. Qui contiendra l'ensemble des documents (FIC), l'ensemble des phrases des documents (PHR) et les codes de concepts détectés dans chaque phrase.

L'index des documents aura une structure relativement simple :

- Une table FIC qui contient une entrée par fichier indexé
- Une table PHR qui contient une entrée par phrase indexée (une phrase qui ne contient aucun concept ne sera pas stockée)
- Une table IDX qui contient une entrée par concept trouvé dans une phrase.

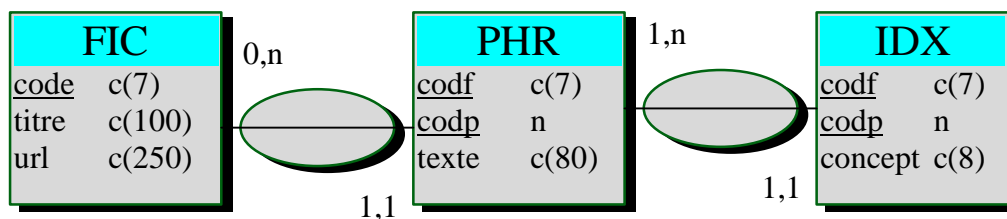


Figure 23 : Schéma de la base de données d'index des documents

Un exemple de représentation est donné sur la Figure 25 : Exemple d'indexation de document (contenu de la base de données) p. 77.

Le processus complet d'indexation d'un document sera le suivant :

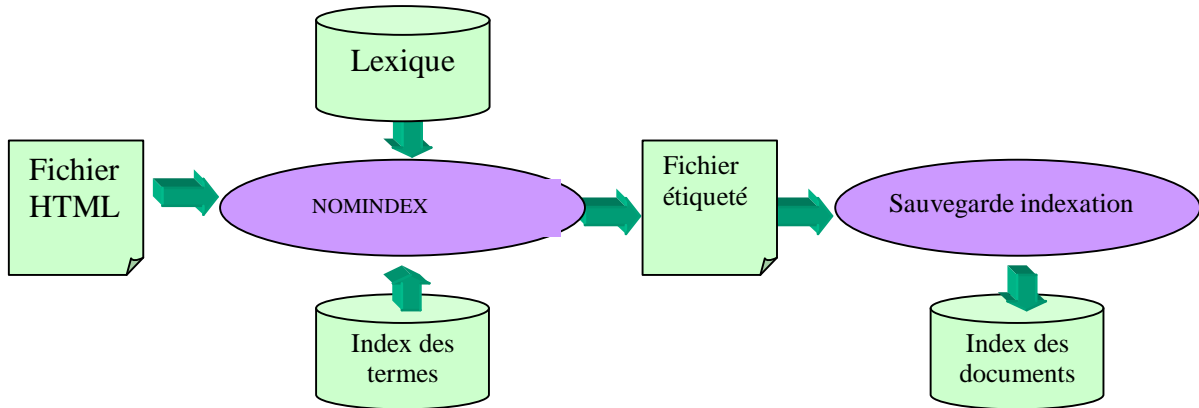


Figure 24 : Processus d'indexation d'un document

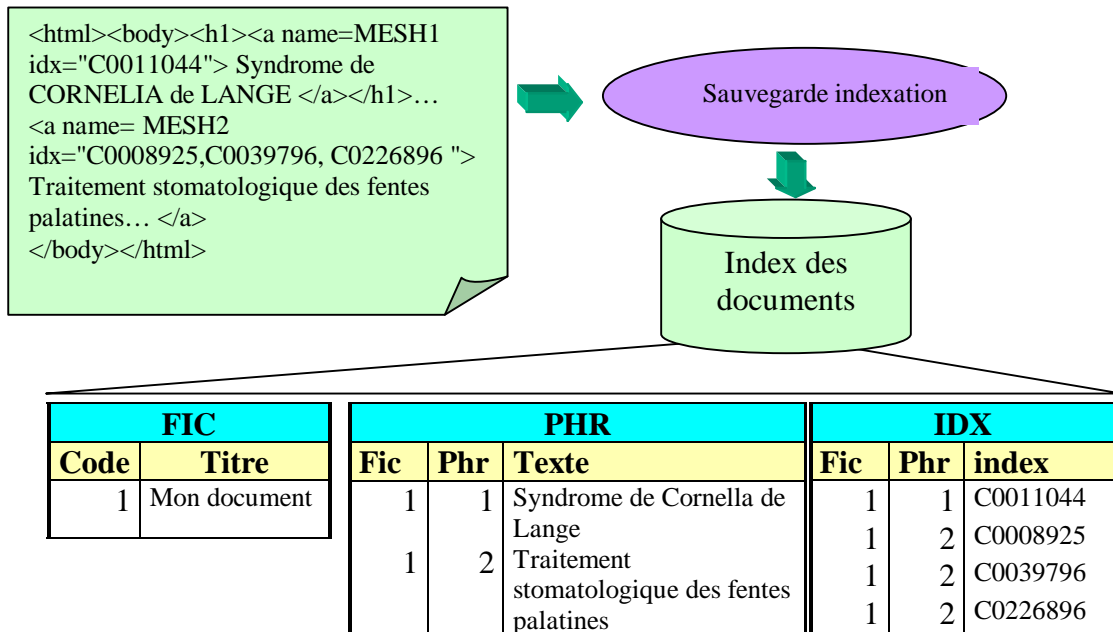


Figure 25 : Exemple d'indexation de document (contenu de la base de données)

L'indexation d'un texte se fera en plusieurs étapes :

- Lecture et éventuelle transformation du texte (1)
- Création d'une entrée dans la table FIC (2)
- Appel de NOMINDEX qui créera un fichier de sortie contenant les balises d'indexation
- Analyse de ce fichier de sortie :
Pour chaque phrase contenant une indexation:
 - Attribution d'un "poids" à la phrase (3)
 - Création d'une entrée dans PHR
 - Pour chaque code concept trouvé : création d'une entrée dans IDX
- Création du poids de chaque concept (4)
- Généralisation : génération automatique des concepts hyperonymes (5)
- Attribution de domaines nosologiques (6)

(1) Ce programme est le plus souvent appelé avec un URL, dans ce cas nous allons "aspirer" le document pour le stocker localement (en utilisant un module de Perl, décrit dans [Wong, 1997]).

Si le document est un URL de type frameset (un document HTML qui n'est qu'un appel à plusieurs frames), nous "aspirons" également les frames incluses, et créons un meta-document qui est la concaténation du frameset et de ses frames.

Si l'analyseur de langue trouve qu'il s'agit d'un document dans une autre langue que le français, nous créons un document qui est le résultat de sa traduction en français (cf. "Traduction automatique pour indexation", p. 101).

Si le document n'est pas dans un format interprétable (texte, RTF, PDF, XML), nous essayons de le traduire en HTML.

(2) et suppression d'une éventuelle ancienne version

(3) Ceci est important pour la pondération des concepts, en effet un concept trouvé dans le titre d'un document est beaucoup plus "important" qu'un concept trouvé dans le corps du texte. Notre analyseur HTML va donc attribuer un poids différent selon que la phrase se trouve dans des balises de type TITLE, META, H1, H2, H3, B (texte écrit en gras). Une phrase habituelle ayant le poids 1, nous avons attribué (de manière empirique) les poids respectifs 6,4,4,3,2,2.

- (4) Ces poids étant calculés selon le contenu du reste du corpus (cf paragraphe Pondération des poids des concepts p. 82). Il faut bien noter que l'insertion d'un nouveau document fausse un petit peu tous les poids dans les autres documents (le document courant n'étant pas pris en compte dans les autres). Cet écart reste cependant négligeable pour un seul document (en adaptant certaines formules, comme attribuer un poids spécial à un concept apparaissant pour la première fois), mais il faut régénérer tous ces poids quand le nombre de documents ajouté atteint une certaine limite.
- (5) Voir le chapitre suivant
- (6) Voir le chapitre "Attribution de domaine" (p. 92).

Notre base de données d'index des documents est maintenant générée (à partir du corpus de documents). Après la présentation d'une évaluation de l'intérêt de la généralisation nous expliquerons maintenant les diverses applications développées qui utilisent cette base d'index.

.VIII. Intérêt de la hiérarchie

Dans certains documents, des concepts sont implicites mais non présents. Pour nous en convaincre, nous avons fait l'étude suivante :

Nous avons étudié les documents EDICERF (cours de radiologie du Collège des Enseignants de Radiologie Français, [Duvaufferier et al., 1995]), ces documents sont en général très longs (plusieurs pages). Nous avons fait une recherche (en texte intégral) du mot "RADIO" sur tous ces textes, à notre grande surprise, quelques-uns ne sont pas apparus. Il s'agit donc bien de cours de radiologie ne contenant jamais le mot Radio (ni radiologue, radiologie, radiologique).

Exemples :

En Neuroradiologie :

- <http://www.med.univ-rennes1.fr/cerf/edicerf/NR/NR009.html> Accidents vasculaires cérébraux ischémiques
- <http://www.med.univ-rennes1.fr/cerf/edicerf/NR/NR012.html> Pathologie infectieuse encéphalique
- <http://www.med.univ-rennes1.fr/cerf/edicerf/NR/NR017.html> Imagerie cérébrale du sujet âgé

En Radiologie Uro-génitale :

- <http://www.med.univ-rennes1.fr/cerf/edicerf/UG/UG22.html> Pathologie fonctionnelle de l'ovaire et dystrophie ovarienne
- <http://www.med.univ-rennes1.fr/cerf/edicerf/UG/UG26.html> Repères anatomiques fœtaux
- <http://www.med.univ-rennes1.fr/cerf/edicerf/UG/UG27.html> Biométrie, âge gestationnel et croissance
- <http://www.med.univ-rennes1.fr/cerf/edicerf/UG/UG28.html> Premier trimestre normal et pathologique de la grossesse
- <http://www.med.univ-rennes1.fr/cerf/edicerf/UG/UG29.html> Bases séméiologiques des dysmorphies fœtales
- <http://www.med.univ-rennes1.fr/cerf/edicerf/UG/UG30.html> Le doppler en obstétrique

En Radio Anatomie :

- <http://www.med.univ-rennes1.fr/cerf/edicerf/RADIOANATOMIE/029.html> Biométrie fœtale

Un guide de radio anatomie ne contenant ni le mot "RADIO", ni le mot "ANATOMIE" !

Sur 94 documents, il y a donc 10 documents qui ne contiennent jamais le mot RADIO. Ce qui signifie qu'une recherche sur ces documents, sans traitement hiérarchique, entraîne 10 % de pertes !

Or, grâce à la hiérarchie (la taxonomie extraite du thésaurus), nous pouvons générer le concept "RADIO" si nous rencontrons des concepts comme : "SCANNER", "TOMO-DENSITOMETRIE", "MAMMOGRAPHIE"...

Cette étude montre bien l'intérêt d'un traitement hiérarchique sur les concepts trouvés dans les documents. Nous avons profité de la présence de liens hiérarchiques entre les différents concepts d'un thésaurus, pour générer les concepts hyperonymes des concepts trouvés dans le document par NOMINDEX.

.IX. *Système de Recherche d'Information*

À ce niveau se posait le problème du choix du modèle de recherche d'information, nous avons choisi le modèle vectoriel qui nous a semblé plus adapté que le modèle booléen. La raison principale est qu'il paraît simpliste d'appliquer une logique binaire à une recherche d'information (un document correspond ou ne correspond pas). Le modèle booléen a l'inconvénient de privilégier les longs documents (un document de 100 pages contient beaucoup de concepts différents et risque donc de correspondre très souvent aux requêtes) contrairement au modèle vectoriel qui pondère le résultat par le nombre et le poids des autres concepts du document. La recherche d'information sur un modèle booléen n'est pas intuitive, il faut parfois formuler des requêtes comme ("fièvre" ou "hyperthermie") et ("mal de tête" ou "céphalée"). De plus le modèle vectoriel permet de calculer des scores de similarité entre documents.

Le modèle vectoriel [Salton, 1971] propose de représenter un document comme un vecteur exprimé sur les dimensions représentées par les mots. Nous l'avons adapté pour représenter un document par un vecteur de concepts.

La critique que l'on peut faire du modèle vectoriel est qu'il considère chaque descripteur comme étant indépendant des autres, or un descripteur tel que "Coronaire" est loin d'être indépendant du descripteur "Cœur". Mais, dans notre système, ces deux descripteurs sont fusionnés, d'autre part, l'utilisation de la hiérarchie permet de ne pas considérer indépendamment deux descripteurs comme "Fente palatine" et "Malformations de la bouche". Par contre, un descripteur comme "Mal de tête" reste indépendant du descripteur "Tête".

Les coordonnées des vecteurs sur chaque dimension peuvent être exprimées en fonction du nombre d'occurrence du concept dans le document, mais nous préférons pondérer le concept en fonction de son importance.

.IX.1. Pondération des poids des concepts

Plutôt que de représenter le vecteur en fonction de la fréquence du concept dans le document, nous utilisons le score TFIDF [Salton et Buckley, 1988]. Ce score permet de donner une importance au concept en fonction de sa fréquence dans le document (TF = Term Frequency) pondérée par la fréquence d'apparition du concept dans tout le corpus (IDF = Inverse Document Frequency). Ainsi un concept très spécifique au document (n'apparaissant

que dans ce document) aura un score correspondant à sa fréquence d'apparition, par contre, un concept apparaissant dans tous les documents du corpus aura une pondération maximale.

La formule est la suivante :

$$TFIDF_{c,d} = TF_{c,d} \cdot IDF_{c,d} \quad \text{(Equation 1)}$$

avec:

- c : un concept
- d : le document

TFIDF voulant dire "Term Frequency, Inverse Document Frequency"

soit :

$$TFIDF_{c,d} = TF_{c,d} \cdot \left(\log_2 \frac{N}{DF_c} + 1 \right) \quad \text{(Equation 2)}$$

avec:

- c : un concept
- d : le document
- $TF_{c,d}$: la fréquence d'apparition du concept dans le document
- DF_c : le nombre de documents du corpus contenant le concept
- N : le nombre de documents du corpus

Ainsi, quand DF_c est égal à 1 (concept n'apparaissant que dans ce document), le TFIDF sera fort, quand DF_c est proche de N (concept apparaissant dans tous les documents), le TFIDF sera faible.

Après la phase d'indexation du corpus de textes, nous calculons donc, pour chaque concept dans un document, son score TFIDF. Nous verrons que, dans toutes les applications de l'indexation, nous utiliserons ce score TFIDF comme métrique de l'importance du concept dans le document.

Par contre, l'ajout d'un nouveau document dans le système nécessite de recalculer tous les scores TFIDF. Il s'avère néanmoins, que, lorsque le nombre de document est élevé, l'ajout d'un nouveau document ne modifie pas beaucoup les autres scores TFIDF. Le recalcul complet peut donc être différé.

.IX.2. Similarité

L'application la plus intéressante de ce modèle vectoriel est de pouvoir calculer le score de similarité entre un ensemble de concepts (extraits d'une phrase ou d'un document) et les autres documents. La première application consiste à utiliser une formule de similarité dans notre moteur de recherche. Ce moteur de recherche permet à l'utilisateur d'entrer une phrase en langage naturel, il indexe cette phrase en mots de référence, recherche les concepts du thésaurus correspondant et, pour chaque document du corpus, calcule le score de similarité avec cet ensemble de concepts. Le résultat de la recherche présentera donc les documents ayant le score le plus élevé.

La mesure de similarité utilisée est la formule *Cosine* qui calcule le cosinus de l'angle entre le vecteur représentant la requête de l'utilisateur et chaque document du corpus [Salton, 1983].

Voici la formule *Cosine* :

$$COSINE(d,r) = \frac{\sum_{c \in d \cap r} TFIDF_{c,d} \cdot TFIDF_{c,r}}{\sqrt{(\sum_{c \in d} TFIDF_{c,d}^2) \cdot (\sum_{c \in r} TFIDF_{c,r}^2)}} \quad \text{(Equation 3)}$$

Avec :

- d : Le document
- r : la requête
- c : un concept

D'autres mesures de similarité existent :

- La plus simple consiste à calculer le nombre de concepts communs entre la requête et le document (s'apparente à une recherche booléenne pondérée). Le défaut principal est d'attribuer le même poids à tous les concepts, or une recherche sur "fièvre et légionellose" devrait, implicitement, donner plus de poids aux documents contenant le concept "légionellose" que ceux contenant le concept "fièvre". L'autre défaut est de privilégier les longs documents (ceux qui contiennent le plus de concepts)
- Un peu plus sophistiquée : calculer la somme des produits TFIDF (qui permet de tenir compte de l'"importance" de chaque concept). Là encore, on privilégie les

longs documents (même si les concepts trouvés ne représentent qu'une petite partie du document)

- La mesure *Okapi*, basée sur un modèle probabiliste, qui permet de calculer les documents similaires à une requête [Robertson, 1994]
- Comme cela est préconisé dans [Wilkinson, 1994], on peut utiliser une combinaison linéaire des mesures précédemment citées.

Parmi ces mesures *Cosine* est la plus couramment utilisée. La comparaison *Cosine Okapi* est parfois contradictoire : dans [Bellot, 2000] elle donne les meilleurs résultats, d'autres expérimentations, comme [Savoy et Picard, 2000], [Savoy et Rasolofo, 2002] ou encore [Steinberger et al., 2002] ont de meilleurs résultats avec *Okapi*.

Nous avons préféré *Cosine* car elle pondère la somme des produits *TFIDF* par la taille de la requête et, surtout, par la taille des documents. Elle sera donc élevée si le document contient principalement les concepts de la requête. La pondération *TFIDF* permet, quant à elle, d'accorder plus d'importance aux concepts fréquents par rapport aux autres. Un autre avantage de *Cosine* sur *Okapi* est de pouvoir fonctionner de la même manière pour comparer des documents.

Cette formule *Cosine* sera également utilisée pour calculer la similarité d'un document par rapport à un autre (cf. chapitre suivant) ou pour calculer la similarité d'une phrase par rapport au reste du document (cf. chapitre "Synthèse automatique de document" p. 89).

Fonctionnement du S.R.I. :

A partir d'une phrase, le système recherche tous les mots de référence. La fonction de correction orthographique est automatiquement mise en œuvre pour les mots inconnus, si un mot est corrigé, un message d'alerte affichera à l'utilisateur la correction effectuée.

Ensuite, les concepts seront extraits de la phrase (ceux dont tous les mots appartiennent à la phrase). Si, parmi les mots de la phrase, un mot n'a contribué à aucun concepts un processus supplémentaire est mis en œuvre : il recherchera tous les termes de notre thésaurus qui contiennent ce mot (à condition que les concepts correspondants soient utilisés dans l'indexation).²⁷

²⁷ Par exemple, pour une phrase comme "peau rose", le système ne trouvera qu'un seul concept ("peau") dont tous les mots sont inclus dans la phrase, mais le mot "rose" n'a contribué à aucun concept, le système proposera

Ensuite, le système calcule la similarité (formule Cosine) entre chaque document de notre corpus et la liste de concepts détectés dans la phrase. Les documents seront présentés par ordre décroissant du score Cosine.

Exemple :

A la requête "Voyage et prévention de la malaria", le S.R.I. répondra :

Titre	URL	Score Cosine
Recommandations canadiennes pour la prévention et le traitement du paludisme...	http://www.hc-sc.gc.ca/hpb/lcdc/publicat/ccdr/00vol26/26s2/index_f.html	62.04%
Edisan Edisan Edisan	http://www.edisan.fr/	56.41%
Information sur la maladie : Paludisme, Programme de médecine des ...	http://www.hc-sc.gc.ca/hpb/lcdc/osh/info/pal_mal_f.html	52.97%
EuroSurveillance 1998; 3(3)	http://www.ceses.org/eurosurveillance/v3n4/en21-12.htm	51.32%
Santé Canada - RMTIC 25-6 Index	http://www.hc-sc.gc.ca/hpb/lcdc/publicat/ccdr/99vol25/rm2506f.html	50.29%
Recommandations sanitaires aux voyageurs Menu entete Recommandations sanitaires...	Http://www.sante.gouv.fr/hm/pointsur/voyageurs/index.htm	49.53%
Bienvenue A Sante Voyages Rouen	http://www.chu-rouen.fr/cap/svhome.html	47.03%
Faire reculer le paludisme	http://www.who.int/inf-fs/fr/am203.html	46.67%
Avis - «Idées fausses sur le paludisme et la méfloquine»	http://www.hc-sc.gc.ca/hpb/lcdc/osh/malradvf.html	46.08%
Lignes Directrices Concernant L'exercice De La Médecine Des Voyages	http://www.hc-sc.gc.ca/hpb/lcdc/publicat/ccdr/99vol25/25sup/dcc6.html	45.78%
SPS le paludisme	http://www.unicef.org/french/ffl/html/malaria.htm	45.21%
Paludisme	http://www.who.int/inf-fs/fr/am94.html	45.20%

Les concepts extraits de la phrase sont : "Paludisme", "Prévention" et "Voyage".

alors tous les concepts (utilisés) dont un des termes contient le mot "rose" (dans le thésaurus ADM : "Macule couleur rose", "peau couleur rose pale", "Pityriasis rose de Gigert").

.X. *Similarité de documents*

L'autre utilisation de l'indexation consiste à calculer, pour chaque document, son score de similarité par rapport à tous les autres, et ainsi de créer un réseau de proximité de documents. L'interface consistera à afficher les documents ayant le score le plus élevé.

Là aussi, nous utiliserons la formule *Cosine* comme un pourcentage de similarité d'un document par rapport à un autre. En effet, dans notre modèle vectoriel, un document est représenté par un vecteur, deux documents seront d'autant plus proches que l'angle des deux vecteurs sera proche de 0. D'où l'utilisation du cosinus de l'angle, dont la propriété est d'être proche de 1 quand l'angle formé par les deux vecteurs est proche de 0. Le cosinus est égal à 0 quand deux documents sont orthogonaux (c'est-à-dire qu'ils ne contiennent aucun concept identique).

La formule *Cosine* pour comparer deux documents d_1 et d_2 :

$$COSINE(d_1, d_2) = \frac{\sum_{c \in d_1 \cap d_2} TFIDF_{c, d_1} \cdot TFIDF_{c, d_2}}{\sqrt{(\sum_{c \in d_1} TFIDF_{c, d_1}^2) \cdot (\sum_{c \in d_2} TFIDF_{c, d_2}^2)}} \quad \text{(Equation 4)}$$

Avec:

- d_1, d_2 : Les documents
- c : un concept

Prenons l'exemple de deux documents (gastroentérologie et diabète) s'exprimant sur les trois dimensions (rein, pancréas et foie) :

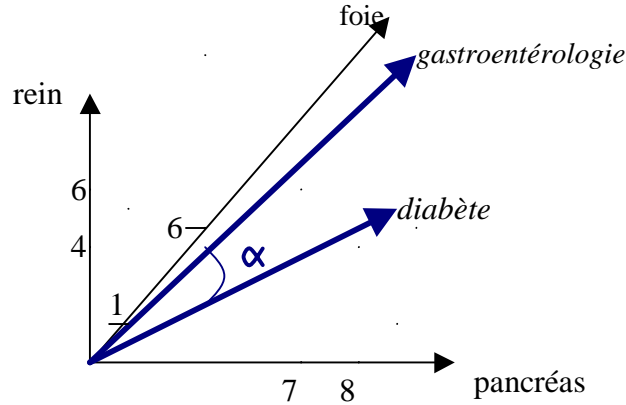


Figure 26 : Exemple de représentation de l'angle de deux vecteurs dans un espace à trois dimensions

Les deux documents seront relativement similaires (La formule *Cosine*, dans cet exemple, donnera $0,835^{28}$ ce qui est proche de 1, l'angle α faisant environ 22°).

Notons que, dans un corpus de N documents, il faudra, pour chacun des documents calculer N similarités, ce qui peut être long quand N est important. Nous avons, dans la base de données, pré calculé une table de similarité.

L'analyse factorielle des correspondances [Benzécri et al., 1973], permet également de représenter synthétiquement les documents sur un graphique à deux dimensions, et donc d'évaluer les distances, selon certains axes qu'il faut interpréter.

²⁸
$$\frac{8 \times 7 + 4 \times 6 + 1 \times 7}{\sqrt{(8^2 + 4^2 + 1^2) \times (7^2 + 6^2 + 7^2)}}$$

.XI. Synthèse automatique de document

Il existe, principalement, deux méthodes de synthèse de texte : par compréhension et par extraction. La première méthode, basée sur les méthodes d'intelligence artificielle et de compréhension automatique, se base sur une représentation sémantique du texte (modèles conceptuels) suivie d'une phase de réduction. La seconde se base sur des calculs statistiques de similarité de chaque phrase avec le document (pour en extraire les plus proches). La première méthode a été d'emblée écartée car nous ne disposons pas d'une représentation conceptuelle des connaissances suffisantes pour "comprendre" chaque document (nous connaissons les concepts des phrases mais pas les relations sémantiques entre ces concepts). La seconde méthode a l'énorme avantage de ne pas nécessiter beaucoup de ressources informatiques, à tel point que la synthèse automatique d'un document peut se faire en ligne (temps inférieur à une seconde).

.XI.1. Méthode

Desclès et Minel [Desclès et Minel, 2000] proposent de calculer pour chaque unité textuelle (phrase) un score de pertinence par rapport au reste du document. Le score le plus couramment utilisé est le fameux TFIDF (qui, rappelons-le, indique la fréquence du terme pondéré par la fréquence du terme dans tout le corpus). Ces auteurs proposent d'appliquer cette formule aux mots extraits du texte, nous l'avons appliquée aux concepts extraits par NOMINDEX. En appliquant la formule du cosinus, on calcule le coefficient de similarité entre la phrase et le document. Plus une phrase a un coefficient élevé, plus elle est "typique" du document. Cette formule *Cosine*, entre un document et une phrase, s'écrit :

$$COSINE(d, p) = \frac{\sum_{c \in d \cap p} TFIDF_{c,d} \cdot TFIDF_{c,p}}{\sqrt{(\sum_{c \in d} TFIDF_{c,d}^2) \cdot (\sum_{c \in p} TFIDF_{c,p}^2)}} \quad \text{(Equation 5 : Phrase/document)}$$

Avec :

- d : le document
- p : la phrase
- c : un concept
- $TFIDF_{c,d}$: score du concept dans le document (p : dans la phrase)

Il suffit, ensuite, de trier chaque phrase par ordre décroissant de similarité, et de n'en garder que les n premières (n pouvant être fixé par l'utilisateur).

À ce niveau, nous avons ajouté une amélioration, en effet, avec cette méthode appliquée aux concepts, il est fréquent que deux phrases distinctes soient codées avec les mêmes concepts. Deux phrases, identiques au niveau des concepts, risquent donc d'être extraites simultanément. Nous avons choisi de n'extraire que la première d'entre elles.²⁹

Les phrases extraites apparaissent à différents endroits du texte, nous les affichons dans l'ordre d'apparition, il est néanmoins très fréquent que le résultat apparaisse décousu, car nous ne tenons pas compte des relations existant entre chaque phrase.

[Kan et Klavans, 1998] proposent une solution à ce problème, ils construisent une segmentation linéaire des phrases, en analysant les co-occurrences de termes dans chaque phrase et la proximité textuelle (la chaîne est rompue si la phrase précédente est éloignée de plus de p phrases de la précédente). Ensuite ils appliquent la même méthode que précédemment, mais, cette fois-ci, sur les segments ainsi constitués. La synthèse résultat sera donc moins décousue.

²⁹ Une autre amélioration aurait pu consister à extraire la phrase la plus significative d'un document, puis à recommencer le processus avec le document duquel on a enlevé la phrase. Ceci afin d'éviter de présenter des phrases qui contiennent toutes les mêmes concepts.

.XI.2. Développement

Le programme que nous avons réalisé permet, pour un document donné, de faire une synthèse automatique de quatre phrases (par défaut, ce paramètre est laissé au choix de l'utilisateur).

Une présentation de cette synthèse automatique se trouve à l'adresse suivante :

<http://www.med.univ-rennes1.fr/cgi-bin/nomindex/resume.pl>

Par exemple, la synthèse du document " Les champignons",
(url: <http://www.expasy.ch/linder/APP/champignons.html>) est la suivante :

En médecine, à part l'effet bénéfique des antibiotiques, les champignons sont plutôt redoutés. Ils sont liés à différentes maladies et l'augmentation de personnes immunodéficient (AIDS, radiothérapie) et l'utilisation d'antibiotiques antibactérien (rôle de la flore normale) ont contribué à une augmentation d'occurrences d'infections par les champignons. Selon un article dans ASM News (Août 99), les infections par *Candida albicans* représentent la quatrième cause des infections nosocomiales. (...)

Dans les infections bactériennes nous avons vu l'utilisation d'antibiotiques qui s'attaquent à des cibles spécifiquement procaryotes. Les champignons sont des vrais eucaryotes et donc beaucoup plus difficile à combattre. (...)

Les infections myceuses ne sont que très rarement du à un contact de personne à personne. Souvent les infections de champignons sont dues à des spores ou des morceaux d'hyphae transporté par l'air

Cet exemple donne un bon aperçu du contenu du document, mais cette "méthode du mauvais élève" ne pourrait pas constituer un véritable résumé du document (même les fautes d'orthographe sont conservées !).

Une utilisation possible de cette synthèse automatique pourrait être d'aider à remplir les champs "Description" des metadata, comme le Dublin-core³⁰ [Darmoni et al., 2001].

³⁰ URL: <http://dublincore.org>

.XII. Attribution de domaine (catégorisation de documents)

.XII.1. introduction

Le but de cette réalisation est de catégoriser les textes médicaux, en attribuant automatiquement un domaine nosologique à chaque document (Endocrinologie, Cardiologie, ORL, ...).

La technique utilisée est celle des cooccurrences. Cette technique, purement statistique, utilise un corpus de référence pré-indexé. L'usage habituel se base sur le prédicat suivant : "Si un élément apparaît souvent dans les textes indexés avec l'index I , alors la probabilité qu'un texte contenant cet élément soit indexé avec I est forte"

Cette technique étant uniquement basée sur un corpus de textes, sa performance dépend évidemment de la qualité de l'indexation du corpus, de la quantité de documents et de sa variété (si l'on choisit comme corpus des comptes-rendus de psychiatrie, les résultats seront mauvais sur des textes de néphrologie).

.XII.2. Phase d'apprentissage

Au LIM, nous disposons de la base de connaissances ADM qui couvre une grande partie de la médecine clinique et sémiologique (hormis le domaine psychiatrique), d'autre part, chaque pathologie décrite dans l'ADM a été indexée par son domaine nosologique.

Nous avons donc ainsi un corpus relativement fiable, qui permet de connaître les domaines nosologiques d'une pathologie. Par extension, nous avons développé un outil qui permet d'attribuer un (ou plusieurs) domaine(s) à un symptôme, en se basant sur un prédicat "un symptôme apparaissant dans une pathologie classée dans le domaine D , est lui-même classé dans le domaine D ".

Nous créons ainsi une matrice Concept ADM / Nosologie. Chaque entrée contient une probabilité de classement du concept dans la nosologie.

Exemple : le concept "Hypoglycémie" est classée principalement en endocrinologie, puis dans les maladies iatrogènes, puis en cancérologie et hématologie. Le concept "Brûlures" est classé en dermatologie et traumatologie.

Pour les concepts de type pathologie (maladie, syndrome, forme clinique) nous prenons toutes les nosologies avec lesquelles la pathologie a été indexée.

Pour les concepts de type descripteur (signe de pathologie) nous prenons toutes les nosologies des pathologies dans lesquelles apparaît ce signe (le poids de chaque nosologie étant fonction du nombre d'apparitions du signe dans cette nosologie).

Restent maintenant les pathologies ADM qui ne sont pas décrites (ou non classées dans une nosologie particulière) et les signes qui ne sont pas utilisés. Nous appliquons deux traitements :

- 1) Si le concept possède, via la hiérarchie ADM, un concept hyperonyme qui est déjà classé, nous lui attribuons les nosologies de son hyperonyme. Exemple : le concept "Bassins déformés atypiques" a pour hyperonyme "Bassin ostéomalacique" qui est indexé dans la nosologie "Obstétrique", on lui attribuera automatiquement ce domaine.
- 2) Pour les autres concepts, on utilise NOMINDEX sur chacun des termes, pour en extraire les concepts inclus (c'est-à-dire les concepts dont tous les mots sont inclus dans le terme), et l'on attribue les nosologies de chacun des concepts inclus. Exemple : "Tumeurs osseuses malignes" inclu les concepts "Cancers" et "Os" qui sont classés dans les nosologies "Cancérologie" pour "Cancers", "Traumatologie" et "Gastro-entérologie" pour "Os". Le concept sera donc classé dans ces trois nosologies.

Pour le thésaurus ADM, il y avait 6 700 concepts non classés. À l'aide du second traitement nous en avons classé 5 330 supplémentaires. Après ces traitements, seulement 0,2 % des concepts ADM ne sont plus classés.

Voici un exemple représentatif du traitement d'attribution de domaine à un concept :







Le concept ADM "HISTOLOGIE TRACHEE" n'était pas utilisé en description de pathologie, notre premier traitement ne pouvait donc pas lui attribuer de nosologie. Avec NOMINDEX, on reconnaît les deux concepts inclus : "Histologie" et "Organe Trachée".

Le concept "Histologie" est classé dans les domaines suivants :






	Pneumologie	21,77%
	Cancérologie	19,35%
	Cardiologie, Angéiologie	4,84%
	Orl, Stomatologie	4,84%

(Suivent huit autres nosologies avec seulement 2,4 %)

Le concept " Organe Trachée" est classé en :

Nosologie	texte	score
	Orl, Stomatologie	30.00%
	Traumatologie	7.50%
	Pediatrie	7.50%
	Rhumatologie	7.50%
	Genet. Med., Cytogenetique	7.50%
	Cancerologie	7.50%

Le résultat sera donc pour "histologie trachée" la moyenne de chaque nosologie des deux concepts (pondérée par un facteur de 80% pour accorder un peu moins d'importance à un tel concept qu'aux autres)³¹ :

Nosologie	texte	score
	Orl, Stomatologie	19.35%
	Cancerologie	14.92%
	Pneumologie	12.10%
	Pediatrie	5.51%
	Genet. Med., Cytogenetique	5.51%

(Suivent huit autres nosologies avec moins de 5 %.)

³¹ Une amélioration de ce calcul consisterait à tenir compte également de l'"importance" de chacun des concepts inclus. Dans cet exemple, le concept "Organe trachée" est plus important que "Histologie" pour déterminer le domaine nosologique de "Histologie trachée". Le plus simple serait de pondérer par le nombre de fois que le concept est utilisé dans un corpus de référence. Ou encore tenir compte du nombre de domaines différents dans lesquels apparaît chacun des concepts.

.XII.3. Extension des domaines au thésaurus MeSH

Dans le MeSH, les concepts ne sont pas classés par nosologie, la méthode précédente ne peut donc pas être appliquée. Citons cependant la méthode de [Bodenreider, 2000] qui extrait de concepts MeSH les catégories de maladies UMLS (qui sont relativement identiques à ce que nous appelons domaines nosologiques). Ici, nous avons réutilisé le classement ADM pour classer les concepts MeSH, en trois étapes :

- 1) Nous faisons la correspondance "concept MeSH" ↔ "concept ADM" (cf. chapitre "Discussion" p. 130), pour les concepts ayant une correspondance nous leur attribuons les nosologies du concept ADM
- 2) Pour les concepts restants, nous appliquons l'amélioration présentée précédemment. C'est-à-dire que nous regardons les concepts ADM inclus dans chaque terme MeSH, et attribuons les nosologies de chaque concept ADM trouvé.

Sur 22 481 concepts MeSH, seulement 2 656 ont une correspondance directe avec un concept ADM. L'amélioration décrite précédemment classe 7 420 concepts supplémentaires.

.XII.4. Attribution d'un domaine à un document

À l'aide de cette matrice (concept / nosologie), nous sommes capables d'attribuer des domaines nosologiques à un ensemble de concepts. En l'occurrence, nous voulons connaître le domaine d'un texte, nous connaissons les concepts du texte, ainsi que leur "importance" (score TFIDF), nous allons donc faire un calcul en multipliant le score TFIDF de chaque concept avec les nosologies associées (en divisant le tout par le nombre de nosologies dans lesquelles apparaît le concept), la somme de ce calcul pour tous les concepts sera un score d'appartenance du texte au domaine.

Exemple:

Un document intitulé "Cours de Néphrologie du troisième cycle" (consultable à l'adresse : http://www.nephrohus.org/3_cycle_index.html) est classé automatiquement dans le domaine "Néphrologie".

En effet les concepts ayant les scores (TFIDF) les plus élevés sont :

HYPERTENSION ARTERIELLE, REIN, DIABETE, INSUFFISANCE RENALE AIGUE, GLOMERULONEPHRITE, VASCULARITE

Il est aussi logique de voir que le second domaine proposé est "Endocrinologie".









Nosologie	score
 Néphrologie	409.00
 Endocrinol., Nutr., Metab.	343.00
 Cardiologie, Angeiologie	221.00
 Maladies Iatrogenes	118.00
 Dermatologie, Venereologie	116.00
 Hematologie ,Immunologie	113.00
 Neurologie	113.00
 Gastro-Enterol., Proctol.	110.00

Figure 27: Classement du document "Cours de néphrologie" en domaines nosologiques

.XIII. Autres utilisations de l'indexation

Deux autres utilisations de l'indexation, et du modèle vectoriel, peuvent être mises en œuvre : la similarité de concepts et la classification de documents.

.XIII.1. Similarité de concepts

Nous pouvons adapter la formule *Cosine* pour calculer la similarité de deux concepts :

$$COSINE(c_1, c_2) = \frac{\sum_{d \in d_{c_1} \cap d_{c_2}} TFIDF_{c_1, d} \cdot TFIDF_{c_2, d}}{\sqrt{\left(\sum_{d \in d_{c_1}} TFIDF_{c_1, d}^2\right) \cdot \left(\sum_{d \in d_{c_2}} TFIDF_{c_2, d}^2\right)}} \quad \text{(Equation 6 : Concept/concept)}$$

Avec :

- c_1, c_2 : Les deux concepts à comparer
- d_{c_1}, d_{c_2} : Les documents contenant le concept c_1 (respectivement c_2)
- d : un document

Il s'agit ici de calculer le cosinus de l'angle formé par deux concepts (chaque concept étant représenté par un vecteur sur l'espace des documents).

Si la formule du cosinus peut s'appliquer directement aux concepts sur l'espace des documents sans recalcul intermédiaire, il est important de noter que le TFIDF n'est pas la meilleure mesure de l'"importance" d'un document pour un concept. TFIDF mesure l'importance d'un concept dans un document et non le contraire, il serait plus judicieux de recalculer ce score avec :

TF = nombre d'occurrence du concept dans le document

DF = Nombre de concepts que contient le document

La formule (équation 6, ci-dessus) peut néanmoins être appliquée aux concepts, bien qu'elle ne tienne pas compte directement du nombre de documents communs aux deux concepts. Nous avons ajouté deux facteurs à cette formule *Cosine*, l'un tenant compte de l'"importance" du concept, l'autre du nombre de documents communs. Cette formule de similarité est détaillée en annexe 4 ("Pondération de la similarité de deux concepts").

Nous avons tout d'abord expérimenté cette fonctionnalité sur les 500 documents du réseau pédagogique rennais... avec peu de succès ! En effet, le nombre de documents était trop peu important pour que les co-occurrences de concepts soient suffisamment significatives. Les co-occurrences sont calculées sur les documents entiers, ce qui réduit encore les chances de trouver des co-occurrences sémantiquement significatives.

Par contre, la récente création d'une base de données de 7000 documents extraits du CISMeF contient des co-occurrences de concepts sémantiquement plus pertinentes. Aucune expérience exhaustive n'a cependant été mise en œuvre, voici néanmoins quelques exemples qui montrent la pertinence de cette fonctionnalité.

Voici deux exemples de similarité de concepts :

Pour le concept "Déformations du pied" :

Concept similaire	Score pondéré	Nb de docs communs
PIED	30.91	26
PIED PLAT	23.63	13
TISSU CONJONCTIF	15.00	26
ORTEIL	14.58	12
MALADIES NEONATALES ET MALFORMATIONS	13.45	24
MALFORMATIONS APPAREIL LOCOMOTEUR	13.40	16
DEFORMATIONS CONGENITALES PIED	13.32	10
REGION CORPS	12.49	30
ANOMALIES CONGENITALES MEMBRES	11.51	13
PIED BOT	8.94	6
SIGNES ET SYMPTOMES, ETATS PATHOLOGIQUES	8.75	30
ORTHOPEDIE	8.65	14
CHAUSSURES	8.49	7
TENDONS	8.30	10
MALFORMATIONS	8.11	21

Table 2 : Concepts similaires à "Déformation du pied"

Pour le concept "Sciatique", on trouve les concepts similaires suivants :

Concept similaire	Score pondéré	Nb de docs communs
NEVRALGIE	22.72	29
INJECTION	16.80	13
HERNIE DISCALE	14.46	6
RACHIS	13.63	24
MANIFESTATIONS NEUROLOGIQUES	12.73	29
NERF SCIATIQUE	11.87	9
LOMBALGIE	11.34	9
TISSU CONJONCTIF	11.24	29
RHUMATISMES	11.11	19
SIGNES ET SYMPTOMES	10.71	29
NEUROLOGIE	10.21	20

Table 3 : Concepts similaires à "Sciatique"

En annexe 5 ("Quelques exemples de similarités de concepts"), on pourra consulter 11 exemples détaillés de similarité, sur les concepts : "Déformations pied", "Ecchymose", "Embolie artérielle", "Fente palatine", "Fièvre Q", "Lèpre", "Néphroblastome", "Protéines de transport", "Sciatique", "Scoliose" et "Variole".

.XIII.2. Classification/catégorisation de documents

Les méthodes utilisées pour la classification (ou catégorisation) automatique de documents auraient pu être utilisées avec les concepts plutôt que les mots. Cela n'a pas été mis en œuvre car les méthodes sont nombreuses, et l'analyse factorielle des correspondances (cf. chapitre correspondant p. 126) qui aurait pu être poursuivie par une classification de documents montre l'amélioration apportée par de notre indexation. L'attribution automatique de domaines nosologiques (méthode statistique de classification de documents par apprentissage sur corpus) est un exemple d'utilisation de notre indexation pour la catégorisation de documents.

.XIV. Traduction automatique pour indexation

.XIV.1. Introduction

Le but initial de notre outil est d'indexer des textes médicaux francophones, or l'usage de l'anglais est de plus en plus fréquent. Même sur le réseau pédagogique rennais, quelques textes anglais sont maintenant indexés (bien que rédigés par des professeurs francophones). Il serait dommage de ne pas pouvoir indexer de tels textes avec notre outil.

Nous disposons d'un lexique uniquement français, nous ne pouvons donc pas l'utiliser sur des textes non francophones. La seule solution consiste à traduire le texte en français automatiquement avant l'indexation. Le recours à des traducteurs automatiques (comme SYSTRAN©) aurait pu résoudre notre problème (sachant que la traduction automatique n'est jamais parfaite). Cependant, le meta-thésaurus UMLS (comme beaucoup d'autres thésaurus médicaux) est, en partie, multilingue. Nous avons, à titre expérimental, voulu exploiter cette information multilingue dans le but de traduire uniquement les termes qui nous serviront ensuite à l'indexation.

En aucun cas, nous n'avons voulu écrire un traducteur du langage naturel.

La construction du lexique de traduction doit se faire de manière automatique.

Voici les "a priori" :

- Nous supposons que tous les textes seront des textes médicaux (par conséquent le vocabulaire est limité)
- Il est impossible de construire nous-même des lexiques.
- Le traducteur se contentera de traduire "mot à mot" les textes. Aucune analyse syntaxique ou linguistique ne sera utilisée.
- L'existence d'un nombre important de mots d'origine gréco-latine en médecine facilite la tâche.

.XIV.2. Constitution des lexiques

L'idée de départ est qu'il existe plusieurs thésaurus médicaux qui sont traduits (les concepts possèdent des termes anglais parallèlement à des termes français, espagnols, allemands ...).

Il devient alors relativement facile de mettre en correspondance un terme français avec son homologue anglais.

Étant donné qu'il s'agit d'une traduction mot à mot (plus exactement "terme à terme"), et que le vocabulaire est volontairement limité au domaine médical, nous avons fait le choix suivant : "à un terme français correspond un et un seul terme anglais". Ce qui est très réducteur, mais, sans cela il aurait fallu mettre en place une véritable analyse linguistique, ce qui nécessiterait un moteur de traduction autrement plus complexe ! (et dépasserait l'objectif purement expérimental de notre outil).

Pour constituer notre lexique, nous nous basons sur l'UMLS qui comporte, pour chaque concept, ses différents termes traduits.

.XIV.3. Première étape

Prendre les termes ne comportant qu'un seul mot pour faire une équivalence :

un mot français \leftrightarrow un mot anglais

(et, pour le lexique inverse : un mot anglais \leftrightarrow un mot français)

On se rend compte rapidement que cela est grandement insuffisant.

Donc on constitue un second lexique, que l'on appellera lexique des mots multiples :

plusieurs mots français \leftrightarrow un terme anglais

qui permet, par exemple, de traduire : "SMALL INTESTIN" en "INTESTIN GRELE"

La reconnaissance des mots multiples est légèrement plus complexe que celle des mots simples...

Par exemple :

"ABDOMEN X-RAY" peut aussi s'exprimer par "X-RAY OF ABDOMEN"

Nous avons défini, pour résoudre ce problème, un ensemble de mots "nuls" (en l'occurrence, en anglais, il s'agit de "THE", "FROM", "OF", "S", "AT") qui ne sont pas pris en compte dans la recherche de mots multiples, la recherche des mots multiples se fait sur l'ensemble des composants quel qu'en soit l'ordre. Les problèmes posés par l'exemple précédent sont donc résolus. L'optimisation de la reconnaissance de mots multiples s'apparente à l'algorithme utilisé dans la reconnaissance de mots multiples (cf. Algorithme 1, p. 61).

.XIV.4. Enrichissement du lexique

L'enrichissement du lexique est relativement simple, il se fait à partir des mots multiples traduits. On essaie de trouver une traduction d'un mot d'après les termes complexes qui le contiennent.

Exemple :

Voici quelques termes traduits :

Français	Anglais
SYSTEME	SYSTEM
GLANDE	GLAND
GLANDE ENDOCRINE	ENDOCRINE GLAND
SYSTEME ENDOCRINE	ENDOCRINE SYSTEM

Table 4 : Exemple de traductions extraites du MeSH

Grâce aux deux premiers termes, on connaît les traductions des deux mots simples "SYSTEM" et "GLANDS", si on suppose ne pas connaître la traduction de "ENDOCRINE"

un traitement d'enrichissement du lexique reconnaîtra, à l'aide des deux derniers termes, que la traduction de "ENDOCRINE" est, tout simplement, "ENDOCRINE". Ce traitement doit bien-sûr être supervisé car ceci est faux dans un certain nombre de cas (exemple: "Intestin grêle" qui est traduit par "Small intestine" risque de suggérer la traduction erronée de "small" en "grêle" !)

.XIV.5. Méthode de traduction

- Aucun effort n'a été fait pour respecter la syntaxe (l'adjectif risque d'être souvent devant le substantif !) - Si un mot est inconnu, on teste tout de même la traduction du pluriel (ou du singulier dans le cas d'un mot déjà au pluriel).
- Le texte traduit comportera en majuscules les mots traduits et en minuscules les mots non traduits.

.XIV.6. Quelques statistiques sur le lexique

Pour les traductions "Anglais ↔ Français", le lexique contient environ 43543 traductions de mots simples qui font en moyenne 9,7 caractères chacun. Il contient 13666 expressions de 2,2 mots (et de 18 caractères) en moyenne, soit un total de 57209 entrées. La vitesse de traduction est d'environ 95 mots à la minute (3 minutes 20 secondes pour un texte de 19062 mots).

Environ 10 à 15 % des mots d'un texte médical sont inconnus du lexique, selon une évaluation que nous avons faite sur 100 textes médicaux extraits du réseau pédagogique rennais.

.XIV.7. Langues disponibles

A l'aide de l'UMLS, nous avons pu constituer des lexiques de traductions pour les couples de langues :

- Anglais ↔ Français
- Espagnol ↔ Français
- Portugais ↔ Français
- Italien ↔ Français

- Allemand ↔ Français

Cependant, nous nous sommes concentré sur le lexique anglais, bien que la traduction fonctionne relativement bien pour les trois autres langues latines (italien, espagnol et portugais). Par contre, et cela mérite d'être souligné, notre système fonctionne très mal pour l'allemand, en effet, les préfixes et suffixes sont si importants qu'un mot aussi simple que "KREBS" ("cancer") est tout simplement absent de l'UMLS (par contre on trouvera : "BASALZELLKREBS" – "cancer basocellulaire" - ou "MAGENKREBS" – "cancer gastrique").

.XIV.8. Exemple de traduction

Pour un texte médical de description d'un cas clinique. Sur 179 mots que compte le texte (116 mots différents) il n'y a que 11 mots inconnus du lexique.

Voici les 11 mots non reconnus : *blotchy, brought, chills, cramping, detailed, driving, epiodes, experiencing, intubated, workup, yo*

("epiodes" étant une faute d'orthographe, "yo" une abréviation de "years-old")

Exemple (première ligne du texte en question):

"Ms. Smith is a 54 yo admitted for nausea, vomiting, diarrhea, and abdominal cramps."

Est traduit automatiquement en :

"M. SMITH EST UN 54 yo ADMIS POUR NAUSEE, VOMISSEMENT, DIARRHEE, ET ABDOMINALE CRAMPES."

Cette traduction ressemble à une phrase d'"Astérix chez les Bretons", par contre l'indexation de la phrase traduite automatiquement sera le plus souvent identique à celle qui aurait été faite sur la même phrase correctement traduite. Ce qui est bien notre but.

.XIV.9. Reconnaissance de la langue d'un texte

Nous disposions de différents lexiques de traduction, nous avons rapidement développé un module de reconnaissance de la langue d'un texte (aussi appelé "devineur de langue") en faisant des statistiques sur les mots du texte appartenant aux différents lexiques (Anglais, Français, Espagnol...). Ce module est intégré au système d'indexation, son seul avantage est d'être très adapté pour le domaine médical (puisque'il a été construit à partir du thésaurus MeSH essentiellement). Par contre, il aurait été plus judicieux d'utiliser des méthodes statistiques de reconnaissance de la langue selon les *N*-grammes ou selon les mots-outils contenus dans le texte. Citons notamment un module Perl créé par [Piotrowski, 2000] (basé sur l'algorithme décrit dans [Dunning, 1994]) qui permet d'"entraîner" la reconnaissance sur un corpus de texte et qui est basé sur une indexation en trigrammes³². Citons également [Giguet, 1995] qui propose une méthode de reconnaissance basée sur les mots-outils d'un texte. Dans [Grefenstette, 1995] la comparaison des deux méthodes (*N*-grammes et mots-outils) montre que la méthode basée sur les *N*-grammes donne des résultats légèrement meilleurs, mais entraîne un temps de calcul plus important.

Bien que ce moteur de traduction automatique ne soit pas très au point, nous l'avons tout de même intégré au moteur d'indexation (quand un document anglophone est inséré dans le réseau pédagogique rennais il est préalablement "traduit" avant d'être indexé par NOMINDEX). Nous l'avons également intégré au moteur de recherche, un utilisateur peut maintenant poser une question en anglais, elle est préalablement traduite en français avant d'en extraire les concepts qui serviront à la recherche. Un exemple est donné sur la figure suivante.

³² Module Lingua::Ident, URL: <http://search.cpan.org/search?dist=Lingua-Ident>

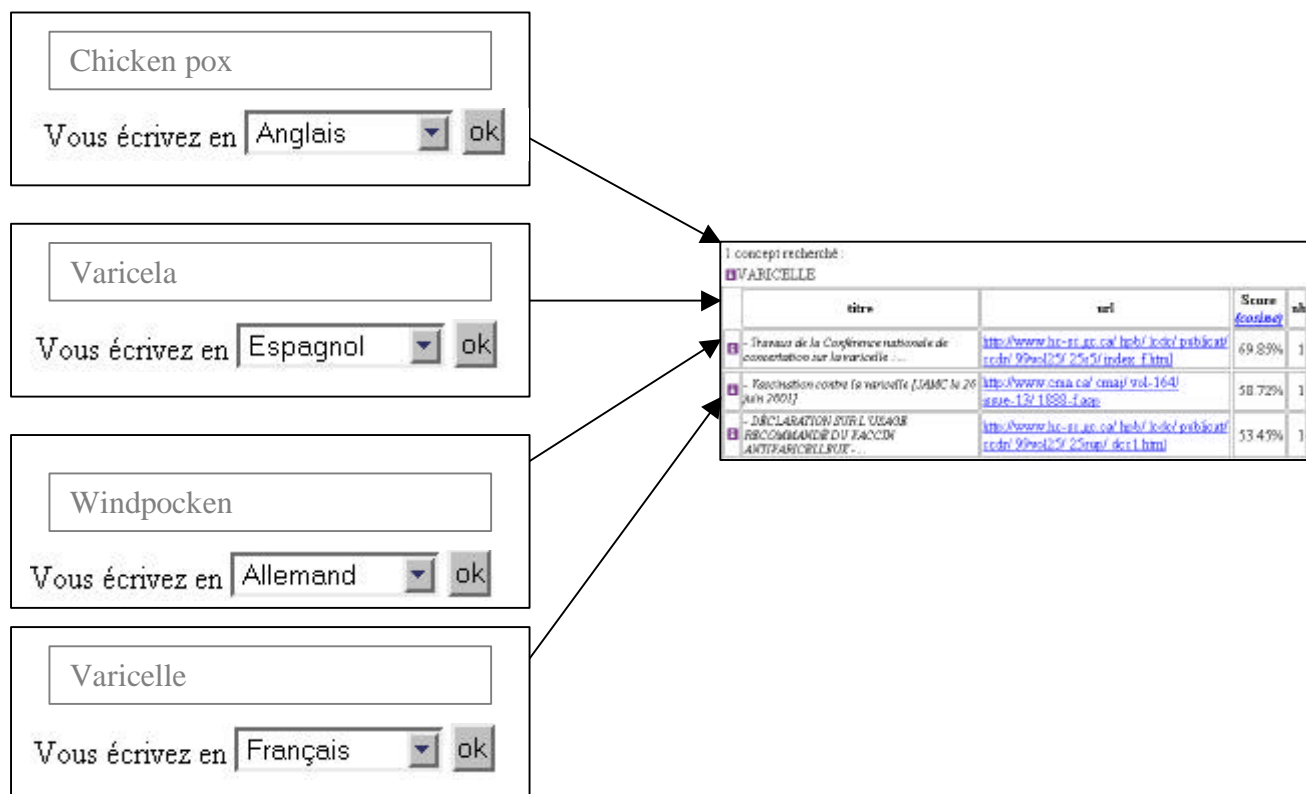


Figure 28 : Exemple d'interrogation multilingue

Après ce tour d'horizons des différentes applications développées, nous avons essayé de faire une évaluation de chacun de ces outils (quand cela était possible).

Résultats

Une base de données d'index utilisant l'ADM comme thésaurus cible a été alimentée par tous les documents du réseau pédagogique rennais (qui contient environ 500 documents). Une autre base de données a été constituée avec cette fois-ci le MeSH comme thésaurus cible, elle a été alimentée par près de 7000 documents (le catalogue des documents est issu du CISMef). Les différents outils utilisant ces index sont disponibles à l'adresse : <http://www.med.univ-rennes1.fr/nomindex/>

Le système d'indexation, d'un point de vue quantitatif, répond aux besoins exprimés, car il peut indexer 100 documents en moins d'une demi-heure, quant aux diverses utilisations, elles sont toutes assez rapides pour être faites en temps réel (par exemple, le temps de réponse d'une recherche sur tout le corpus est inférieur à trois secondes). La création d'un index de près de 7000 documents montre que notre produit est robuste. Les performances sont bien sûr affectées par le nombre de documents du corpus, mais restent tout à fait convenables (sur un serveur qui est par ailleurs déjà très sollicité par environ 75000 connexions web par jour)³³.

La réalisation principale, la plus visible par l'utilisateur du système, est un moteur de recherche couplé à un "baladeur" disponibles tous les deux sur une interface Web. L'utilisateur entre une requête, les documents correspondants sont affichés (cf. chapitre "Système de recherche d'information"), puis il peut accéder à la description de chaque document, à savoir :

- Son titre, son URL, le nombre de phrases indexées, le nombre de concepts différents trouvés...
- Les mots-clés détectés
- Les domaines nosologiques automatiquement attribués
- La possibilité de consulter les documents identiques (fonction de similarité de documents)
- La possibilité de faire une synthèse automatique du document
- Quand le thésaurus cible de l'indexation est l'ADM, la possibilité d'évoquer un diagnostic sur les concepts les plus importants du document
- Puis, pour chaque phrase du document, les concepts détectés.

³³ Statistiques du serveur www.med.univ-rennes1.fr disponibles à l'URL : http://www.med.univ-rennes1.fr/stat/med/stat_actuelle

Quand un concept est présenté à l'utilisateur, celui-ci peut accéder à sa description, à savoir:

- Les termes du concept
- La décomposition en mots de chaque terme
- La hiérarchie liée au concept (concepts hyperonymes, concepts hyponymes)
- Les documents du corpus qui contiennent ce concept (une autre option consiste à afficher les phrases dans lesquelles ce concept apparaît)
- Les domaines nosologiques de ce concept
- Les concepts proches

Un exemple d'affichage est développé en annexe 6 ("Présentation du "baladeur" NOMINDEX").

Un outil permet de comprendre le fonctionnement de NOMINDEX, il s'agit d'une interface Web qui interprète une phrase (pouvant être écrite dans une autre langue que le français) et qui se contente d'afficher les concepts détectés³⁴, par exemple :

Une interrogation avec "My teeth are painful", sera préalablement traduite puis interprétée par NOMINDEX, qui affichera :

Phrase originale	Phrase traduite	Concepts détectés
"my teeth are painful"	mon dent sont douloureux	DENTITION DOULEUR DENT ODONTALGIE

Avec, pour chacun des concepts, un hyperlien permettant d'accéder au baladeur.

³⁴ Accessible à l'adresse : <http://www.med.univ-rennes1.fr/cgi-bin/nomindex/exemple.pl>

.I. Evaluation

L'évaluation de la qualité de l'indexation elle-même est une tâche énorme, pour ne pas dire impossible. Par contre les utilisations de cette indexation sont plus proches du raisonnement humain, et donc, plus faciles à évaluer.

Pour évaluer la qualité des systèmes de recherche d'information ou l'attribution de mots-clés, les critères utilisés sont le silence et le bruit (critères à minimiser) ou leurs compléments : le rappel et la précision (critères à maximiser). Par contre les performances d'autres outils comme la similarité de documents, la synthèse automatique ou la traduction sont plus difficiles à quantifier, l'évaluation restera donc très subjective.

.I.1. Évaluation du lexique

L'idée est d'observer le fonctionnement du lexique sur des interrogations faites par des acteurs ne connaissant rien de sa structure (mais l'utilisant dans un contexte médical).

Comme nous l'avons vu lors de la présentation des outils existants (p. 46). Diverses interrogations sont faites sur notre moteur de recherche en texte intégral. Ces interrogations se font en langage libre. Chaque requête laisse une "trace" dans le fichier de journalisation de notre serveur Web. Nous avons créé une routine qui extrait ces interrogations et qui ré-interroge le lexique d'indexation (dans son état actuel) avec chacune de ces phrases. Ensuite, nous avons fait une évaluation de ces résultats. Un travail a récemment consisté à ajouter dans le lexique d'indexation les mots qui étaient inconnus du MeSH français. Nous avons alors fait une seconde évaluation avec cette nouvelle version du lexique.

Voici, en résumé, l'analyse statistique de ces résultats :

Sur 2331 interrogations (faites pendant le premier semestre 2001), 150 contenaient des mots inconnus du lexique, et 198 contenaient des mots automatiquement corrigés par notre outil.

Sur cet ensemble de phrases, il y en a donc **6,43 %** (150/2331) qui ne sont pas reconnues par le lexique. Mais un examen plus approfondi nous montre que les phrases en question comportent beaucoup de fautes de frappe ou d'orthographe, il est donc normal que le lexique ne les connaisse pas.³⁵

³⁵ Etrangement, peu de noms propres ont été utilisés dans ces interrogations

Sur les 198 mots corrigés automatiquement (par notre fonction de correction automatique, cf. p. 63), il s'agit effectivement, la plupart du temps de fautes d'orthographe (ou de frappe) :

- Grinsser des dents la nuit
- Plaquerouge
- Diarre

Notons que dans 42 cas il s'agissait de mots correctement orthographiés, mais inconnus du lexique, et corrigés à tort (comme "indexés" corrigé en "indènes").

Par contre, sur les 150 mots inconnus, 92 mots étaient des fautes de frappe ou d'orthographe (qui n'ont pas pu être corrigées automatiquement). Notons, au passage, que cette évaluation montre qu'il y a 248 fautes de frappe (198-42+92) sur 2331 requêtes (soit 10,63% d'erreurs!).

Reste donc 58 phrases qui sont orthographiquement correctes et qui ne sont pas reconnues

Exemples :

- Bosse de buffle
- Coliforme
- Cheville enflée
- Serrement de l'estomac
- Sialophagie
- Ennui
- Anxiolytique
- Douleur en avalant

Il est intéressant de noter que, depuis la nouvelle version du lexique, seuls "Buffle", "Serrement", et "Ennui" sont maintenant inconnus. Cf. le chapitre "Mots inconnus" p. 63, pour la méthode utilisée.

Le résultat de cette première évaluation montrait donc que **2,6 %** (58/2331) des interrogations n'étaient pas reconnues par le lexique (hors fautes d'orthographe).

Avec la nouvelle version du lexique, il ne reste maintenant plus que 102 phrases ayant des mots inconnus, sur ces 102 phrases seules 12 ne contiennent pas de fautes d'orthographe. Nous avons donc maintenant un taux de **0,5 %** de mots inconnus.

Les résultats, s'ils ne sont pas parfaits, montrent bien que notre lexique "comprend" très bien ces requêtes. Notons tout de même, que nous n'avons en aucun cas essayé d'évaluer notre lexique avec des textes entiers en langage naturel (le résultat, sans aucun doute, serait désastreux), car notre but n'est pas d'avoir un lexique qui "comprend" tous les mots utilisés dans la langue française, mais bien d'avoir un lexique qui reconnaît les termes médicaux utilisés dans un texte.

Il est important de noter que, parmi les mots non reconnus, la plupart sont des mots du français courant (hors du domaine médical).

.I.2. Évaluation du système de recherche d'information

Il est toujours difficile d'évaluer un S.R.I. sans disposer d'élément de comparaison. Nous avons choisi d'évaluer notre S.R.I. en le comparant avec un moteur de recherche en texte intégral afin de mesurer l'amélioration apportée par notre outil. Nous sommes bien conscients que ceci ne constitue pas une évaluation complète du système, nous n'avons pas évalué la validité des réponses c'est à dire le bruit, (il aurait fallu consulter chaque document pour vérifier s'il correspond réellement à chaque requête).

Nous avons extrait toutes les interrogations qui ont été faites sur notre moteur de recherche en texte intégral et nous avons interrogé notre nouveau moteur de recherche avec ces mêmes requêtes, afin de restreindre notre évaluation nous n'avons considéré que les requêtes apparues plus d'une fois, et nous avons interrogé les documents se trouvant dans la spécialité "pédiatrie" du réseau pédagogique rennais. Les résultats de cette évaluation sont détaillés en annexe (voir annexe 7, "Comparaison de la recherche en texte intégral et la recherche par concepts NOMINDEX").

Grâce à l'apport du dictionnaire ADM les regroupements de mots diminuent considérablement le silence dans les réponses. Voici quelques exemples de recherches fructueuses grâce à l'apport des synonymes :

<i>Requête formulée par l'utilisateur:</i>	<i>Mots trouvés dans les textes retournés par NOMINDEX</i>
coloscopie	Coloscopie, Endoscopie colon
athérosclérose	Athérome, Athéromateux, Athéromateuse, Athéromatose
VEINE POUMON (noter que la recherche en texte intégral ne trouve aucun texte)	Veine pulmonaire, Veines pulmonaires "Retour <u>veineux pulmonaire</u> anormal" "L'eau passe vers le sang <u>veineux pulmonaire</u> " "Surcharge <u>veineuse pulmonaire</u> ".
transfusion sanguine	Transfusion sanguine, transfusion sang , Transfusion de produits sanguins, Les transfusions de sang, "On peut reconnaître la <u>transfusion</u> foëto-maternelle par l'existence d'hématies foëtales dans le <u>sang</u> de la mère"

S'il reste bien quelques requêtes ne fonctionnant pas sur le nouveau moteur ("Foie", "Prématuré", "Vitamine D", cf. annexe 7), elles sont, en proportion, peu fréquentes (cette analyse permet d'ailleurs de mettre en évidence certaines carences du lexique, que l'on pourra corriger par la suite).

Comme cela est décrit en annexe 7, une évaluation de la précision des requêtes sur plus de 100 documents pédiatriques du réseau pédagogique rennais montre une très nette amélioration avec le nouveau moteur de recherche.

L'ancien moteur retournait **60%** des documents pertinents à ces requêtes, NOMINDEX en retourne maintenant **93%**.

Plus précisément, sur l'évaluation faite, le moteur de recherche retourne 92,2 % des documents pour NOMINDEX-ADM, et 93,6 % pour NOMINDEX-MeSH. La comparaison des deux thésaurus comme cible de l'indexation est à peu près équivalente, cependant les requêtes formulées sont assez générales et ne représentent pas de manière exhaustive toutes les requêtes formulées (car nous avons sélectionné les questions fréquentes). Notons que l'ADM semble mieux fonctionner que le MeSH pour des questions sur des pathologies ou symptômes ("Maladie de Hurler", "Myélome"), et le MeSH pour des questions plus générales comme "Infections nosocomiales", "Chirurgie". Le thésaurus idéal pourrait être le résultat de l'union des deux thésaurus...

Il aurait été intéressant de comparer NOMINDEX à un moteur de recherche en texte intégral plus sophistiqué que celui que nous avons développé. Mais une telle évaluation est lourde à mettre en œuvre et les moteurs de recherche sont nombreux.

Noter également qu'une recherche en texte intégral est presque irremplaçable quand un utilisateur recherche des noms propres (un étudiant recherchant les cours écrits par un professeur par exemple). C'est pourquoi, pour le réseau pédagogique rennais, quand des mots de la requête sont inconnus ou ne participent à aucun concept, le système propose de lancer la même requête sur le moteur de recherche en texte intégral.

.I.3. Extraction de mots-clés

.I.3.1. Comparaison avec un autre outil pour 70 documents

Notre outil a été utilisé pour comparer l'indexation automatique avec l'indexation manuelle de documents médicaux sur le portail de site médicaux francophones du CISMéF [Darmoni et al., 2000]. Dans ce cadre, il a été comparé avec un outil commercial (Medikeo³⁶), qui utilise une méthode statistique d'apprentissage par co-occurrences de mots.

L'équipe CISMéF indexe systématiquement chaque nouveau document médical par un certain nombre de mots-clés, ces mots-clés sont tous issus du MeSH (ce qui facilitera une comparaison entre l'indexation manuelle et une indexation automatique), dans ce cadre nous avons indexé ces mêmes documents avec le MeSH comme thésaurus cible.

Une première évaluation (sur 70 documents) a été réalisée par l'équipe CISMéF, elle est détaillée dans l'annexe 8 ("Evaluation des mots-clés proposés par NOMINDEX, comparaison avec l'indexation manuelle du CISMéF"), en voici cependant un résumé:

L'objectif a été de connaître l'apport des outils Medikeo et NOMINDEX par rapport à une indexation manuelle de différentes ressources. Parmi les sites indexés manuellement dans le catalogue CISMéF, 70 documents ont été extraits aléatoirement par l'équipe de documentalistes et ont été soumis aux deux moteurs. Une documentaliste de l'équipe CISMéF a procédé à une comparaison systématique entre les concepts MeSH de l'indexation manuelle faite par l'équipe CISMéF, et la liste de termes proposée par les deux outils (Medikeo ou NOMINDEX). Seuls les 20 premiers concepts ont été pris en compte. A noter que le nombre de concepts pertinents mais absents de l'indexation manuelle a permis de mesurer le silence de cette indexation manuelle.

Les résultats sont :

Pour Medikeo : 85 % de bruit, 15 % de pertinence, **0,8 %** de silence de l'indexation manuelle

Pour NOMINDEX : 83 % de bruit, 17 % de pertinence, **3,2 %** de silence de l'indexation manuelle

Le taux de bruit important des deux systèmes nous paraît justifié du simple fait de prendre les vingt premiers concepts proposés, étant donné qu'en moyenne, il y a seulement 2,47 concepts indexés manuellement par document.

³⁶ <http://www.medikeo.com/>

Le troisième chiffre nous paraît le plus important dans le cadre de l'utilisation de l'outil comme aide au codage de l'information. Il s'agit là de concepts auxquels l'indexeur n'avait pas pensé. Plus important, une étude de ces nouveaux concepts montre que, dans près d'un document sur deux, un concept pertinent mais non codé est proposé par notre système (ici dans 34 des 70 documents étudiés).

Rappelons que l'outil NOMINDEX n'a jamais été "entraîné" sur les sites du CISMeF. Son fonctionnement, par extraction et non par apprentissage, en fait un outil plus adapté à de nouveaux documents, utilisant de nouveaux concepts. En effet, les programmes fonctionnant par apprentissage sont "entraînés" sur un corpus de textes pré-indexés, ils créent ainsi une base de données contenant la probabilité d'appartenir à un concept selon les mots que contient chaque texte. Ils sont donc, de par leur fonctionnement, incapables de fonctionner sur des concepts n'ayant jamais été rencontrés dans la phase d'apprentissage. NOMINDEX possède un avantage certain à ce niveau, la comparaison des performances des deux moteurs doit donc aussi en tenir compte (bien que l'évaluation montre que NOMINDEX se révèle déjà un peu supérieur à Medikeo).

Voici, à titre d'exemple, l'étude détaillée sur deux exemples :

Premier document: "Notions d'hygiène hospitalière", se trouvant à l'adresse suivante:

<http://www.md.ucl.ac.be/entites/esp/hosp/cours/HH0.htm>

Nous présentons, par ordre d'apparition les 11 premiers concepts trouvés par NOMINDEX

Validité	Concept détecté	Score TFIDF	Commentaires
non adapté	SEPSIS	89,64	Sepsis est, très abusivement, étiqueté comme synonyme de "infecté" dans le dictionnaire ADM
ok ++	INFECTION HOSPITALIERE	64,04	Ce concept est pertinent mais non codé par l'indexeur CISMeF
Oui	HYGIENE	51,57	
Oui	DESINFECTANTS	36,73	
non adapté	HOPITAL	32,82	
non adapté	AIDE FINANCIERE	24,01	
Oui	PREVENTION	23,88	
trop précis	SOINS INTENSIFS	20,70	apparaît dans le texte, mais n'est pas à mettre en mot-clé
Oui	DESINFECTION	19,77	
trop précis	URINAIRE, INFECTION	19,01	apparaît dans le texte, mais n'est pas à mettre en mot-clé
Oui	STERILISATION	14,41	

Il y a donc quatre concepts proposés par NOMINDEX qui étaient codés manuellement, un concept manuel supplémentaire manque dans cette liste: "Bactériologie". Ce dernier apparaît bien, mais en 26ème position. Noter que le concept "Infection hospitalière" était absent de l'indexation manuelle, bien que pertinent.³⁷

³⁷ Le résultat complet de l'indexation de ce document est consultable à l'adresse : http://www.med.univ-rennes1.fr/cgi-bin/nomindex/info_idx.pl?prefixe=CF&code=R000822&max_key=50

Second document : "L'anorexie mentale" se trouvant à l'adresse suivante: <http://www-sante.ujf-grenoble.fr/sante/corpmcd/Corpus/corpus/question/pedi258.htm>

Validité	Concept détecté	Score TFIDF
trop fin	ANOREXIE	124,88
mal adapté	NUTRITION	61,68
mal adapté	COMPORTEMENT	60,15
mal adapté	REGIME ALIMENTAIRE	59,48
mal adapté	GENETIQUE	42,01
trop fin	FAMILLE	42,01
Oui	ADOLESCENCE	36,99
ok ++	COMPORTEMENT ALIMENTAIRE	36,53
mal adapté (trop fin)	FAIM	34,89
mal adapté	MANGER	34,89
trop fin	DEPRESSION	32,14
trop général	PSYCHIATRIE	31,26
mal adapté	PSYCHOTHERAPIE	31,11
mal adapté	SEXUALITE	29,23
mal adapté	SEXE	29,23
mal adapté	HOSPITALISATION	27,80
Oui	ANOREXIE MENTALE	25,09
mal adapté	PRONOSTIC	24,11
mal adapté	THERAPEUTIQUE	22,46
trop fin	VOMISSEMENT	22,09
trop fin	PARENTS	20,76
trop fin	PUBERTE	20,65
mal adapté	MALADIE	19,34
mal adapté	ETUDIANT	18,95
mal adapté	PSYCHOLOGIE	18,03
trop fin	TROUBLES PERSONNALITE	17,82
Oui	PEDIATRIE	17,20

Il manque le concept "Enfant" codé manuellement mais non proposé par NOMINDEX (et pour cause: NOMINDEX ne fait pas la différence, trop précise pour un tel moteur, entre le concept "Enfant" et le concept "Pédiatrie"). Ici encore, on observe la présence d'un concept pertinent, mais absent de l'indexation manuelle ("Comportement alimentaire").³⁸

Les résultats sont jugés très performants par l'équipe CISMef elle-même, et fera l'objet d'une publication prochaine dans une revue médicale.

³⁸ Résultat consultable à l'adresse : http://www.med.univ-rennes1.fr/cgi-bin/nomindex/info_idx.pl?prefixe=CF&code=R006810&max_key=50

.I.3.2. Comparaison avec l'indexation manuelle sur 7000 documents

Une autre étude, plus exhaustive mais moins détaillée a été alors mise en œuvre, il s'agit d'indexer plus de 7000 documents qui ont été codés manuellement par l'équipe CISMéF. Un fichier contenant le codage manuel (utilisant lui aussi le MeSH) de 7097 documents nous a été transmis, nous avons indexé chacun de ces documents avec NOMINDEX (toujours avec le MeSH comme thésaurus cible), parmi ces documents certains n'étaient pas téléchargeables³⁹, NOMINDEX en a néanmoins indexé 6773. Une fois l'indexation faite, nous avons comparé l'indexation manuelle avec l'indexation automatique.

Nous avons choisi de ne travailler que sur les textes ayant plus d'un certain nombre de concepts détectés par NOMINDEX. Il s'agit d'un petit biais, mais il est inutile d'essayer de proposer des mots-clés sur un document contenant moins de 10 concepts. Ces documents, contenant moins de 10 concepts, peuvent être des erreurs dues au téléchargement automatique, par exemple le document <http://www.concertation.org/init.asp>, quand il est lu automatiquement retourne un texte d'erreur ("Type mismatch: 'CDBI", ou " Invalid procedure call or argument: 'Mid'...")⁴⁰. D'autres documents ne contiennent quasiment pas de texte (la page d'accueil contenant une animation ou un menu présenté de manière non textuelle), NOMINDEX ne peut en extraire une quelconque information, exemple: <http://www.ciml.univ-mrs.fr> dont le seul texte extractible est "CIML - Centre d'Immunologie de Marseille Luminy".

Le filtrage précédent retient 4573 documents (soit 68% des 6773). Les résultats sont affichés ici sous la forme de précision et rappel, ces métriques sont présentées par "rang", car, ne connaissant pas l'indexation manuelle, le système ne peut décider par lui-même de n'afficher que les N premiers mots-clés. Les deux courbes sont donc présentées en fonction du rang. Un utilisateur pourra ensuite se limiter aux premiers rangs s'il privilégie la précision au rappel, ou, au contraire, vouloir consulter les 50 premiers concepts s'il privilégie le rappel.

³⁹ Pour plusieurs raisons: certains documents sont dans un format qui n'est pas compris par nomindex, d'autres ne sont plus accessibles avec l'URL indiquée, d'autres enfin contiennent des javascripts ou des filters qui empêchent une lecture automatique.

⁴⁰ Parfois NOMINDEX y découvre quand même des concepts, dans cet exemple les concepts "OR", ou "Cal osseux" (le mot "Call" est corrigé automatiquement en "Cal" qui génère le concept "Cal osseux"...).

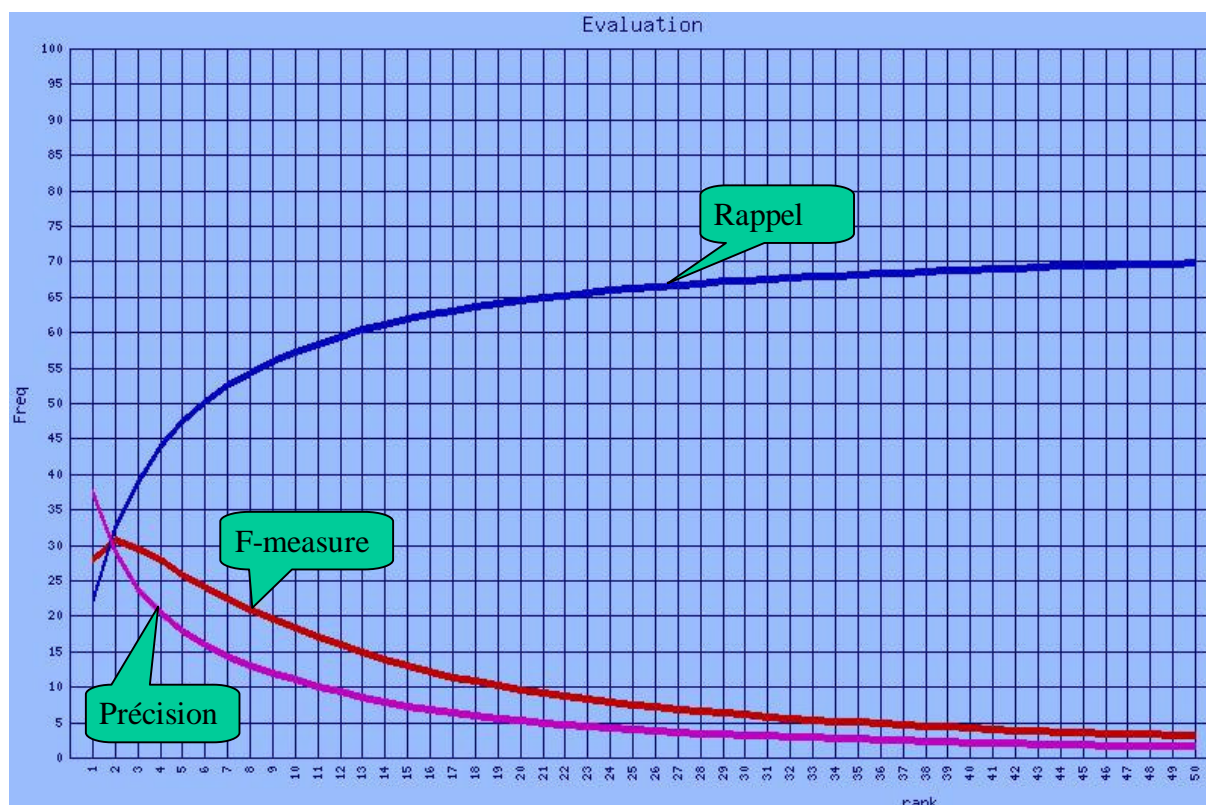


Figure 29 : Courbes de précision/rappel en fonction du rang (nombre de concepts proposés)

La table suivante (Table 5) donne les chiffres exacts obtenus selon le rang, la table complète est présentée en annexe 9 ("Comparaison des mots-clés de NOMINDEX et CISMef pour 7000 documents").

Rang	Précision	Rappel	F-measure
1	37	22	28
2	29	33	31
3	24	39	29
4	20	44	28
5	18	47	26
7	14	52	22
10	11	57	18
20	5	64	9
50	2	70	3

Table 5: Rappel, précision et F-measure de NOMINDEX pour 4573 documents en fonction du rang

La première appréciation de ces résultats est qu'ils n'ont rien de miraculeux. En effet la précision et le rappel sont incompatibles avec un codage uniquement automatique. Au rang 10

par exemple, seulement 57 % des concepts manuels sont détectés. Mais cette seconde étude ne quantifie plus le "silence CISMéF" (le taux de concepts pertinents détectés mais non codés manuellement), qui constitue, à notre avis, le plus gros apport de notre outil. Mais ce rappel (de 57 % au rang 10) est déjà un très bon chiffre si l'on tient compte du fait que l'outil n'a pas été construit dans ce but et n'a jamais reçu la moindre amélioration pour répondre à cet objectif.

Une autre méthode d'attribution automatique de mots-clés pourrait être développée en aval de notre produit, elle consisterait à appliquer une méthode par apprentissage (par exemple sur ce corpus de 7000 textes). Il faudrait, comme nous l'avons expérimenté pour l'attribution automatique de domaine nosologique, créer une base de données de probabilité d'appartenance à un mot-clé en fonction des concepts détectés dans le document par NOMINDEX. Ainsi, le programme pourrait attribuer un ensemble de mots-clés pour un nouveau document en fonction des concepts détectés par NOMINDEX.

A titre d'information, au Centre Commun de Recherche de la Communauté Européenne, ma nouvelle équipe d'accueil, nous avons utilisé une méthode d'attribution automatique de mots-clés pour des textes généraux [Steinberger, 2001], [Steinberger et al., 2002]. Cette méthode, entraînée sur 3318 documents, donne des résultats de l'ordre de 84% de précision au premier rang et 60% de rappel au rang 10. Pour avoir moi-même contribué à la construction de cet outil, je peux affirmer que cette méthode, si elle était appliquée en aval de NOMINDEX sur les documents du CISMéF, augmenterait très sensiblement les résultats indiqués sur la table 3. Les deux corpus d'entraînement étant comparables.

.I.3.3. Autres évaluations

Une étude a été faite sur l'utilisation de l'outil pour extraire des mots-clés de thèses de pharmacie. La méthode et les résultats sont détaillés dans [Mary et al., 2002]. Bien que notre lexique médical n'ait jamais été construit dans le but de traiter des documents de pharmacie, les résultats sont meilleurs que prévus, et rejoignent les remarques faites précédemment: beaucoup de bruit, mais NOMINDEX propose beaucoup de concepts pertinents non indexés manuellement. En résumé 10% de silence, 50% de bruit mais, surtout, près de 7 concepts nouveaux proposés par NOMINDEX pour chaque document.

En parallèle, une étude a porté sur 28 résumés de sortie de patients en cardiologie, combinant une reconnaissance vocale et l'indexation de NOMINDEX [Happe et al., 2002]. Les résultats sont à peu près identiques pour le silence: 12%. Le bruit est moins élevé, 25%, ce qui se comprend facilement quand on sait que le dictionnaire ADM est plus exhaustif en cardiologie qu'en pharmacologie.

.I.4. Traduction pour l'indexation

Cet outil étant expérimental, nous n'allons pas l'évaluer de manière détaillée, voici cependant un aperçu de son fonctionnement sur un exemple précis.

Nous avons choisi un document bilingue (disponible à l'adresse : http://europa.eu.int/comm/health/ph/programmes/cancer/index_en.htm). Ce document a été choisi car il contient exactement la même information conceptuelle dans les deux langues. Nous avons indexé le document français, et la traduction automatique du même document en version anglaise. Les résultats sont présentés sur la table suivante.

Phrase du texte original en français	Concepts extraits	Phrase du texte original en anglais	Phrase traduite	Concepts extraits
Cette année, la semaine européenne contre le cancer, qui se déroulera du 9 au 15 octobre, attirera l'attention du public sur le rôle que peuvent jouer de bonnes habitudes alimentaires dans la prévention du cancer.	PUBLICITE ,ATTENTION ,TUMEUR MALIGNE ,REGIME ALIMENTAIRE ,EUROPE ,ALIMENT ,NUTRITION ,PREVENTION ,ROLE ,Cancers	This year's Europe against Cancer Week, from 9 through 15 October, is focussing public attention on proper eating habits as a means to prevent cancer.	Ce an's Europe contre Cancer Semaine 9 travers 15 Octobre, est focussing publics attention sur convenable manger habitude a a moyen pour empeche cancer.	PUBLICITE ,APPETIT ,ATTENTION ,TUMEUR MALIGNE ,MANGER ,EUROPE ,EUROPE EST ,FAIM ,Cancers
En mangeant davantage de légumes et de fruits frais ainsi que des céréales a haute teneur en fibres, on diminuera le risque de contracter un grand nombre de cancers de différents types.	TUMEUR MALIGNE ,CEREALES ,FRUITS ,GIGANTISME ,RISQUE ,Cancers Synonyme abusif de grand	Eating more vegetables and fresh fruit, and cereals with high fibre content will reduce the risk of getting many different types of cancers.	Manger plus legumes et frais fruits, et cereales a haut fibre teneur va allger risque getting nombreux differentes types cancer.	APPETIT ,TUMEUR MALIGNE ,CEREALES ,MANGER ,FRUITS ,FAIM ,RISQUE ,Cancers mange
On contribuera aussi à prévenir cette maladie mortelle en évitant l'excès de poids, en augmentant son activité physique et en mangeant moins d'aliments gras.	MORT ,REGIME ALIMENTAIRE ,MATI ,MAL "Mangeant" est inconnu ,MATI ,ALIMENT ,GIGANTISME ,NUTRITION ,MALADE ,PHYSIQUE ,PREVENTION ,PRISE POIDS	Avoiding being overweight, increasing physical activity and eating less fatty foods will equally contribute to keeping the deadly disease at bay.	Evitant etant surpoids, croissante physique activite et manger moins grasses aliment va equally contribuer gardant deadly maladie baie.	APPETIT ,REGIME ALIMENTAIRE ,MATIERE GRASSE ALIMENTAIRE ,MALADIE ,MANGER ,MATIERES GRASSES ,ALIMENT ,CROISSANCE ,FAIM ,NUTRITION ,MALADE ,PHYSIQUE ,PRISE POIDS
Des éléments scientifiques solides et concordants montrent que des régimes a haute teneur en fruits et légumes réduisent le risque de contracter un grand nombre de cancers, en particulier les cancers de la bouche, du pharynx, de l'œsophage, des poumons et de l'estomac, et probablement d'autres types de cancers.	TUMEUR MALIGNE ,REGIME ALIMENTAIRE ,ELEMENTS ,OESOPHAGE ,ALIMENT ,FRUITS ,POUMON ,TUMEUR POUMON ,Cancer de l'estomac ,NUTRITION ,PHARYNX ,PNEUMOLOGIE ,RISQUE ,SCIENCE ,ESTOMAC ,CANCER DE L'OESOPHAGE ,TUMEUR MALIGNE DE LA BOUCHE ,TUMEUR MALIGNE DU LARYNX ,BOUCHE ,Autres tumeurs malignes ,Cancers ,Autres tumeurs malignes ,Autres cancers ,CANCER GASTRIQUE	There is strong and consistent scientific evidence showing that diets high in vegetables and fruits decrease the risk of many cancers, notably of cancers of mouth, pharynx, oesophagus, lung, and stomach, and probable or possible for other types of cancers.	La est forte et consistent scientific argument montrant celui regime alimentaire haut dans legumes et fruits diminue risque nombreux cancer, notablement cancer bouche, pharynx, oesophage, poumon, et probable ou eventuel pour autre types cancer.	TUMEUR MALIGNE ,REGIME ALIMENTAIRE ,OESOPHAGE ,ALIMENT ,FRUITS ,POUMON ,TUMEUR POUMON ,Cancer de l'estomac ,NUTRITION ,PHARYNX ,PNEUMOLOGIE ,RISQUE ,ESTOMAC ,CANCER DE L'OESOPHAGE ,TUMEUR MALIGNE DE LA BOUCHE ,TUMEUR MALIGNE DU LARYNX ,BOUCHE ,Autres tumeurs malignes ,Cancers ,Autres tumeurs malignes ,Autres cancers ,CANCER GASTRIQUE

Table 6 : Comparaison des concepts extraits depuis un texte traduit automatiquement

Nous pouvons observer, sur cet exemple, que les concepts trouvés ne sont pas si différents entre le texte original français et le texte traduit de l'anglais.

Cet exemple nous indique un moyen d'évaluer de manière intensive notre outil de traduction (couplé avec l'indexation). Il faudrait constituer un corpus de documents parallèles (chaque document français a un équivalent traduit en anglais). Nous pourrions ainsi indexer le document français par NOMINDEX et indexer la traduction automatique du document anglais, ensuite nous pourrions évaluer (avec les critères habituels rappel/précision) la performance de l'indexation de la traduction.

Ce programme de traduction automatique a été utilisé pour le transcodage d'information et l'enrichissement du meta-thésaurus UMLS par l'ADM [Le Duff et al., 2000]. Les résultats de traductions étaient le plus souvent corrects (malgré un bruit tout de même important).

Une autre évaluation de la performance de la traduction consiste à quantifier le nombre de mots du dictionnaire ADM qui sont inconnus de notre traducteur français ↔ Anglais. Les résultats sont très décevants car, sur 60000 mots que comporte notre lexique, seulement 18000 sont traduits (soit seulement un tiers). Ce résultat doit cependant être relativisé car, si l'on ne tient compte que des mots utilisés plus de 10 fois (dans le thésaurus ADM), seulement 4,5% sont inconnus. Cette dernière évaluation montre néanmoins que ce moteur de traduction reste insuffisant pour une utilisation intensive.

.II. Analyse factorielle des correspondances (application aux concepts)

Notre étude porte sur un échantillon aléatoire de 104 documents issus du CISMéF [Darmoni et al., 2000].

À titre de comparaison, nous avons effectué une AFC sur une indexation en texte intégral, et sur une indexation par concepts sur ces documents.

.II.1. AFC en texte intégral

Le nombre de documents est très limité, et, de plus, les co-occurrences de mots dans des textes de plusieurs pages perdent un peu de leur information sémantique. Il n'est pas très surprenant d'observer que l'AFC sur ce corpus de textes soit décevante. Nous avons étudié les dix premiers axes résultant de cette analyse, mais les mots pertinents extraits étaient pollués par des mots sans importance (Navigateur, Web, Serveur, lien ...). Comme nous nous y attendions, le nombre de documents était insuffisant pour extraire une information sémantique intéressante.

.II.2. AFC sur les concepts extraits

Nous avons construit une matrice comportant 104 documents et quelques 200 concepts avec leur fréquence d'apparition dans chaque document. Puis nous avons lancé une analyse des correspondances avec le logiciel QNOMIS II⁴¹. La Figure 30 présente le résultat sur les deux premiers axes (en italique apparaissent les documents, en écriture normale apparaissent les concepts). La première remarque est que les deux premiers axes ne nous fournissent pas beaucoup d'information sur notre corpus sinon qu'un document est particulièrement "atypique" : "M.S.T. en Afrique". Trois autres documents (en bas à gauche) semblent également atypiques mais sur un axe dont la signification n'est pas évidente (un axe partant des concepts "extension", "maladie", "récidive", "syndrome" et "diagnostic" jusqu'à "établissement", "descripteur", "enseignement et éducation"), cet axe semble permettre de différencier les documents plus orientés diagnostic des documents plus orientés pédagogie.

⁴¹ Plus exactement la chaîne de logiciels BI©LogInserm, ADDAD©, QNOMIS II©LogInserm

Cette hypothèse est vérifiée si l'on affiche plus de concepts ⁴², on trouve à gauche "récidive", "syndrome" "rémission", et, à droite : "publication", "rédaction", "science", "sciences de l'information".

Notons le concept atypique "virus" qui est, très justement, partagé entre le document "MST en Afrique" et les documents orientés diagnostic.

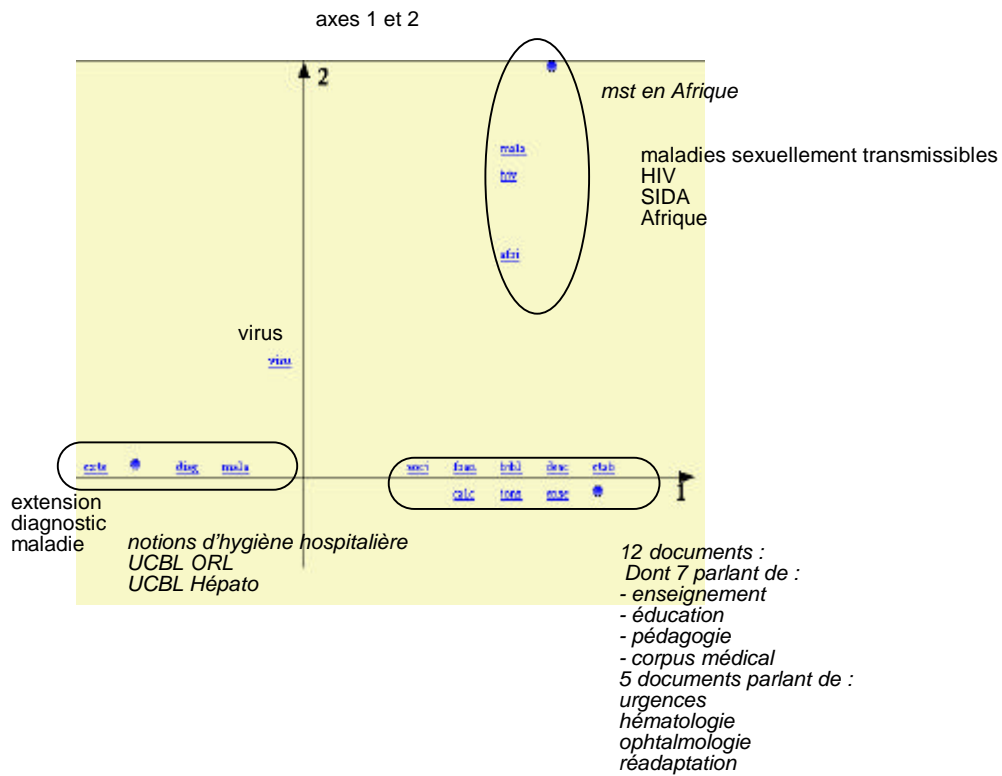


Figure 30 : AFC, visualisation des axes 1 et 2 sur 104 documents

⁴² des concepts ayant une contribution à la valeur propre plus faible que ceux précédemment cités

Nous avons poursuivi l'analyse des axes, mais nous nous sommes rendu compte qu'un document en Anglais avait été indexé par erreur. Notre programme a donc essayé de reconnaître des termes français sur un document anglophone, et celui-ci possédait des concepts très spécifiques "IF" ou "THE" (conjonctions et articles très courants en anglais, mots rares en français, "THE" étant interprété comme "Thé"). Ce document étant très "atypique" il est sur-représenté sur les axes 4 et 5.

Nous avons analysé les autres axes proposés jusqu'à trouver une répartition plus "nosologique" des documents, les axes 4 et 5 étant "pollués", nous avons fini par trouver ce plan de représentation sur les axes 3 et 6 :

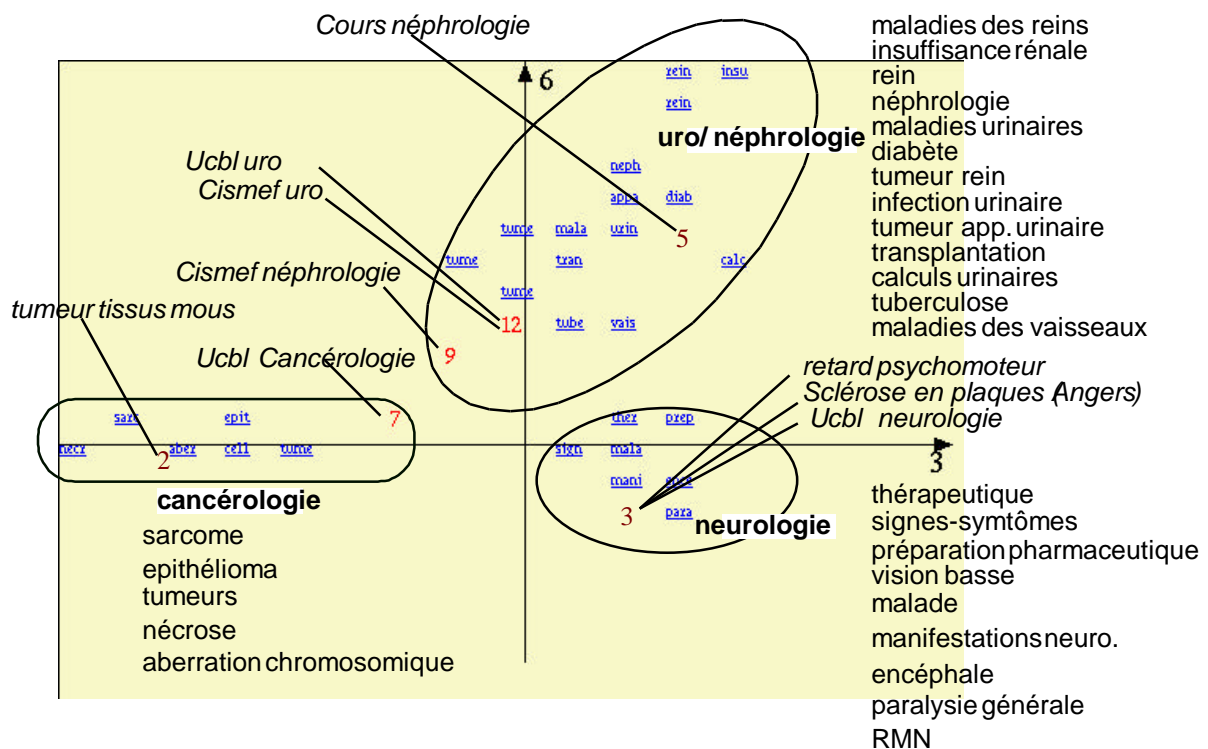


Figure 31 : AFC, visualisation des axes 3 et 6 sur 104 documents

Cette figure montre une nette séparation entre les documents appartenant aux trois domaines nosologiques : neurologie, uro-néphrologie, et cancérologie. Les documents appartenant aux trois domaines semblent bien classés (leur titre, en italiques sur le schéma, suffit pour s'en apercevoir), les concepts également.

Cette expérience nous montre toute la complexité de l'AFC, l'analyse d'un corpus de documents n'est, en aucun cas, immédiate et simple. Mais, on le voit bien sur cet exemple, les

différentes représentations sont rarement aberrantes (le tout est de "comprendre" à posteriori, pourquoi des concepts et documents sont proches et selon quel axe...).

.II.3. Perspectives

Les deux graphiques présentés ci-dessus nous donnent, d'une manière synthétique, une bonne idée du contenu de notre corpus.

Cette méthode, utilisée le plus souvent sur une matrice documents-mots a été ici appliquée sur une matrice documents-concepts. L'avantage de notre usage est de mettre en évidence très rapidement les proximités sémantiques des documents.

L'analyse factorielle des correspondances est une méthode statistique qui commence à donner de bons résultats quand le corpus de textes est suffisamment étendu pour que les co-occurrences de mots soient suffisamment significatives, cela fonctionne d'autant mieux que les unités textuelles sont courtes, comme en bibliographie (c. f. [Kerbaol et Bansard, 1999]). Par contre, lorsque les textes sont de longueur importante, les co-occurrences de mots perdent de leur signification (ce qui impose de segmenter les textes). Cependant les co-occurrences de concepts sont moins sensibles à la taille des textes.

Nous aurions pu poursuivre notre analyse afin de classifier automatiquement les documents en observant les axes sur lesquels les documents s'"expriment". Les deux graphiques montrent bien qu'il serait possible, par exemple, de classer automatiquement les documents "pédagogiques" et les documents "description clinique".(axe 1, Figure 30), ou de les classer par domaine nosologiques (axes 3 et 6, Figure 31). Mais, comme nous l'avons vu en introduction, la classification automatique de documents est un vaste domaine de recherche que nous n'avons pas souhaité aborder.

En conclusion, et par comparaison avec la même méthode sur un tableau lexical simple, cela tend à démontrer que notre indexation est sémantiquement pertinente.

Si les méthodes statistiques, ignorant les informations linguistiques, ne deviennent performantes que lorsque le corpus est suffisamment important, leurs performances peuvent être très sensiblement améliorées par une extraction préalable de concepts importants.

Discussion

Afin de placer notre outil dans son contexte, nous allons axer notre discussion en trois points : ce qui est, ou pourrait être, placé en amont de NOMINDEX, discussion sur l'outil lui-même, et ce qui pourrait être fait, ou amélioré, en aval de l'outil.

.I. En amont

Notre outil est relativement figé de par l'utilisation d'un lexique spécialisé (issu essentiellement du dictionnaire ADM). Mais il serait possible de fournir à NOMINDEX le résultat d'outils d'apprentissage de vocabulaire à partir de corpus. Le corpus des documents médicaux disponibles sur Internet est maintenant suffisamment étendu pour pouvoir constituer un lexique des termes médicaux conséquent (voire pour constituer de nouveaux thésaurus). Cet outil aurait alors suffisamment de connaissances pour pouvoir indexer de manière très acceptable tout document médical. L'autre alternative serait de fournir comme lexique d'entrée une source existante (Wordnet [Miller, 1995] par exemple), à plus forte raison si l'on veut tester l'outil avec un lexique en anglais (ou dans toute autre langue).

Il faut bien souligner que l'outil ne permettra pas d'indexer des documents avec une classification telle que la CIM10 ou la CdAM, les entrées de ces classifications sont trop générales (il s'agit de classes, non de concepts). D'autre part, les termes utilisés sont souvent "des expressions d'un métalangage plutôt que les expressions que l'on pourrait trouver dans des textes naturels" [Zweigenbaum, 1999]. NOMINDEX ne pourrait pas reconnaître un terme aussi complexe que "Autres incisions sur la vulve et le périnée, agrandissement de l'orifice vaginal sans autres indications, à l'exclusion de l'ablation de corps étranger sans incision". Il s'agit bien là de "classifications", et non de thésaurus (lire ci-dessous une méthode proposée pour indexer des documents avec une classification).

L'outil fonctionne sur un thésaurus, ce qui laisse la possibilité de l'utiliser pour un domaine non médical. La difficulté est, dans ce cas, de constituer un lexique qui ne soit plus orienté médecine.

De même, il n'y a pas de contre-indication pour utiliser NOMINDEX dans une autre langue que le français, en constituant un lexique dans la même langue. Notons qu'il existe des outils pour la détection de mots composés, par contre il n'en existe pas pour la constitution de

mots associés, qui sont une originalité de NOMINDEX (mais ceci constitue un axe de recherche intéressant⁴³).

Le système d'indexation est aussi figé par le thésaurus, un concept n'appartenant pas au thésaurus ne sera jamais indexé dans nos documents. Il est important de tenir à jour la version du thésaurus (mettre à jour le MeSH par exemple), sans quoi les néologismes ne seront jamais indexés. Un autre axe de recherche serait d'observer les questions des utilisateurs et repérer, par une méthode statistique classique, les concepts fréquemment utilisés mais qui sont inconnus de NOMINDEX, ceci afin de cibler les mises à jour en fonction des demandes des utilisateurs.

⁴³ Un programme statistique pourrait détecter, dans un thésaurus, les synonymies "mot simple" – "mot complexe", si le mot complexe n'est pas un mot composé il pourrait être proposé comme mot associé candidat...

.II. L'outil NOMINDEX

Les mots associés (réelle originalité de notre processus) mériteraient d'être segmentés en différentes sous-catégories (ceux dont l'inversion des mots est possible, nombre et nature des mots intermédiaires permis, dérivations possibles de chaque composants...). Ici encore, l'étiquetage des mots associés pourrait être le résultat d'un apprentissage sur corpus...

La polysémie des mots et des termes entraînent souvent des erreurs du système d'indexation (par exemple le concept MeSH "Lutte" est souvent détecté dans les documents, or, dans ce thésaurus, ce terme désigne le sport). Polysémie d'autant plus importante que notre lexique n'est pas accentué (une interrogation sur "côte" entraîne beaucoup de bruits, car l'entrée du lexique est identique pour "cote", "côte" et "côté"). La ré-accentuation du lexique permettrait donc de résoudre une partie de ces problèmes. Un lemmatiseur, ou un étiqueteur syntaxique, permettrait de différencier certaines ambiguïtés selon leur classe grammaticale ("la côte" serait différenciée automatiquement de "le côté").

Il serait aussi possible d'utiliser des notions de contextes pour désambigüiser des mots. Voir, par exemple, l'expérience de [Sébillot et Pichon, 1999]. On pourrait, par exemple, classer le mot "lutte" dans la famille de "lutteur" quand le concept "sport" a été trouvé (notons que ceci ne constitue pas une solution universelle, car un document pourra parler de "Lutte contre le dopage sportif").

Le système NOMINDEX est ainsi fait que toute modification du thésaurus, doit être suivie d'une ré-indexation complète des documents du corpus. A plus forte raison, la modification du lexique nécessite le recodage du thésaurus et la ré-indexation de tous les documents. Il faudrait pouvoir disposer d'un outil permettant de mettre à jour quelques entrées du thésaurus ou du lexique, et qui assurerait un minimum de cohérence avec les documents déjà indexés.

.III. En aval

L'outil d'extraction de mots de références pourrait être utilisé pour comparer/fusionner des thésaurus médicaux, comme cela a été expérimenté dans [Le Duff et al., 2000]. En effet, deux termes de thésaurus peuvent être orthographiés différemment, cet outil pourrait alors les mettre automatiquement en correspondance, l'outil de traduction automatique pourrait même être utilisé pour mettre en correspondance des thésaurus multilingues. L'outil NOMINDEX pourrait être utilisé pour proposer des relations d'inclusion de concepts (méthode que nous avons utilisée pour l'attribution de domaine), c'est à dire que l'on peut connaître les concepts dont tous les mots sont inclus dans le terme. Par exemple, dans la nomenclature médicale CIM10, nous avons le concept I82 "Autres embolies et thromboses veineuses", NOMINDEX pourrait y reconnaître les concepts MeSH suivants : C0013922 EMBOLIE, C0031542 PHLEBITE⁴⁴, C0040046 THROMBOPHLEBITE, C0040053 THROMBOSE, C0042449 VEINES, C0085307 EMBOLIE ET THROMBOSE.

Lors de l'évaluation du S.R.I., nous avons observé que l'outil se comportait légèrement différemment selon que l'on utilise l'ADM ou le MeSH comme thésaurus cible. Il serait très intéressant de fusionner ces deux thésaurus (et d'y ajouter d'autres thésaurus français comme SNOMED). Cela permettrait de disposer d'une liste de concepts plus conséquente, l'outil fonctionnerait d'autant mieux sur ce nouveau thésaurus.

Les expérimentations de la méthode d'Analyse Factorielle des Correspondances pourraient être les prémices d'un outil de cartographie et de classification des documents plus ambitieux.

Si le but est d'indexer des textes par des classifications médicales (objectif particulièrement intéressant pour le PMSI), l'outil NOMINDEX pourrait se révéler très utile pour étiqueter les concepts contenus dans des textes et y appliquer une méthode d'apprentissage statistique sur un corpus indexé manuellement. Nous pourrions utiliser la même méthode que pour l'attribution de domaines pour attribuer automatiquement des codes de nomenclature (par exemple CIM10) à des textes en langage naturel. En résumé, à partir d'un corpus de résumés de sortie de patient indexés manuellement avec la nomenclature CIM10, en extraire les concepts MeSH, et calculer les co-occurrences de concepts avec chaque code CIM10. Ainsi nous pourrions tenter d'attribuer automatiquement des codes CIM10 à de nouveaux résumés de sortie de patients.

Nous pourrions utiliser l'étiquetage de NOMINDEX sur les documents pour créer automatiquement un réseaux de liens hypertextes entre les différents chapitres d'un document (permettre d'accéder aux autres documents contenant les mêmes concepts que ceux contenus dans le chapitre), comme cela avait été expérimenté avec les documents de radiologie EDICERF [Duvauferrier et al., 1997]. L'indexation par le MeSH permettrait aussi de retrouver automatiquement des références bibliographiques sur le même sujet (via pubmed qui est interrogeable par le MeSH⁴⁵).

Nous avons expérimenté un usage de l'information contenue dans la base de connaissances ADM: la possibilité d'utiliser les concepts détectés dans un texte médical pour interroger le système d'évocation de diagnostics de l'ADM. Le fonctionnement est relativement simple, nous prenons les concepts les plus importants du document (scores TFIDF les plus élevés), et nous lançons une évocation de diagnostics sur ces concepts. Le résultat sera une liste de pathologies contenant tout ou partie de ces concepts. Cet outil étant purement expérimental, il est décrit en annexe 10 ("Evocation de diagnostics pour un document").

Enfin, il est important de souligner que si notre système permet de reconnaître la plupart des concepts exprimés dans un document, il ne permet pas de reconnaître les relations sémantiques existant entre ces concepts. Il pourrait cependant être utilisé comme une aide à la construction d'une modélisation de l'information plus ambitieuse, comme le formalisme des graphes conceptuels [Sowa, 1984]. Il serait en effet plus simple de construire une représentation sémantique adéquate d'un texte si l'on en connaît déjà quelques concepts.

⁴⁴ Phlébite est quasi-synonyme de "Thrombose veineuse" dans notre lexique

⁴⁵ Notamment à l'aide du module Perl WWW::Search::PubMed

Conclusion

Les résultats sont jugés très satisfaisants, mais nous ne disposons pas réellement d'éléments de comparaison, sauf pour la recherche d'information où notre moteur de recherche en texte intégral avait environ 60% de précision, nous en avons maintenant 93% en moyenne (voir le chapitre "Évaluation du système de recherche d'information", p. 113). L'autre comparaison est celle effectuée par l'équipe CISMeF sur l'attribution de mots-clés, là encore les résultats sont très encourageants (voir le chapitre "Extraction de mots-clés", p. 115). Nous sommes conscients des limites de telles évaluations, mais elles ont permis de mettre l'accent sur les qualités et les défauts du système.

La plupart des défauts constatés peuvent être corrigés en intervenant sur le lexique, ou, parfois, en intervenant sur certains termes ou concepts du thésaurus. Les défauts inhérents à la construction des thésaurus sont une limite du système, mais gageons que le meta-thésaurus UMLS corrigera peu à peu ces défauts...

D'autres défauts ne pourront être corrigés qu'avec le recours à la linguistique, notamment le recours à un étiquetage lexical semble nécessaire, mais cela nécessite un travail considérable sur le dictionnaire ADM (afin d'ajouter l'information grammaticale), le recours à des outils statistiques ou à d'autres lexiques existants permettrait de faciliter la tâche.

Le dictionnaire ADM, sur lequel repose notre système d'indexation, est uniquement en français, ce qui exclut toute utilisation sur des thésaurus anglophones non traduits. Cependant, afin de pouvoir néanmoins indexer des textes multilingues, nous avons expérimenté une traduction préalable des termes en français, traduction faite à partir des traductions de termes de l'UMLS, l'outil se révèle pratique, mais engendre beaucoup de bruits. L'autre perspective, plus ambitieuse, serait de construire un lexique entièrement anglais (qui pourrait être alimenté par les données lexicales de l'UMLS).

Ce système serait applicable à d'autres domaines que la médecine, à la condition de construire un lexique approprié. Ce qui n'est pas tâche facile, mais des outils existent pour constituer un lexique de termes à partir, soit d'un thésaurus (comme [Zweigenbaum et Grabar, 2000]), soit du corpus de textes (comme [Jacquemin et Tzoukermann, 1999]). La simplicité de la construction du lexique peut alors devenir un atout, car il ne sera pas nécessaire de définir les informations grammaticales, syntaxiques ou sémantiques de chaque entrée du lexique, et permet donc l'utilisation d'outils purement statistiques (ou mixtes), ce qui facilitera aussi la création de lexiques non francophones.

Les limites et défauts du système sont, dans le cadre de l'utilisation qui en a été faite, compensés par sa simplicité de mise en oeuvre, et ses performances quantitatives.

Il est prévu d'utiliser cet outil de manière intensive pour indexer chaque nouveau document inséré dans le CISMéF (40 nouveaux documents par semaine). Deux buts sont visés : aide à l'indexation (l'outil permettra de proposer des mots-clés à un documentaliste pour l'aider à indexer le document sur le catalogue CISMéF), et la recherche d'information (le CISMéF disposerait ainsi de notre moteur de recherche en texte libre dans tous les documents indexés). Le fait que cet outil soit utilisé pour aider à indexer tous les documents du plus grand répertoire des sites francophones est, selon nous, le meilleur résultat dans l'état actuel de l'outil. Même si nous sommes bien conscients que beaucoup de travail reste à accomplir.

Bibliographie

(108 références)

- [**Abeillé et Blache, 2000**] Abeillé A. et Blache P., 2000, *Analyse syntaxique*, Ingénierie des langues, Hermes
- [**Arene et al., 1998**] Arene I., Ahmed W., Fox M., Barr C.-E., Fisher K., 1998 , *Evaluation of quick medical reference (QMR) as a teaching tool.*, MD Comput 1998, 15(5), p. 323-6
- [**Aronson et al., 2000**] Aronson A.R., Bodenreider O., Chang H.F., Humphrey S.M., Mork J.G., Nelson S.J., Rindflesch T.C., Wilbur W.J., 2000, *The NLM Indexing Initiative*, Proc AMIA Symp 2000;:17-21
- [**Baeza et Ribeiro, 1999**] Baeza-Yates A., Ribeiro-Neto B., 1999, *Modern information retrieval*, ACM Press books, Addison-Wesley.
- [**Bellot, 2000**] Bellot P., 2000, *Méthodes de classification et de segmentation locales non supervisées pour la recherche documentaire*, Thèse de doctorat, Université d'Avignon.
- [**Benzécri et al., 1973**] Benzécri J.-P. et al., 1973, *La taxinomie*, Vol. I ; L'analyse des correspondances, Vol II, Dunod, Paris
- [**Berners-Lee et al., 2001**] Berners-Lee Tim, Hendler James, Lassila Ora, 2001, *The Semantic Web*, Scientific American
- [**Bloksma et al., 1996**] Bloksma, L., Díez-Orzas, Vossen, P., 1996, *The User-Requirements and Functional Specification of the EuroWordNet-project*, EuroWordNet deliverable D001, LE2-4003, University of Amsterdam, Amsterdam
- [**Bodenreider et al., 1998**] Bodenreider O., Burgun A., Botti G. et al. 1998 *Evaluation of the Unified Medical Language System as a Medical Knowledge Source* JAMIA 1998; 1 (5) : p. 76 - 87.
- [**Bodenreider, 2000**] Bodenreider Olivier, 2000, *Using UMLS semantics for classification purposes*, Proc. AMIA Symposium 2000; p. 86-90.
- [**Bodenreider et Zweigenbaum, 2000**] Bodenreider Olivier, Zweigenbaum Pierre, 2000, *Identifying proper names in parallel medical terminologies*, Medical Infobahn for Europe - Proceedings of MIE2000 and GMDS2000, Amsterdam, IOS Press, p. 443-447

- [Bourigault, 1996]** Bourigault D., 1996, *LEXTER, a Natural Language Processing tool for terminology extraction*, Proceedings of the 7th EURALEX International Congress, Goteborg
- [Bourigault et Jacquemin, 2000]** Bourigault D. et Jacquemin C., 2000, *Construction de ressources terminologiques*, Ingénierie des langues, Hermes.
- [Brandt et Nadkarni, 1999]** Brandt Cynthia, Nadkarni Praksh, 2001., *Web-based UMLS concept retrieval by automatic text scanning: a comparison of two methods*, Comput Methods Programs Biomed, 64(1): p. 37-43.
- [Buckley, 1985]** Buckley Chris, 1985, *Implementation of the SMART Information Retrieval System*. Technical Report TR85-686. Cornell University, Ithaca, NY.
- [Burgun et al., 1992]** Burgun A., Le Beux P., Bremond M., Lenoir P., 1992, *Translation from one medical nomenclature into another: a categorization of the problems*. Medinfo 1992: p. 1458-61
- [Burgun et al., 1996]** Burgun A., Botti G., Bodenreider O., et al., 1996, *Methodology for using the UMLS as a background knowledge for the description of surgical procedures*. Int J Biomed Comput;43(3): p. 189-202
- [Choueka et Zampoli, 1992]** Choueka Y., et Zampoli. A., 1992, *Responsa: An Operational Full-Text Retrieval System with Linguistic Components for Large Corpora: Computational Lexicology and Lexicography: a Volume in Honor of B. Quemada*. Pisa: Giardini Press.
- [Church, 1995]** Church Kenneth Ward. 1995. *One Term or Two?* Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval 1995: 310–318.
- [CIM 1977]** *Classification Internationale des Maladies - 9^{ème} Révision*, 1977, OMS, Genève.
- [Cimino et al., 1993]** Cimino J.J., Johnson S., Peng P., Aguirre A., 1993, *From ICD9-CM to MeSH using the UMLS: a how-to guide*. Proc Annu Symp Comput Appl Med Care,; p. 730-4
- [Cleret et al, 2001]** Cleret M., Le Duff F., Fresnel A., Le Beux P., 2001, *Diamed: a probabilistic diagnostic aid system on the web*. Medinfo, 10(Pt 1):429-33

- [Cote, 1995] Cote R, 1995, ed. Systematized Nomenclature of Human and Veterinary Medicine (SNOMED International), version 3.1. College of American Pathologists, Northfield (IL), American Veterinary Medical Association, Schaumburg (IL).
- [CPT, 1996] Physician's current procedural terminology. (4th ed.) Chicago (IL): American Medical Association, 1996.
- [Daille, 1994] Daille B., *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*, Thèse en informatique fondamentale, université de Paris VII, 1994
- [Darmoni et al., 2000] Darmoni S., Leroy J.-P., Thirion B., Baudic F., Douyere M., Piot J. 2000 *CISMeF : a structured Health resource guide*, Methods of Information in Medicine, Janvier 2000; 39(1) p. 30-35.
- [Darmoni et al., 2001] Darmoni S.-J., Thirion B., Leroy J.-P., Douyère M., Piot J., 2001, *The use of Dublin Core metadata in a structured health resource guide on the Internet*. - Bulletin of the Medical Library Association, Juillet 2001; 89(3) p. 297-301
- [David et Plante, 1990] David S. et Plante P., *De la nécessité de l'approche morpho-syntaxique dans l'analyse de textes*, Intelligence Artificielle et Sciences Cognitives au Québec, 3 (3), p. 140-154, 1990.
- [Deerwester et al., 1990] Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R., 1990, *Indexing by latent semantic analysis*. Journal of the American Society for Information Science, 41, p. 391-407
- [Descartes et Bunce, 2000] Descartes A. et Bunce T., 2000, *Programming the Perl DBI, Database programming with Perl*, ed. O'Reilly & Associates
- [Desclès et Minel, 2000] Desclès J.-P., Minel J.-L., 2000, *Résumé automatique et filtrage sémantique de textes*, Ingénierie des langues, Hermes
- [Dunning, 1994] Dunning Ted, 1994, *Statistical identification of language*, Technical report CRL MCCS-94-273, Computing Research Lab, New Mexico State University.
- [Duvauferrier et al., 1995] Duvauferrier R., Rambeau M., André M., Denier P., Le Beux P., Coussement A., Caillé J. M., Robache P., Morcet N., 1995, *Iconothèques et ouvrages multimédia sur serveur et cd-rom en imagerie médicale (l'expérience française)*. J. Radiol 1995; 76 (12), p. 1079-85.

- [Duvauferrier et al., 1997]** Duvauferrier R., Le Beux P., Pouliquen B., Séka L.-P., Morcet N., Rolland Y., 1997, *Intérêt de l'Indexation médicale automatique d'une Iconothèque et d'une Bibliothèque Radiologiques numériques*. Journal de Radiologie.
- [El-Beze, 1995]** El-Beze M., Merialdo B., Rozeron B. et Derouault A.-M., 1994, *Accentuation automatique de textes par des méthodes probabilistes*, Technique et science informatiques, 13 (6), p. 797-815,
- [Enguehard et Pantera, 1995]** Enguehard C. et Pantera L., 1995, *Automatic natural acquisition of a terminology*, Journal of Quantitative Linguistics, 2 (1), p. 27-32
- [Felber, 1987]** Felber H.-L., 1987, *Manuel de terminologie*, Unesco, Paris,
- [Fieschi et Joubert, 1994]** Fieschi M., Joubert M, 1994, *Intégration de bases d'informations dans des systèmes d'informations médicaux : apport de l'intelligence artificielle*. Actes des 5ièmes journées francophone d'informatique médicale. Genève, Suisse
- [Fluhr, 2000]** Fluhr C., 2000, *Indexation et recherche d'information textuelle*, Ingénierie des langues, Hermes
- [Frakes, 1992]** Frakes W.-B., 1992. *Stemming Algorithms*. In Information Retrieval, edited by W. B. Frakes, and R. Baeza-Yates. Englewood Cliffs, NJ: Prentice Hall: p. 131–160.
- [Fresnel et al., 1996]** Fresnel A., Pouliquen B., Riou C., Delamarre D., Le Beux P., 1996, *Computer Assisted Medical Diagnosis - European Congress of the Internet in Medicine, Hospital*. European Congress of the Internet in Medicine, Brighton 17 oct. 1996, p. 14-
- [Fresnel et al., 1998]** Fresnel A., Jarno P., Burgun A., Delamarre D., Denier P., Cleret M., Courtin C., Seka LP, Pouliquen B., Cleran L., Riou C., Leduff F., Lesaux H., Duvauferrier R., Le Beux P., 1998, *A first evaluation of a pedagogical network for medical students at the University Hospital of Rennes*. Med Inform (Lond). 1998, 23(3), p. 253-64.

- [**Giguet, 1995**] Giguet.E., 1995, *Categorization according to Language: A step toward combining Linguistic Knowledge and Statistic Learning*. In Proceedings of the International Workshop on Parsing Technologies (IWPT'95), Prague - Karlovy Vary, Czech Republic, p. 20-24.
- [**Grefenstette, 1995**] Grefenstette G., 1995, *Comparing Two Language Identification Schemes*. In the Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data (JADT'95), Rome, Italy, Rome, Italy
- [**Gross, 1986**] Gross M., 1986, *Lexicon-grammr, The representation of compound words*, proceedings of COLING-86, Bonn, p. 1-6
- [**Gundavaram, 1996**] Shishir Gundavaram, 1996, *Programmer des CGI sur World Wide Web*, ed. O'Reilly & Associates
- [**Guttman, 1941**] Guttman L., 1941, *The Quantification of a Class of Attributes*. In "The prediction of personal adjustment, P.Horst ed., SSCR (New York)
- [**Habert et Jacquemin, 1993**] Habert B., Jacquemin C., 1993, *Noms composés, termes, dénominations complexes : problématiques linguistiques et traitements automatiques*, Traitement Automatique des Langues (TAL), n° 2, p. 5-41.
- [**Habert et al., 1997**] Habert B., Nazarenko A., Salem A., 1997. *Les linguistiques de corpus*. Paris: Armand Colin.
- [**Happe et al., 2002**] Happe André, Pouliquen Bruno, Burgun Anita, Cuggia Marc, Le Beux Pierre, 2002, *Combining voice recognition and automatic indexing of medical reports*, MIE 2002, Budapest, 25-29Août 2002, (in press)
- [**Harbeck et Ohler, 1999**] Harbeck S. et Ohler U., 1999, "Multigrams for Language Identification", Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech 99), Budapest, Hungary
- [**Hayashi, 1956**] Hayashi C., 1956, *Theory and Examples of Quantification (II)* Proc. of the Institute of Stat. Math., 4 (2) p. 19-30
- [**Ihaka et Gentleman, 1996**] Ihaka Ross et Gentleman Robert, 1996, *R: A Language for Data Analysis and Graphics*, Journal of Computational and Graphical Statistics, vol. 5, n. 3, p. 299-314

- [ISO-latin, 1987]** ISO 8859-1, 1987, *Information processing — 8-bit single-byte coded graphic character sets — Part 1: Latin alphabet No. 1.*
- [Jacquemin et Tzoukermann, 1999]** Jacquemin C. et Tzoukermann E., 1999, *NLP for Term Variant Extraction : Synergy between Morphology Lexicon and Syntax*, Natural Language Information retrieval, T. Strzalkowski (ed.), Kluwer.
- [Jacquemin et Zweigenbaum, 2000]** Jacquemin C. et Zweigenbaum P., 2000, *Traitement automatique des langues pour l'accès au contenu des documents.* In J. Le Maître, J. Charlet and C. Garbay, editors, *Le document Multimédia en Sciences du Traitement de l'Information*, CÉPADUÈS-Éditions, Toulouse, p. 71-110.
- [Kan et Klavans, 1998]** Kan M.Y., Klavans J., 1998, *Linear segmentation and segment signifiance*, draft, soumission ID#542,A5-L1,RP, Columbia University, USA,.
- [Kerbaol et al., 1997]** Kerbaol M., Bansard J.-Y., Favier L., 1997, *Sélection de la bibliographie de "maladies rares", une approche expérimentale*, 4^{ème} séminaire d'Obernai DIDDOC INSERM
- [Kerbaol et Bansard, 1999]** Kerbaol M., Bansard J.-Y., 1999, *Pratique de l'analyse des données textuelles en bibliographie*; Ecole MODULAD SFdS, INRIA, Bases de données et statistiques, Editeur sous presse
- [Lancaster, 1998]** Lancaster F.-W., 1998, *Indexing and abstracting in theory and practice.* Library Association Publishing. London.
- [Laporte, 2000]** Laporte Eric, 2000, *Mots et niveau lexical*, Ingénierie des langues, Hermes, p. 25-49
- [Le Beux et al., 1995]** Le Beux Pierre, Burgun Anita, Denier Patrick, Delamarre Denis, 1995, *Du codage à l'information médicale, des signes aux concepts et relations sémantiques.* Actes Cristal's. "Une Nomenclature Unique : Comment Faire et Pour Quoi Faire", Juin 1995, St Malo

- [**Le Beux et al., 2000**] Le Beux P., Le Duff F., Fresnel A., Berland Y., Beuscart R., Burgun A., Brunetaud JM, Chatellier G., Darmoni S., Duvauferrier R., Fieschi M., Gillois P., Guille F., Kohler F., Pagonis D., Pouliquen B., Soula G., Weber J., 2000, *The French Virtual Medical University.*, Stud Health Technol Inform. 2000; 77, p. 554-62.
- [**Le Duff et al., 2000**] Le Duff F., Burgun A., Cleret M., Pouliquen B., Barac'h V., Le Beux P. 2000, *Knowledge acquisition to qualify Unified Medical Language System interconceptual relationships*, Proc AMIA Symp, p. 482-6
- [**Lebart et al., 1995**] Lebart L., Morineau A., Piron M., 1995, *Statistique exploratoire multidimensionnelle*, Dunod, Paris
- [**Lenoir et al., 1981**] Lenoir P., Michel J.-R., Frangeul C., Chales G., 1981, *Réalisation, développement et maintenance de la base de données A.D.M.* Médecine informatique, vol. 6, N° 1, p. 51-56
- [**Lindberg et al., 1993**] Lindberg DAB, Humphreys BL., Mc Cray AT., et al., 1993, *The Unified Medical Language System* Meth Inform Med; 4 (32) : p. 281-91
- [**London, 1998**] London S., 1998, *DXplain: a Web-based diagnostic decision support system for medical students*, Med Ref Serv Q, 17(2), p.17-28
- [**Mani et al., 1999**] Mani I., House D., Klein G., Hirschmann L., Firmin T. et Sunfheim B., *The TIPSTER SUMMAC Text Summarization Evaluation*, Proceedings of the 9th Conference of the European Chapter of the Association of Computational linguistics (EACL'99), Bergen, ACL, 1999, p. 77-85
- [**Mary et al., 2002**] Mary Vincent, Pouliquen Bruno, Le Duff Franck, Darmoni Stefan J., Segui Alain, Le Beux Pierre, 2002, *Automatic conceptual indexing of French Pharmaceutical theses*, MIE 2002, Budapest, 25-29Août 2002, (in press)
- [**Mc Cray et al., 1993**] Mc Cray A.T., Arondson A.R., Browne A.C., Rinflesh T.C., Razi A., Srinivasan S., 1993, *U.M.L.S. knowledge for biomedical language processing*. Bulletin of medical library association. 81(2) : p. 184-193

- [Meadeb et al., 1986] Meadeb J., Chales G., Lenoir P., 1986, *Apprentissage du diagnostic médical par simulation sur ordinateur (AEDM)*, MED.INFORM., vol 11, n°2, p. 167-175
- [MeSH, 1986] NATIONAL LIBRARY OF MEDICINE, 1986, *Medical Subject Headings* Bethesda, Maryland
- [Miller, 1995] Miller G.A., 1995, *WordNet: A Lexical Database for English*.
Communications of the ACM 11
- [Mizzaro, 1997] Mizzaro S., 1997, *Relevance : the whole history*, Journal of the American Society for Information Science, vol. 48, n° 9, p. 810-832.
- [Murphy et al., 1996] Murphy G.-C., Friedman C.-P., Elstein A.-S., Wolf F.-M., Miller T., Miller J.-G., 1996, *The influence of a decision support system on the differential diagnosis of medical practitioners at three levels of training*. Proc AMIA Annu Fall Symp 1996, p. 219-23
- [Piotrowski, 2000] Piotrowski Michael, 2000, *NLP-Supported Full-Text Retrieval*, Master's Thesis, Erlangen-Nürnberg Institut für Germanistik Abteilung für computerlinguistik, April 28, 2000
- [Pouliquen et al., 1995] Pouliquen B., Riou C., Denier P., Fresnel A., Delamarre D., Le Beux P., 1995, *Using World Wide Web Multimedia in Medicine*, Proc. of MEDINFO'95, IMIA, Eds Greenes, Peterson, Protti, p. 1519-1523
- [Riou et al., 1990] Riou C., Le Beux P., Rammal M., Cador F., Frangeul C., Lenoir P., 1990, *A computer assisted instruction diagnosis aid using a large medical knowledge base*, IMIA International Conference on Medical Informatics and Medical Education. Prague, Czechoslovakia
- [Riou et al., 1994] Riou C., Le Beux P., Lenoir P., 1994, *L'AEDM - EAO de simulation de cas cliniques : fonctionnalités actuelles et tendances évolutives*. Actes Colloque International Informatique appliquée à la santé au travail et dans l'environnement. Amphoux M. Ed Commission des Communautés Européenes EUR 14772, p. 114-119

- [Riou et al., 1999]** Riou C., Pouliquen B., Le Beux P., 1999, *A computer assisted Drug prescription System: the Model and its implementation in the ATM knowledge base.* Methods of Information in Medicine
- [Robertson, 1994]** Robertson S. E., Walker S., Hancock-Beaulieu M., Gatford M., 1994. *Okapi in TREC-3*, Text Retrieval Conference TREC-3, U.S. National Institute of Standards and Technology, Gaithersburg, USA. NIST Special Publication 500-225, p. 109-126.
- [Rouault, 1997]** Rouault Philippe, 1997, *Conception et réalisation d'un système d'indexation automatique multilingue de textes spécialisés à partir d'un dictionnaire médical.* Mémoire de diplôme d'ingénieur CNAM en génie informatique. Conservatoire des Arts et Métiers, Rennes.
- [Ruch et al., 1999]** Ruch P., Wagner J., Bouillon P., Baud R.H., Rassinoux A.M., Scherrer J.R., 1999, *MEDTAG: tag-like semantics for medical document indexing*, Proc AMIA Symp., p. 137-41.
- [Salton, 1971]** Salton G., 1971, *The SMART retrieval system. Experiment in automatic document processing.* Prentice Hall. Englewood Cliffs. New Jersey.
- [Salton, 1983]** Salton G., 1983, *Introduction to Modern Information Retrieval*, McGraw-Hill.
- [Salton, 1986]** Salton G., 1986, *Another look at automatic text retrieval systems*, Commun. ACM, 29 (7), p. 648-656
- [Salton et Buckley, 1988]** Salton G., Buckley C., 1988, *Term weighting approaches in automatic text retrieval*, Information Processing and Management, vol. 24, n° 5, 1988, p. 513-523
- [Savoy et Picard, 2000]** Savoy J., Picard J., 2000, *Report on the TREC-8 Experiment: Searching on the Web and in Distributed Collections.* Proceedings TREC'8, NIST publication #500-246, Gaithersburg (MD), p. 229-240
- [Savoy et Rasolofo, 2002]** Savoy J., Rasolofo Y., 2002, *Report on the TREC-10 Experiment: Distributed Collections and Entrypage Searching.* Proceedings TREC'10, NIST, Gaithersburg (MD)

- [Sébillot et Pichon, 1999]** Sébillot Pascale, Pichon Ronan, juillet 1999, *Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience*, TALN'99 (Traitement automatique des langues naturelles), Cargèse, France
- [Sébillot et Pichon, 1997]** Sébillot Pascale, Pichon Ronan, 1997, *Acquisition automatique d'informations lexicales à partir de corpus : un bilan*, INRIA , No. RR-3321
- [Séka et al., 1995]** Séka L.-P., Fresnel A., Delamarre D., Riou C., Burgun A., Pouliquen B., Duvauferrier R., Le Beux P., 1997, *Computer assisted medical diagnosis using the Web*. Int J Med Inf ; 47(1-2), p. 51-56
- [Sheridan et Ballerini, 1996]** Sheridan Páraic, et Ballerini Jean Paul., 1996, *Experiments in Multi-lingual Information Retrieval using the SPIDER System*. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR 96, p.58–65.
- [Simard, 1996]** Simard M., *Automatic Restoration of Accents in French Text*, Centre d'innovation en technologies de l'information, Laval, Canada 1996 : 9 pages
- [Smadja, 1993]** Smadja F., 1993, *Retrieving collocations from text: Xtract*, Computational Linguistics, 19 (1), p. 143-177
- [Sowa, 1984]** Sowa J.-F., 1984, *Conceptual structures. Information processing in mind and machine*. Readings, Massachusetts: Addison-Wesley
- [Steinberger, 2001]** Steinberger Ralf, 2001, *Cross-lingual Keyword Assignment*. Proceedings of the XVII Conference of the Spanish Society for Natural Language Processing (SEPLN'2001). Jaén, Spain
- [Steinberger et al., 2002]** Steinberger Ralf, Pouliquen Bruno, Hagman Johan, 2002, *Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus Eurovoc*. In: A. Gelbukh (ed.) *Intelligent Text Processing and Computational Linguistics. Lecture Notes in Computer Science for the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'2002)*. Mexico-City, Mexico, 17-23 Février 2002. Springer, Heidelberg.
- [Unicode, 1997]** The Unicode Consortium, 1997, *The Unicode Standard, Version 2.0*. Addison-Wesley.

- [Van Rijsbergen, 1979]** Van Rijsbergen, C.J., 1979, *Information retrieval* (second edition), Butterworths, London
- [Volot et al., 1993]** Volot F., Zweigenbaum P., Bachimont B., Ben Said M., Bouaud J., Fieshi M, Boisvieux J.-F., 1993, *Structuration and acquisition of medical knowledge : using U.M.L.S. in the conceptual graph formalism*. Proceedings of the 17th symposium computer of medical care (SCAMC). Ed. Mc Graw-Hill. Washington, District of Columbia.
- [Wall et al., 1996]** Wall L., Christiansen T., Schwartz R., 1996, *Programmation Perl*, ed. O'Reilly & Associates
- [Wilkinson, 1994]** Wilkinson R., 1994, *Effective retrieval of structured documents*, ACM/SIGIR'94 Conference on Research and Development in Information Retrieval, Dublin, Ireland, p. 311 à 317
- [Wong, 1997]** Wong Clinton, 1997, *Web Client Programming with Perl*, Automating Tasks on the Web, ed. O'Reilly & Associates
- [Zweigenbaum, 1999]** Zweigenbaum P., 1999, *Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances*.ISIS, (2-3), p. 27-47.
- [Zweigenbaum et al., 1994]** Zweigenbaum P., Consortium Menelas. 1994, *MENELAS: An Access System for Medical Records Using Natural Language*. Computer Methods and Programs in Biomedicine, 45, p. 117-120.
- [Zweigenbaum et Grabar, 2000]** Zweigenbaum P. et Grabar N., 2000, *Expériences d'acquisition automatique de connaissances morphologiques par amorçage à partir d'un thésaurus*, Actes du 12 congrès Reconnaissance des Formes et Intelligence Artificielle, Paris, p. II-101-II-110

.I. Index des illustrations

Figure 1 : Processus d'indexation.....	6
Figure 2 : Fonctionnement général d'un Système de Recherche d'Information	13
Figure 3 : AFC: exemple de projection sur un plan	18
Figure 4 : AFC: Graphe résultant de la projection.....	18
Figure 5 : Evaluation, classification des documents	23
Figure 6 : Evaluation, bruit et silence.....	23
Figure 7 : Evaluation, schéma idéal	25
Figure 8 : Exemple d'interrogation de l'ADM	32
Figure 9 : Fonctionnalités essentielles de l'ADM	35
Figure 10 : Exemple de famille de mots dans le dictionnaire ADM.....	36
Figure 11 : Exemple de contenu du thésaurus ADM (concepts, termes, hiérarchie et relation sémantique)	41
Figure 12 : Un réseau sémantique basé sur UMLS.....	44
Figure 13 : Schéma en trois couches.....	53
Figure 14 : Fonctionnement du système d'indexation NOMINDEX.....	54
Figure 15 : Exemple de fonctionnement du système NOMINDEX	55
Figure 16 : Schéma de base de données des mots.....	56
Figure 17 : Un exemple d'entrée du dictionnaire ADM	59
Figure 18 : Fichier XML d'export du dictionnaire ADM, et visualisation.....	60
Figure 19 : Principe d'indexation d'un thésaurus	65
Figure 20 : Exemple d'indexation de thésaurus	66
Figure 21 : Schéma de la base de données des concepts et de leur indexation en mots	67
Figure 22 : Exemple d'étiquetage de document HTML	74
Figure 23 : Schéma de la base de données d'index des documents.....	76
Figure 24 : Processus d'indexation d'un document	77
Figure 25 : Exemple d'indexation de document (contenu de la base de données).....	77
Figure 26 : Exemple de représentation de l'angle de deux vecteurs dans un espace à trois dimensions	88
Figure 27: Classement du document "Cours de néphrologie" en domaines nosologiques	96
Figure 28 : Exemple d'interrogation multilingue	107
Figure 29 : Courbes de précision/rappel en fonction du rang (nombre de concepts proposés)	120
Figure 30 : AFC, visualisation des axes 1 et 2 sur 104 documents.....	127
Figure 31 : AFC, visualisation des axes 3 et 6 sur 104 documents.....	128

.II. Index des tables

Table 1 : Tableau lexical d'un exemple simple	17
Table 2 : Concepts similaires à "Déformation du pied"	98
Table 3 : Concepts similaires à "Sciatique"	99
Table 4 : Exemple de traductions extraites du MeSH	103
Table 5: Rappel, précision et F-mesure de NOMINDEX pour 4573 documents en fonction du rang	120
Table 6 : Comparaison des concepts extraits depuis un texte traduit automatiquement	124

.III. Index des algorithmes

Algorithme 1 : Reconnaissance de mots dans une phrase	61
Algorithme 2 : Fonction de recherche de mots multiples.....	62
Algorithme 3 : Reconnaissance de concepts dans une phrase	71
Algorithme 4 : Fonction de correspondance de NOMINDEX	72

.IV. Index des équations

$$TFIDF_{c,d} = TF_{c,d} \cdot IDF_{c,d} \quad (\text{Equation 1}) \dots\dots\dots 83$$

$$TFIDF_{c,d} = TF_{c,d} \cdot \left(\log_2 \frac{N}{DF_c} + 1 \right) \quad (\text{Equation 2}) \dots\dots\dots 83$$

$$COSINE(d,r) = \frac{\sum_{c \in d \cap r} TFIDF_{c,d} \cdot TFIDF_{c,r}}{\sqrt{\left(\sum_{c \in d} TFIDF_{c,d}^2 \right) \cdot \left(\sum_{c \in r} TFIDF_{c,r}^2 \right)}} \quad (\text{Equation 3}) \dots\dots\dots 84$$

$$COSINE(d_1, d_2) = \frac{\sum_{c \in d_1 \cap d_2} TFIDF_{c,d_1} \cdot TFIDF_{c,d_2}}{\sqrt{\left(\sum_{c \in d_1} TFIDF_{c,d_1}^2 \right) \cdot \left(\sum_{c \in d_2} TFIDF_{c,d_2}^2 \right)}} \quad (\text{Equation 4}) \dots\dots\dots 87$$

$$COSINE(d, p) = \frac{\sum_{c \in d \cap p} TFIDF_{c,d} \cdot TFIDF_{c,p}}{\sqrt{\left(\sum_{c \in d} TFIDF_{c,d}^2 \right) \cdot \left(\sum_{c \in p} TFIDF_{c,p}^2 \right)}} \quad (\text{Equation 5 : Phrase/document}) \dots\dots\dots 89$$

$$COSINE(c_1, c_2) = \frac{\sum_{d \in d_{c_1} \cap d_{c_2}} TFIDF_{c_1,d} \cdot TFIDF_{c_2,d}}{\sqrt{\left(\sum_{d \in d_{c_1}} TFIDF_{c_1,d}^2 \right) \cdot \left(\sum_{d \in d_{c_2}} TFIDF_{c_2,d}^2 \right)}} \quad (\text{Equation 6 : Concept/concept})$$

Glossaire

Terme	Définition
Acronyme	<i>Sigle</i> , prononcé sans être épilé (ex: CIM10 pour "Classification Internationale des Maladies").
AFC	Cf. <i>Analyse Factorielle des Correspondances</i> .
Analyse Factorielle des Correspondances (AFC)	Méthode statistique d'analyse des correspondances de mots et de documents permettant, notamment, de représenter graphiquement les correspondances entre les mots et les documents (Ici entre les concepts et les documents).
Antonyme	Relation sémantique exprimant qu'un concept est le contraire d'un autre ("plein" est antonyme de "vide").
Bigramme	Séquence de deux caractères consécutifs (cf. <i>N</i> -gramme).
Candidat terme	Mot, ou groupe de mot, susceptible d'être retenu dans une terminologie
CGI	Common Gateway Interface, méthode permettant à un serveur Web d'exécuter des programmes afin de générer des pages HTML dynamiques
Concept	Une idée générale, intelligible et transmissible par différents usagers. Dans notre travail, un concept sera désigné comme une entrée d'un thésaurus (de moins de cinq mots). Un concept est représenté par un ou plusieurs <i>termes</i> .
Corpus (de textes/documents)	Ensemble de textes (documents) relatifs à un domaine donné. Ici, les documents qui seront indexés.
Cosine	<i>Mesure de similarité</i> exprimant le cosinus de l'angle formé par deux vecteurs (deux documents) dans notre <i>espace vectoriel</i> .
Bruit	Réponses erronées parmi les résultats fournis par un programme d'analyse textuelle (le plus souvent un moteur de recherche).
Descripteur	Désigne un mot, ou groupe de mots servant à représenter le contenu d'un texte (document, ou partie de document). Dans notre contexte, l'information extraite de document servant d'index (mots, lemmes, n-grammes, termes, concepts...), aussi appelé <i>unité élémentaire</i> .

Document	Dans notre contexte, nous désignons par document tout texte destiné à être indexé.
Etiqueteur	Processus qui ajoute dans un texte des annotations (morphologiques, syntaxiques, grammaticales, sémantiques...) (eng. "tagger")
Famille de mots	Ensemble de synonymes, quasi-synonymes, flexions et dérivation d'un même mot.
Flexion	Caractéristique d'un mot contenant un indice de genre ou nombre. Ici nous désignons par "flexion" un mot au pluriel ou au féminin
Forme canonique	Représentant unique pour toutes les formes fléchies d'un mot (singulier pour un nom, masculin singulier pour un adjectif, infinitif pour un verbe ...). Voir aussi <i>lemme</i> .
Fréquence (d'un <i>descripteur</i> dans un texte)	Nombre de fois qu'apparaît un mot donné dans un texte
Grammaire	Règles d'écriture d'une langue.
Homonymie	Ressemblance entre deux mots ayant des sens différents (dans le langage écrit, on parle d' <i>homographe</i>)
Homographe	Forme commune (même orthographe) de plusieurs mots (exemple, "couvent").
Holonyme	Relation sémantique exprimant qu'un concept est un tout d'un autre ("main" est holonyme de "doigt").
HTML	HyperText Markup Language, langage de description des documents destinés à être publiés sur des serveurs web.
Hyperonyme	Se dit d'un terme qui est générique d'un autre (son <i>hyponyme</i>). On le désigne aussi par terme générique. "Fruit" est hyperonyme de "pomme".
Hyponyme	Se dit d'un terme particulier d'un autre terme plus général. On le désigne aussi par "terme spécifique". "Pomme" est hyponyme de "fruit".

IDF	Inverse Document Frequency, facteur fonction du nombre de documents dans lesquels apparaît le concept. IDF est faible pour un concept très utilisé, fort pour un concept peu utilisé.
Indexation automatique	Procédure permettant l'extraction de <i>descripteurs</i> d'un texte. Ces descripteurs sont ensuite stockés de manière à être utilisés à de multiples fins: recherche, classification...
Index	Résultat d'une indexation (qui peut être automatique ou manuelle), désigne souvent la base de données stockant cet index.
Langage naturel	Langage parlé naturellement par les humains. Par extension désigne aussi le langage écrit.
Latent semantic indexing (LSI)	Méthode statistique consistant à indexer des documents par l'information sémantique implicite contenue dans les co-occurrences de descripteurs dans les documents.
Lemmatisation	Processus linguistique (morphologique) qui, à l'aide d'un lexique des flexions, permet d'enlever les flexions de mots, pour ne garder que la forme grammaticale de base (la racine du mot).
Lemmatiseur	Outil de <i>lemmatisation</i> .
Lemme	Résultat d'une lemmatisation, racine de mots, forme du mot sans la partie flexionnelle. Toutes les flexions d'un mot ont le même lemme.
LSI	<i>Cf. Latent Semantic Indexing</i>
Méronymie	Relation sémantique exprimant qu'un concept est une "partie" d'un autre concept (le concept "myocarde" est méronyme de "Cœur", "Cœur" est holonyme de myocarde).
Mesure de similarité	Cf. score de similarité.
Modèle booléen	Modèle de recherche d'information répondant à des questions exprimées sous la forme de combinaison d'opérateurs booléens ("et", "ou", "sauf").
Modèle vectoriel	Modèle représentant un document comme un vecteur de mots (dans notre contexte, un document comme un vecteur de concepts).
Morphologie	Etude de la formation des mots et de leurs variations de forme.
Mot	Le plus souvent, en <i>TALN</i> , un mot est défini comme une suite de

	caractères séparés par des espaces (on parle alors d'unité lexicale ou de mot simple). Ce qui s'éloigne de la définition linguistique (unité autonome d'une langue), dans notre contexte, un mot peut être composé de plusieurs unités lexicales.
Mot associé	Mot formé par plusieurs unités lexicales qui ne sont pas obligatoirement contiguës, et dont la signification est déduite de celle de ses composants ("mal de tête", "douleur rachis dorsal"), on utilise un mot associé pour exprimer sa synonymie avec un mot simple
Mot-clé	Un mot (ou groupe de mots) désignant une notion essentielle d'un texte, permettant d'en faciliter la compréhension (voire la classification).
Mot composé	Mot formé par plusieurs unités lexicales, et dont la signification est souvent différente de celle de ses composants ("angine de poitrine", "pomme de terre")
Mot de référence	Mot utilisé pour représenter une <i>famille de mots</i> . Le mot le plus communément utilisé pour représenter ses flexions, synonymes, ou dérivations (eng. "synnet").
Mot multiple	Mot composé de plusieurs unités lexicales, dans notre contexte, il peut être, soit un <i>mot composé</i> , soit un <i>mot associé</i> . On parle aussi d'unité lexicale complexe
Mot nul	Mot n'ayant aucune information sémantique, le plus souvent il s'agit de <i>mot-outil</i> . Ainsi dénommé car on les enlève de la phrase avant la recherche.
Mot simple	Unité lexicale, un mot ne contenant pas d'espaces, par opposition à <i>mot multiple</i>
Mot-outil, mot-stop	Mot dont la fonction est purement grammaticale (eng. "Stop word")
Moteur de recherche	Cf. S.R.I. (eng. "search engine")

<i>N</i> -gramme	Séquence de <i>N</i> caractères consécutifs (pouvant être utilisés comme <i>descripteurs</i> de document).
Polysémie	Propriété d'un terme qui présente plusieurs sens
Pondération (de poids de concept)	Facteur permettant d'augmenter le poids d'un concept dans un document en fonction de son usage dans les autres documents (par exemple: <i>TFIDF</i>).
Pragmatique	Etude du contexte d'énonciation.
Racinisation	Processus qui consiste à enlever des mots les derniers caractères (considérés comme décrivant la partie flexionnelle des mots), par exemple, enlever le "s" des mots au pluriel. Aussi appelé Stemmatization. (eng. " <i>stemming</i> ")
Recherche documentaire, recherche d'information	Cf. S.R.I.
Réseau sémantique	Désigne une représentation des relations entre les concepts, ces relations sont exprimées en fonction de critères sémantiques particuliers. Les relations les plus couramment utilisées sont les <i>taxonomies</i> (relation de type générique vers spécifique), ou les <i>méronymies</i> (relation de type "partie_de"), mais on peut exprimer d'autres relations comme "s'exprime_par", "est_un_symptome_de"...)
R.I.	Recherche d'Information.
S.R.I.	Système de Recherche d'Information. Système permettant la recherche de documents (préalablement indexés dans la plupart des cas) répondant à une requête de l'utilisateur
Scores de similarité	Métrique de mesure de similarité entre deux documents, entre un document et un descripteur, entre un document et une requête de l'utilisateur, entre deux descripteurs.
Segmentation	Découpage d'un texte en différents segments, en général, la phrase ou le paragraphe.
Sémantique	Etude du sens.

SGBDR	Système de Gestion de Base de Données Relationnel (par exemple, Oracle, Sybase, Mysql, PostgreSQL, ...). (eng. "RDBMS")
SGML	Standard Generalised Markup Language, norme ISO 8879 de 1986. Langage de description de documents (sous forme de balises).
Sigle	Abréviation, en général composée des initiales de chacun des composants (ex: A.V.C. pour "Accident Vasculaire Cérébral").
Silence	Réponses correctes mais néanmoins absentes des résultats fournis par un programme d'analyse textuelle
Similarité	Indique à quel point deux documents sont semblables, dans notre contexte exprime à quel point les concepts utilisés dans les deux documents sont les mêmes.
Stemming, stemmatisation	Cf <i>racinisation</i> .
Stop-word	Cf. <i>mot-outil</i>
Synonyme	Qui a un sens identique ou très voisin. "Céphalée" est un synonyme de "Mal de tête".
Syntaxe	Etude de l'agencement des mots et à leurs relations structurelles dans un texte.
Tableau lexical	Matrice représentant, pour chaque mot (ici chaque concept), sa fréquence d'apparition dans chaque document. Aussi appelé tableau de contingence
Tagger	Cf. étiqueteur
TALN	Cf. <i>Traitement Automatique du Langage Naturel</i> .
Taxonomie	Relation sémantique exprimant qu'un concept est <i>hyperonyme</i> d'un autre (relation terme générique ↔ terme spécifique)
Terme	Représentant linguistique d'un concept dans un domaine de connaissance, résultat d'un processus d'analyse terminologique.
Terme de référence, Terme vedette	Terme le plus communément utilisé pour désigner un concept
Terminologie	Recensement des concepts et termes d'une langue, ou d'un domaine particulier.

Texte intégral	Système indexant les textes selon les mots qu'ils contiennent. Sans aucun traitement préalable. (eng. "Full-text")
TF	Term Frequency, fréquence d'apparition d'un mot (ici un concept) dans un document.
TFIDF	Term Frequency Inverse Document Frequency. Fréquence d'apparition d'un concept (<i>TF</i>) pondérée par sa fréquence d'utilisation dans le corpus (<i>IDF</i>)
Thésaurus	Répertoire de termes normalisés souvent associé à un <i>réseau sémantique</i> . Sert à décrire et organiser des informations dans un domaine donné.
Traitement automatique du langage naturel (TALN)	Désigne l'ensemble des programmes qui traitent le langage naturel, afin, par exemple de rechercher de l'information, de résumer, classifier, traduire, corriger ou représenter des documents...
Trigramme	Séquence de trois caractères consécutifs (cf. <i>N</i> -grammes).
Unicode	Norme de codage des caractères sur 16 bits, permettant de représenter tout caractère (quel que soit l'alphabet)
Unité d'indexation	Partie de texte choisie pour être indexée. Selon les modèles, pourra être le document entier, le paragraphe, la phrase (résultats d'une <i>segmentation</i>)
Unité élémentaire	Cf. <i>descripteur</i> .
Unité linguistique	Cf. <i>unité d'indexation</i>
Unité lexicale	Une suite de caractères séparés par des espaces (ou ponctuation), nous parlons ici de <i>mots simples</i> .
URL	"Uniform Resource Locator" : Identifiant d'une page Web
XML	eXtensible Markup Language : langage de description de documents conçu à partir de la norme <i>SGML</i> (plus simple, et maintenant plus répandu que <i>SGML</i>).

Annexes

[Note provisoire destinée aux lecteurs de PDF]

Voir second document, temporairement à l'adresse :

<http://www.med.univ-rennes1.fr/~poulique/th/annexe.pdf>