Construction automatique de hiérarchies sémantiques à partir du Trésor de la Langue Française informatisé (TLFi) : application à l'indexation et la recherche d'images

Inga Gheorghita

Construction automatique de hiérarchies sémantiques à partir du *Trésor de la langue française informatisé* (*TLFi*) : application à l'indexation et à la recherche d'images

THÈSE

présentée et soutenue publiquement par

Inga GHEORGHITA

le 17 février 2014

en vue de l'obtention du titre de

Docteur de l'Université de Lorraine

Spécialité : Linguistique Informatique

préparée au laboratoire ATILF (UMR 7118)

dans le cadre de l'École Doctorale Langages, Temps, Sociétés

*** Directeur de thèse : Jean-Marie PIERREL ***

Jury :

Rapporteurs :

Béatrice DAILLE, Professeur à l'Université de Nantes
Brigitte GRAU, Professeur à l'ENSIIE

Examinateurs :

Cyril MARCH, Ingénieur, Directeur général de Xilopix
Jean-Marie PIERREL, Professeur à l'Université de Lorraine
Alain POLGUÈRE, Professeur à l'Université de Lorraine
Yannick TOUSSAINT, Chargé de Recherche HDR à l'INRIA

ANALYSE ET TRAITEMENT INFORMATIQUE DE LA LANGUE FRANÇAISE

Remerciements

Cette thèse CIFRE est le fruit d'une collaboration entre le laboratoire ATILF et l'entreprise Xilopix. Cette organisation m'a permis de profiter des compétences et de l'expérience des personnes du monde académique et professionnel que je tiens à remercier ici.

Je remercie Jean-Marie Pierrel, mon directeur de thèse. Je tiens à lui exprimer toute ma reconnaissance pour la qualité d'encadrement fourni tout au long de la thèse avec autant de disponibilité, pour la confiance et la liberté qu'il m'a accordée, ainsi que pour son soutien permanent sur tous les plans. Je le remercie aussi pour la patience et l'attention avec laquelle il a relu et corrigé mon manuscrit. Pour tout cela, je tiens à lui témoigner toute ma gratitude.

Je remercie avec la même sincérité Cyril March, mon encadrant à Xilopix, de m'avoir fait partager son expérience dans le monde professionnel, son soutien ainsi que sa bonne humeur permanente. Je lui suis reconnaissante de m'avoir donné la chance d'intégrer la société Xilopix après la finalisation de la thèse, et ainsi de réaliser ma première expérience dans le monde de l'entreprise.

Je tiens à saluer Éric Mathieu pour m'avoir accueillie au sein de son entreprise et m'avoir accordé toute sa confiance pour l'évolution du projet Xilopix.

Merci à William Del Mancino, initiateur de cette thèse CIFRE.

Je tiens à remercier les membres du jury pour l'honneur qu'ils m'ont fait en acceptant d'évaluer mes travaux et mon mémoire de thèse.

J'ai aussi une pensée à Jean Véronis qui a accepté d'être rapporteur de cette thèse et qui nous a quitté bien trop tôt.

J'exprime mes remerciements à l'ANRT pour le soutien financier apporté à cette thèse CIFRE.

En ce qui concerne le laboratoire,
Je remercie Sandrine et Étienne pour l'aide qu'ils m'ont apportée au début de la thèse, en mettant à ma disponibilité les ressources lexicales de l'ATILF.
Je remercie également Olivier et William pour le dépannage lorsque mon ordinateur tombait en panne, jamais au bon moment.
Comment ne pas remercier mes collègues et anciens collègues de l'ATILF, Aurore, Coralie, Cornelia, Cécile, Crépin, Jingjing, Émilienne, Gina, Lolita, Lucia, Magali, Mihaela, Ulrike, avec qui cette thèse a été menée dans une ambiance très conviviale. Je remercie aussi Évelyne, Pascale, Jean-Marc pour les discussions que nous avons eues à propos du *TLFi*.

Merci à Laurence et Aurore pour l'aide qu'elles m'ont apportée lors de la préparation des cours.
Enfin je remercie tous les membres du laboratoire ATILF pour leur accueil chaleureux, pour les échanges qu'on a pu avoir à la cafétéria ou à la cuisine.

En ce qui concerne l'entreprise,
Je remercie l'équipe informatique pour l'aide qu'ils m'ont apportée lors du développement de mes algorithmes.
Je remercie l'équipe éditoriale d'avoir participé à l'évaluation de mes résultats. Merci à Hanène pour sa disponibilité et ses conseils.
Je remercie aussi tous les collègues xilopixiens pour leurs sourires, petits « Bonjour, ça va !! » et les bons moments partagés lors des repas de midi.

Je remercie aussi mes relecteurs Maria, Lolita, Aurore, Stéphanie, Hope pour leurs remarques et leurs conseils.

Je remercie aussi toutes les personnes, que j'ai rencontrées pendant ces trois ans, lors de conférences et d'écoles d'été ESSLLI, avec qui j'ai pu partager une très bonne expérience.

À mes amis Ghizlane, Igor, Lilia, Mirabela, Darina, Petru, Victor, Nicoleta, Sana, Sergiu, Youma, je témoigne une grande reconnaissance pour tout les bons moments passés ensemble.

Je remercie la famille Kozak de m'avoir accueillie chez eux avec tant de bienveillance et de chaleur.

Enfin, mes remerciements les plus forts vont à mes parents, ma famille et mes proches pour leur amour inconditionnel, leur réconfort et leurs encouragements. Merci d'avoir cru en moi et de m'avoir soutenue dans mes ambitions. Cette thèse est dédiée à vous!

Je remercie toutes les personnes qui ont directement ou indirectement contribué à la réalisation de ce travail.

Merci à tous ! Mulțumesc la toți !

Спасибо всем ! Дякую всім! Thank you to all!

Toutes les bonnes choses ont une fin. C'est donc avec une certaine émotion que je prends conscience que l'épisode le plus passionnant et enrichissant de ma vie d'étudiante prend fin. Ces souvenirs inoubliables resteront gravés dans ma mémoire pour toujours !

« Veni, vidi, vici »

Sommaire

Introduction générale 13
Contexte de la thèse 14
Problématique et orientations de recherche 15
Organisation du manuscrit 17

Chapitre 1. Le *Trésor de la langue française informatisé* (*TLFi*) et sa structure. 21

1.1. Du *TLF* vers le *TLFi* 21
1.2. Macrostructure et microstructure du *TLFi* 24
1.3. Modèle informatique des données du *TLFi* 29
1.4. Les divers accès possibles aux données du *TLFi* en s'appuyant sur sa structure 31
1.5. Les divers types de définitions lexicographiques 34
1.5.1. Définition hyperonymique (logique) 35
1.5.2. Définition synonymique 36
1.5.3. Définition dérivative 36
1.5.4. Définition méréonymique 37
1.5.5. Définition approximative 38
1.6. Repérage des modes définitoires dans le *TLFi* 38
1.7. Le *TLFi* et les projets connexes 40
1.7.1. Projet SEMEME 41
1.7.2. Projet Definiens et ses résultats 43
1.7.3. Utilisation du *TLFi* pour l'enrichissement d'une ontologie 47
1.7.4. Projet RELIEF 48
1.8. Conclusion 48

Chapitre 2. Fondements de la recherche d'information et spécificités de la recherche d'images 53

2.1. Architecture d'un SRI 54
2.1.1. Phase d'indexation 54
2.1.2. Phase de recherche et aperçu de principaux modèles classiques de RI 58
2.2. Les systèmes d'indexation et de recherche d'images 60
2.2.1. Indexation textuelle manuelle d'images 61
2.2.2. Indexation textuelle automatique d'images 63
2.2.3. Recherche textuelle d'images 65
2.2.4. Indexation et recherche d'images par le contenu 68
2.3. Évaluation des SRI 71
2.3.1. Mesures de la qualité de l'indexation 71
2.3.2. Mesures de la qualité de la recherche 73

2.3.3. Problèmes spécifiques à l'évaluation des systèmes de recherche d'images 75
2.4. Conclusion ... 77

Chapitre 3. Construction de ressources lexicales et leur utilisation pour l'indexation et la recherche d'images ... 81

3.1. Lexiques informatiques ... 81
3.1.1. WordNet ... 81
3.1.2. ConceptNet ... 86
3.1.3. DBpedia ... 87
3.2. Thésaurus ... 88
3.2.1. Thésaurus — outil d'indexation et de recherche ... 89
3.2.2. Les relations dans le thésaurus ... 92
3.2.3. Les modes de représentation des thésaurus ... 93
3.3. Ontologies ... 96
3.3.1. Structures et types d'ontologies ... 97
3.3.2. Utilisation d'ontologies dans le domaine d'indexation et de recherche d'images ... 100
3.4. Construction automatique de hiérarchies sémantiques ... 102
3.4.1. Les études de construction de hiérarchies sémantiques à partir de textes 102
3.4.2. Les études de construction de hiérarchies sémantique à partir de définitions lexicographiques ... 105
3.4.3. Questions posées par l'extraction de connaissances à partir de dictionnaires. ... 108
3.5. Conclusion ... 113

Chapitre 4. Extraction automatique d'informations pertinentes des définitions du *TLFi* ... 117

4.1. Définition du corpus de travail et son analyse ... 117
4.1.1. Analyse des définitions ... 119
4.1.2. Analyse des domaines ... 122
4.1.3. Analyse des locutions ... 124
4.2. Propositions de pondérations des noms dans le *TLFi* ... 124
4.2.1. Analyse des critères de pondération des noms dans une définition ... 125
4.2.1.1. Analyse de la fréquence du nom dans les définitions du *TLFi* ... 127
4.2.1.2. Analyse du nombre de définitions ... 128
4.2.1.3. Analyse de la position du nom dans la définition ... 128
4.2.1.4. Analyse de l'appartenance des noms aux expressions métalinguistiques ... 129
4.2.2. Normalisation du corpus de travail ... 130
4.2.2.1. Normalisation des domaines ... 130
4.2.2.2. Traitement des expressions métalinguistiques ... 132

4.2.3. Pondérations retenues pour déterminer l'importance des noms dans les définitions 133
4.2.3.1. Pondération locale 133
4.2.3.2. Pondération globale 134
4.2.3.3. Pondération par position 135
4.3. Évaluation de chaque facteur de pondération par rapport aux CC et CP du projet Definiens 136
4.4. Analyse de l'influence de chaque facteur de pondération sur le calcul de poids des noms 137
4.5. Conclusion 141

Chapitre 5. Construction automatique de hiérarchies sémantiques à partir du *TLFi* 145

5.1. Hiérarchisation des noms 146
5.1.1. Règles d'inclusion 146
5.1.2. Règles d'association 148
5.1.3. Règles de hiérarchisation 148
5.2. Méthodologie de construction automatique de hiérarchies sémantiques 149
5.2.1. Vue globale de l'approche 149
5.2.2. Description détaillée 149
5.3. Évaluation manuelle des relations hyperonymiques 159
5.3.1. Présentation du corpus d'évaluation 160
5.3.2. Analyse des résultats d'évaluation 161
5.3.2.1. Analyse des résultats positifs de l'évaluation manuelle 162
5.3.2.2. Analyse des résultats négatifs de l'évaluation manuelle 163
5.4. Comparaison des hiérarchies sémantiques avec le thésaurus Xilopix 166
5.4.1. Description du corpus d'évaluation 166
5.4.2. Analyse des résultats d'évaluation 167
5.5. Proposition de méthodologie d'enrichissement du thésaurus Xilopix 171
5.6. Conclusion 172

Chapitre 6. Exploitation des hiérarchies sémantiques du *TLFi* pour l'indexation et la recherche d'images 177

6.1. Intégration des relations *is-a* dans un algorithme simple d'indexation textuelle automatique d'images 177
6.1.1. Évaluation des performances du prototype d'indexation 180
6.1.2. Discussion sur les termes d'indexation déterminés 182
6.2. Prise en compte de relations d'association dans un algorithme simple d'indexation textuelle automatique d'images 188
6.2.1. Analyse des résultats d'évaluation de ce second algorithme 189
6.2.2. Bilan sur l'indexation d'images 192
6.3. Exploitation de l'indexation dans un algorithme simple de recherche d'images 193

6.3.1. Évaluation des résultats de recherche 195
6.3.2. Analyse des résultats d'évaluation 198
6.3.2.1. Domaines erronés 198
6.3.2.2. Relations d'association 200
6.3.3. Avantages de notre approche de recherche d'images 201
6.3.3.1. Structuration des résultats de recherche selon les domaines 201
6.3.3.2. Recherche associative d'images 203
6.3.3.3. Requêtes complexes 204
6.3.4. Bilan sur la recherche d'images 205
6.4. Conclusion 205
Conclusion générale et perspectives 207
Contributions 207
Limites et perspectives 209
Liste des figures 213
Liste des tableaux 217
Index des notions 219
Références bibliographiques 223
Annexe 1. Exemples de hiérarchies sémantiques 235

Introduction générale

Avec l'arrivée de l'Internet, le marché de l'image numérique a progressé de manière exponentielle. L'offre d'illustrations n'a jamais été aussi grande. Sachant qu'une agence photo gère habituellement entre un et vingt millions d'images, qu'un satellite météo envoie plusieurs gigaoctets de données chaque jour, qu'un possesseur d'appareil photo numérique actif prendra de l'ordre de cent mille photos en trente ans (Gros, 2007), l'accès et la recherche dans cette masse énorme d'informations posent de nouveaux défis. L'organisation non structurée des images provoque un très grand désordre et une extraordinaire confusion lorsqu'on cherche à les identifier ou les repérer d'autant que leurs descriptions textuelles associées ne sont souvent pas suffisamment explicites pour permettre leur structuration. Afin de gérer et d'utiliser efficacement de telles bases d'images, un système d'indexation et de recherche est donc indispensable. C'est pour cette raison que l'indexation et la recherche d'images sont devenues des sujets très actifs dans la communauté internationale et ont connu un véritable engouement au cours des deux dernières décennies.

L'objectif d'un système d'indexation et de recherche d'images est d'organiser l'information selon certains critères et de permettre aux utilisateurs un accès rapide et une recherche d'images qui correspondent au mieux à leurs besoins d'information. Il existe deux grandes approches concernant l'indexation et la recherche d'images : soit *par le contenu visuel,* soit par *le contenu textuel* de leur description

La plupart des systèmes d'indexation et de recherche d'images, notamment ceux existant sur le Web tels que Google Images[1], Flickr[2], Fotolia[3], etc., sont basés sur la recherche par le contenu textuel en utilisant des mots-clés. Les procédures d'accès aux données photographiques peuvent alors être schématisées ainsi : l'utilisateur formule une requête composée de termes langagiers et le système lui propose en réponse des images qu'il considère comme proches des termes de sa requête. Pour ce faire, le système, le plus souvent à l'aide de calculs statistiques, détermine la similarité entre les termes de la requête et les termes associés aux images.

[1] http://images.google.fr/
[2] http://www.flickr.com/
[3] http://fr.fotolia.com/

Toutefois, les images proposées à l'utilisateur ne semblent pas toujours correspondre à sa requête initiale. Ces écarts sont liés essentiellement à l'opacité des systèmes de recherche par rapport à la sémantique : dans la plupart des cas, aucune analyse du contenu sémantique de la requête n'est en fait effectuée pour déterminer sa signification. On se trouve donc confronté aux phénomènes de polysémie (une même unité lexicale peut représenter des concepts différents) ou de synonymie (le même concept peut être formulé de plusieurs façons) des mots-clés. De plus, la description textuelle de l'image est souvent trop succincte pour décrire entièrement son contenu. Enfin, une autre difficulté réside dans les aspects subjectifs du contenu d'une image, qui dépendent du domaine de connaissance et de la perception de celui qui la regarde, et qui déterminent la diversité de description d'une image.

Contexte de la thèse

Notre travail de thèse se situe à la croisée entre les besoins d'une jeune entreprise, Xilopix[4], et les compétences et ressources du laboratoire ATILF[5] et s'est effectué dans le cadre d'une CIFRE (Convention Industrielle de Formation par la Recherche).

Jusqu'en 2011, Xilopix proposait une encyclopédie photographique et coopérative offrant des usages tant professionnels (fonctions de photothèque) que personnels (e-cartes, fonds d'écran, cartes postales, etc.). Ce fonds d'images était enrichi chaque jour par des auteurs devant indexer manuellement leurs images en utilisant un thésaurus développé par l'équipe de documentalistes de l'entreprise. Pour ce faire, l'entreprise proposait, à partir d'une table d'indexation, une recherche multicritère à base de mots-clés en s'appuyant sur la hiérarchie des descripteurs du thésaurus développé manuellement. Les problèmes principaux qui se posaient alors étaient, d'une part, pour l'indexation « automatique », d'être capable de mettre en correspondance la description d'une image fournie par son auteur, à travers un ensemble de mots-clés ou un court texte de description, avec les descripteurs hiérarchiques du thésaurus Xilopix ; d'autre part, pour la consultation, d'être capable de mettre en correspondance la formulation du besoin d'un utilisateur, exprimée à travers un ensemble de mots-clés, avec les

[4] http://fr.xilopix.com/
[5] www.atilf.fr

descripteurs hiérarchiques du thésaurus Xilopix. Depuis lors, les travaux de l'entreprise, qui se sont orientés uniquement vers l'élaboration d'un moteur de recherche d'images universel, restent fortement liés à cette problématique d'indexation et de recherche d'images.

Pour sa part, l'ATILF, Laboratoire d'Analyse et de Traitement Informatique de la Langue Française (www.atilf.fr), a une compétence éprouvée en analyse et traitement de la langue et dispose d'un ensemble de ressources lexicales remarquables sur la langue française (voir le portail lexical du Centre National de Ressources Textuelles et Lexicales : CNRTL www.cnrtl.fr). En particulier, l'ATILF dispose des ressources lexicales nécessaires pour mettre en correspondance sémantiquement des contenus textuels.

Ainsi, la complémentarité entre les besoins de l'entreprise et les compétences du laboratoire a conduit à la mise en place d'une coopération, support de ma convention CIFRE.

Problématique et orientations de recherche

Nous pensons que les informations lexicales peuvent améliorer le processus d'indexation et de recherche d'images. Ce fait a déjà été démontré dans plusieurs travaux à travers l'utilisation de ressources lexicales comme WordNet, DBpedia, etc. Toutefois, il existe peu de ressources lexicales fiables et celles déjà existantes, telles que les dictionnaires de langue, ne sont souvent pas suffisamment formalisées pour être exploitées dans un domaine d'application particulier. Au début des années 80, plusieurs travaux, notamment ceux d'Amsler (1980), Chodorow, Byrd et Heidorn (1985), ont montré qu'il était possible d'extraire des connaissances à partir d'un dictionnaire de langue informatisé. Malgré ce fait, plusieurs problématiques posées par les définitions lexicographiques comme la circularité, l'information incomplète, la polyhiérarchie, etc. ont été soulevées par Ide et Véronis (1995) et aucun travail n'a montré l'utilité de telles connaissances pour les systèmes d'indexation et de recherche d'images. C'est pour cette raison que plusieurs travaux s'intéressent plutôt à construire de nouvelles ressources lexicales, ce qui est une tâche coûteuse en termes de temps et d'argent, plutôt que d'exploiter des ressources existantes.

Or, à travers le *Trésor de la langue française informatisé* (*TLFi*)[6] (ATILF, 2004), le laboratoire ATILF dispose d'une remarquable ressource lexicale qui, c'est notre conviction, devrait pouvoir améliorer les processus d'indexation et de recherche d'images. Dans le cadre de cette thèse, notre objectif est de proposer une méthodologie d'exploitation automatique de la sémantique lexicale des connaissances lexicographiques du *TLFi* afin de les rendre utilisables pour une application en recherche d'images. En poursuivant ces recherches, nous cherchons à prouver que la sémantique lexicale extraite à partir des définitions du *TLFi* permet d'améliorer l'indexation et la recherche d'images. Pour pouvoir expliciter l'information contenue implicitement dans le *TLFi,* dans un premier temps, nous proposons une formule de pondération de noms dans les définitions du *TLFi* en faisant l'hypothèse que les noms de poids maximal représentent de bons candidats hyperonymes possibles des lexèmes. Dans un deuxième temps, nous proposons une approche de construction automatique de hiérarchies sémantiques à partir du *TLFi* pour l'enrichissement d'un thésaurus construit manuellement au sein de l'entreprise Xilopix. Afin de montrer que de telles hiérarchies sémantiques peuvent être exploitables pour l'indexation automatique d'images à partir de leurs descriptions textuelles associées, nous fournissons les résultats de la mise en œuvre d'un prototype. Ces résultats montrent aussi que l'exploitation d'une telle ressource dans le domaine de recherche d'images améliore la précision de la recherche en structurant les résultats selon les domaines auxquels les concepts de la requête de recherche peuvent faire référence.

Ainsi dans cette thèse nous nous sommes centrée sur diverses questions de recherche précisant :

- Comment expliciter les connaissances lexicographiques du *TLFi* afin de les rendre utilisables par les applications du TAL ;
- Comment à partir des définitions d'un dictionnaire de langue, tel le *TLFi*, enrichir un thésaurus servant de base pour l'indexation et la recherche d'images ;
- La valeur ajoutée à l'indexation et à la recherche d'images rendue possible grâce à l'utilisation d'un dictionnaire de langue.

Sur le plan pratique, pour valider nos propositions nous terminerons cette thèse par la présentation des résultats d'un prototype simple de recherche d'images qui exploite les

[6] www.atilf.fr/tfli

connaissances sémantiques du *TLFi*. Les résultats d'évaluation des performances d'un tel prototype valident notre thèse de départ et confirment la possibilité d'exploitation de ressources lexicales existantes, comme des dictionnaires de langue, plutôt que de construire de nouvelles ressources pour améliorer les processus d'indexation et de recherche d'images.

Organisation du manuscrit

Ce mémoire de thèse est organisé en six chapitres.

- Dans le chapitre 1, nous présentons la principale ressource lexicale utilisée dans notre recherche, le *TLFi*. Nous réalisons une analyse de sa macro- et microstructure et de la typologie des définitions lexicographiques qu'il contient. Nous présentons aussi les principaux projets de recherche qui utilisent le *TLFi*.
- Le chapitre 2 présente les fondements de la recherche d'informations textuelles. Nous y présentons un aperçu de principaux modèles qui sont à la base de systèmes de RI actuels, l'architecture générale d'un système de RI ainsi que les mesures d'évaluation utilisées. Ensuite, nous portons notre attention sur les systèmes de recherche d'images, en décrivant les principaux types d'indexation et de recherche utilisés.
- Le chapitre 3 est consacré à l'utilisation des ressources lexicales pour l'indexation et la recherche d'images. Nous décrivons les principaux types de ressources lexicales existantes ainsi que les approches d'exploitation de ces ressources lexicales dans le domaine de l'indexation et de la recherche d'images. Ensuite, nous présentons les études de construction de hiérarchies sémantiques à partir de textes, en nous focalisant particulièrement sur celles qui exploitent des définitions lexicographiques et concluons ce chapitre par une présentation des questions de recherche que pose une telle approche.
- Dans le chapitre 4, nous abordons la question de l'extraction automatique d'informations pertinentes au sein des définitions du *TLFi*. Dans ce but, nous analysons plusieurs critères de pondération et proposons une formule de pondération des noms des définitions. Enfin, nous décrivons les évaluations

effectuées afin de valider la formule proposée ainsi que le travail de normalisation des domaines des définitions du *TLFi* qui a été effectué.

- Dans le chapitre 5, nous nous intéressons à la façon d'organiser les noms des définitions de tous les lexèmes d'un même vocable en nous appuyant sur leur poids. Ensuite, nous décrivons la méthodologie de construction automatique de hiérarchies sémantiques pour les lexèmes d'un vocable du *TLFi* que nous proposons. Enfin, nous présentons les résultats des évaluations des relations hyperonymiques et proposons une méthodologie d'enrichissement du thésaurus existant en utilisant des hiérarchies sémantiques générées à partir du *TLFi*.
- Le chapitre 6 correspond à la dimension applicative de la thèse et présente l'application des hiérarchies sémantiques obtenues à partir du *TLFi* pour l'indexation et la recherche d'images. Nous y décrivons les algorithmes d'indexation automatique et de recherche d'images proposés et y présentons l'analyse des résultats d'évaluation à travers un prototype simple qui a été mis en place.

En conclusion, nous dressons un résumé des principales contributions effectuées dans le cadre de cette thèse et discutons enfin des limites et des perspectives d'application de ce travail.

CHAPITRE 1

Le *Trésor de la langue française informatisé* (*TLFi*) et sa structure

Sommaire

1.1. Du *TLF* vers le *TLFi* 21
1.2. Macrostructure et microstructure du *TLFi* 24
1.3. Modèle informatique des données du *TLFi* 29
1.4. Les divers accès possibles aux données du *TLFi* en s'appuyant sur sa structure 31
1.5. Les divers types de définitions lexicographiques 34
1.5.1. Définition hyperonymique (logique) 35
1.5.2. Définition synonymique 36
1.5.3. Définition dérivative 36
1.5.4. Définition méréonymique 37
1.5.5. Définition approximative 38
1.6. Repérage des modes définitoires dans le *TLFi* 38
1.7. Le *TLFi* et les projets connexes 40
1.7.1. Projet SEMEME 41
1.7.2. Projet Definiens et ses résultats 43
1.7.3. Utilisation du *TLFi* pour l'enrichissement d'une ontologie 47
1.7.4. Projet RELIEF 48
1.8. Conclusion 48

Chapitre 1. Le *Trésor de la langue française informatisé* (*TLFi*) et sa structure

Dans ce premier chapitre, nous nous attachons à présenter les connaissances du *Trésor de la langue française informatisé (TLFi)* que nous souhaitons exploiter pour construire automatiquement des hiérarchies sémantiques pouvant tout à la fois (cf. chapitre 5) enrichir, d'une part, les mots-clés utilisés comme descripteurs d'images et d'autre part, le thésaurus existant construit manuellement et utilisé actuellement pour l'indexation et la recherche d'images au sein de Xilopix.

Nous commençons par une brève présentation du *TLF* ainsi que son processus d'informatisation. Nous analysons ensuite sa macro- et microstructure et les divers accès aux données du *TLFi* rendus possibles par l'exploitation de sa structure. Ensuite, nous analysons les principaux types de définitions lexicographiques contenues dans le *TLFi*. Enfin, nous concluons ce chapitre par une présentation de divers projets de recherche qui ont également fait appel au *TLFi* et dont certains serviront aussi d'appui à notre travail.

1.1. Du *TLF* vers le *TLFi*

Le Trésor de la langue française (*TLF*) est le grand dictionnaire de langue française rédigé en seize volumes par l'Institut national de la langue française (INaLF, laboratoire du CNRS)[7]. La réalisation de ce dictionnaire a duré presque trente ans[8]. Les sept premiers volumes (1971-1979) ont été dirigés sous la direction du recteur Paul Imbs et les neuf derniers (1980-1994) sous celle de Bernard Quemada.

Le *TLF* est un dictionnaire de langue, il a comme objectif d'aider son utilisateur à lire, interpréter et décoder correctement les messages de communication orale ou écrite et aussi à écrire et à parler correctement. Un dictionnaire de langue par opposition à un dictionnaire encyclopédique, s'attache en effet à définir une unité lexicale à travers des

[7] Le laboratoire ATILF (Analyse et Traitement Informatique de la Langue Française) est le successeur de l'INaLF à Nancy.

[8] En effet, c'est en 1957, pendant le colloque organisé par Paul Imbs à Strasbourg, que le croquis du grand projet qui s'appelait le *Trésor général de la langue française* a été présenté (Pruvost, 2002).

informations concernant sa nature grammaticale, son orthographe, son étymologie, son histoire, ses significations, ses spécialisations, etc. Par ailleurs, un dictionnaire de langue peut se différencier selon l'attitude adoptée sur le lexique de la langue par les rédacteurs. Le *TLF,* pour ce qui le concerne, est destiné selon P. Imbs à un public constitué des « hommes cultivés […] aujourd'hui appelés les cadres supérieurs ou moyens de la société », ce qui suppose un certain niveau culturel. C'est pour cette raison que le vocabulaire du *TLF* est général, tout en intégrant un vocabulaire plus spécifique à travers des définitions pour des domaines plus spécialisés. Dictionnaire institutionnel de référence du français, il se donne comme objectif de définir les usages de la langue écrite. Notons par ailleurs que, comme l'indique J. -M. Pierrel dans la préface du *TLFi,* « le *Trésor de la langue française* est le premier dictionnaire de langue se fondant sur une méthodologie systématique d'analyse des usages effectifs des mots de notre langue à travers l'exploitation d'une vaste base de données textuelles dont la saisie a débuté dès les années 60 et dont le but premier était de fournir des données organisées aux rédacteurs du dictionnaire *TLF ».* L'utilisation de concordanciers a permis aux rédacteurs de cerner les contextes d'emploi des unités lexicales et ensuite de sélectionner des exemples utilisés dans le dictionnaire. Selon Pruvost (2002, p. 77) le corpus documentaire du *TLF* était constitué à 80 % de textes littéraires et à 20 % de textes scientifiques issus essentiellement de ce qui est devenu aujourd'hui la base textuelle Frantext[9] (Bernet & Pierrel, 2005).

Après de longues années de travail mené par une importante équipe de lexicographes et de chercheurs, l'ouvrage entier du *TLF* contient « 23 000 pages, 100 000 mots, 450 000 entrées, 500 000 citations précisément identifiées » (Pruvost, 2002, p. 78).

L'apparition du dernier volume du *TLF* (1994) correspond à une période où, avec le progrès de l'informatique, plusieurs dictionnaires sont déjà proposés sur support informatique. Dès que le travail sur le *TLF* s'est achevé, un autre problème s'est donc posé à l'équipe de l'INaLF, celui de son informatisation. Ainsi, grâce à Jacques Dendien, « un thaumaturge de l'informatique » (Pruvost, 2002, p. 86), et son équipe, une étude de son informatisation est lancée dès le milieu des années 90 et le *TLFi* apparaît sur Internet en 2001[10].

[9] www.frantext.fr
[10] www.atilf.fr/tlfi

Pour que le *TLF* soit accessible sur Internet, un gros travail a été fait et beaucoup de problèmes ont été résolus en plusieurs étapes. Nous présentons ci-dessous uniquement les principales étapes de l'informatisation du *TLF* (Dedien & Pierrel, 2003) :

1. *Saisie initiale du dictionnaire.*

Les volumes du *TLF* existaient sous différents formats, le premier objectif a donc été de tous les transcrire sur un support informatique unique. Ainsi, la saisie des huit premiers volumes a été effectuée par une entreprise privée grâce au financement de la Bibliothèque Nationale de France. Les tomes 9 à 16 ont été traités par une entreprise spécialisée dans le traitement des archives de photocomposition car ils existaient en formats de photocomposition de trois types différents. Seuls les volumes 9 à 10 et 11 à 13, parce qu'ils étaient en mauvais état, ils ont été remis en ordre et reconstitués au sein de l'INaLF.

2. *Balisage du dictionnaire.*

Ce processus visait à introduire les balises textuelles de type SGML[11]. Il a été automatisé grâce à des automates de reconnaissance de différentes formes. Au terme d'un processus itératif, le taux de réussite des automates a atteint de l'ordre de 99,8 %. Plus récemment, le *TLF* a été converti aux normes de balisage XML. Au final, le *TLFi* contient 36 613 712 balises XML de plusieurs types : balises d'article, de lemme, de code grammatical, de définition, de synonymes, etc[12].

3. *Moteur de recherche Stella.*

Un moteur de recherche, Stella, a été élaboré par Jacques Dendien et permet l'exploitation de la structure XML du *TLFi*. Grâce à ce moteur, le *TLFi* peut être interrogé au moyen de requêtes plus ou moins complexes en spécifiant les types des objets recherchés, ainsi que les relations hiérarchiques entre eux, par exemple « Chercher, dans le domaine de l'agriculture, les définitions relatives à un instrument et empruntées au dictionnaire de l'Académie » (Dedien & Pierrel, 2003).

4. *Hypernavigation.*

Une possibilité d'hypernavigation dans le *TLFi* permet à l'utilisateur, en sélectionnant une unité lexicale, de déclencher une recherche dans des bases textuelles différentes

[11] SGML (Standard Generalized Markup Language) est un langage de codage de données, utilisé avant l'apparition du XML (Extensible Markup Language).
[12] Nous définissons ces notions dans la section 1.2.

comme Frantext[13], la huitième édition du *Dictionnaire de l'Académie française* (édition du 1935), etc.

5. Élaboration d'une base de données phonétiques.

Une base de données phonétiques a été créée en associant à chaque graphie de la base la représentation phonétique de sa prononciation. Cette base de données permet une recherche dans le *TLFi* selon la forme phonétique d'une unité lexicale dont la graphie n'a pas été correctement saisie par l'utilisateur. Ainsi, à partir de *jenero,* le moteur trouve dans la base la forme *généraux* et, après la lemmatisation, propose à l'utilisateur le lemme *général*.

Le *TLF* est passé par toutes ces étapes avant d'apparaître sur Internet. Un travail important a été effectué et beaucoup de problèmes techniques qui paraissaient sans solution au début des années 1990 ont été résolus grâce au collectif de recherche du laboratoire. En 2004 est publiée la version CDROM du *TLF* informatisé (*TLFi*)[14].

1.2. Macrostructure et microstructure du *TLFi*

Le *TLFi*, image exacte du *TLF,* est donc un dictionnaire de langue et, comme tous les dictionnaires de ce type, est organisé en suivant une certaine structure. Ainsi, la *macrostructure* d'un dictionnaire désigne l'organisation de l'ensemble des *vocables*[15], tandis que la *microstructure* désigne l'organisation des informations présentes dans chaque *article*[16] de vocable. La macrostructure d'un dictionnaire comprend une *nomenclature* qui est constituée d'une liste de tous les vocables des articles du dictionnaire, nommés *mots-vedettes* ou *entrées*.

Pour illustrer tous les composants relevant de la macrostructure, nous allons examiner tout d'abord les vocables homonymes[17] AVOCAT[1] et AVOCAT[2] du *TLFi* (cf. figure 1.1.). Chaque acception particulière du vocable, nommée *lexie*, est présentée dans l'article du vocable. Au sein de l'article, chaque lexie est généralement numérotée et

[13] Frantext est une base de données de textes français : textes littéraires et philosophiques, mais aussi scientifiques et techniques (environ 10 %), développée et maintenue au sein de l'ATILF-CNRS.
[14] http://atilf.atilf.fr/tlf.htm
[15] Un vocable est un regroupement de lexies qui sont associées aux mêmes signifiants et liées par un lien sémantique évident (Polguère, 2008, p. 59).
[16] Un article de dictionnaire représente le bloc de texte qui décrit un vocable donné (Polguère, 2008, p. 233).
[17] Les vocables homonymes sont les vocables dont les lexies sont associées aux mêmes signifiants mais ne représentent aucune relation sémantique entre elles.

correspond à un sous-article qui la décrit, appelé aussi « article de lexie ». En effet, une lexie est une unité lexicale, soit un *lexème*[18] (ex. AVOCAT I., AVOCAT I.A.) soit une *locution*[19] (ex. AVOCAT CONSEIL), associée à un sens donné. Ainsi, le vocable AVOCAT[1] est un vocable polysémique car il contient plusieurs lexies (AVOCAT[1]1 « Personne défendant les intérêts », AVOCAT[1]2 « Personne qui, étant inscrite au barreau », etc.), tandis que le vocable AVOCAT[2] est un vocable monosémique (AVOCAT[2] « Baie comestible en forme de poire »). Nous tenons à préciser ici que dans la suite de manuscrit nous allons utiliser cette terminologie dans un cadre lexicographique. Dans un contexte plus général, étant donné que le terme *mot* est fortement ambigu, nous utilisons à sa place soit le terme *mot-forme* pour désigner un signe linguistique, soit le terme *lexème* pour l'unité lexicale. Le terme *mot-clé*, quant à lui, est utilisé lors des recherches d'informations.

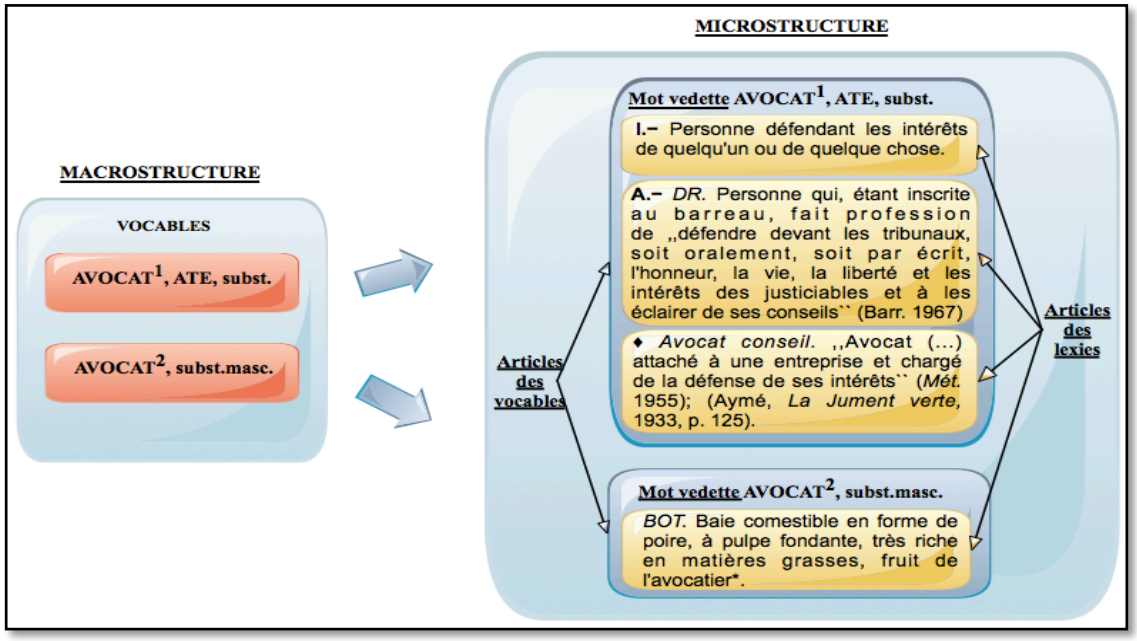


Figure 1.1. Exemple d'organisation générale de macro- et microstructure des vocables AVOCAT[1] et AVOCAT[2] dans le *TLFi*

[18] Lexème, par exemple AVOCAT[1]1 [**I.** *Personne défendant les intérêts de quelqu'un ou de quelque chose.*], est une unité mono-lexicale structurée autour d'un sens exprimable par un regroupement de mots-formes (*avocat, avocats, avocate, avocates*) que seule distingue la flexion.
[19] Locution, par exemple AVOCAT CONSEIL [*Avocat (...) attaché à une entreprise et chargé de la défense de ses intérêts*], est une unité pluri-lexicale structurée autour d'un sens exprimable par un regroupement de syntagmes figés que seule distingue la flexion.

Ainsi, la macrostructure du *TLFi* contient 93 697 vocables organisés par ordre alphabétique, dont 55,26 % appartiennent à la catégorie nom (cf. figure 1.2.) ; le nombre des articles étant de 50 471.

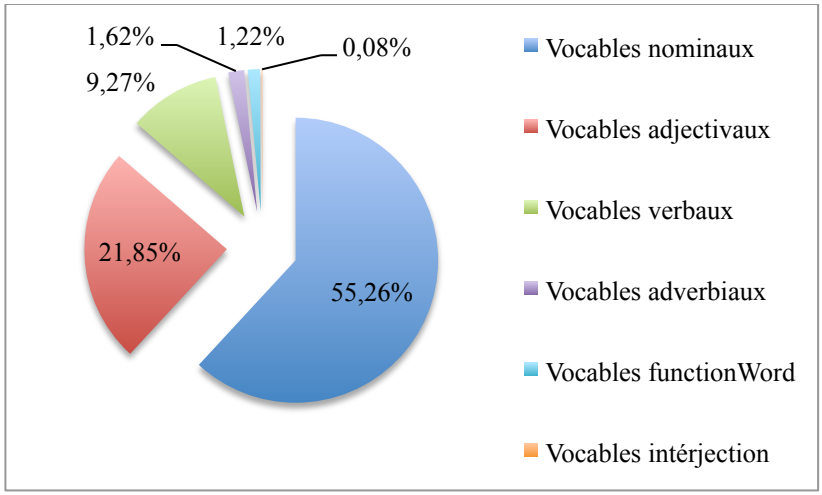


Figure 1.2. Taux des vocables du *TLFi* appartenant aux différentes catégories grammaticales

La microstructure du *TLFi* (cf. figure 1.3.) est hiérarchisée et donne accès à une information riche et hétérogène que nous décrivons ci-dessous :

1. Dans le *TLFi*, un vocable peut renvoyer vers un autre vocable. Par exemple, le vocable PARTNER renvoie vers l'article du vocable PARTENAIRE (ex. PARTNER, voir PARTENAIRE).
2. La vedette principale d'un article représente le vocable traité dans l'article de dictionnaire, c'est-à-dire le mot-vedette et sa catégorie lexicale (ex. *LION, LIONNE, subst.*).
3. Dans le *TLFi*, l'information synchronique[20] est organisée selon une structure hiérarchique qui contient :
   - 3.1 Une marque de plan, telle que *I* ou *a.*
   - 3.2 Un bloc logique d'information encapsule un ensemble d'informations structurellement cohérentes. Il peut contenir tous les objets textuels qui sont interrogeables dans l'interface du *TLFi* et d'autres aussi :

[20] La situation d'une langue à un moment donné.

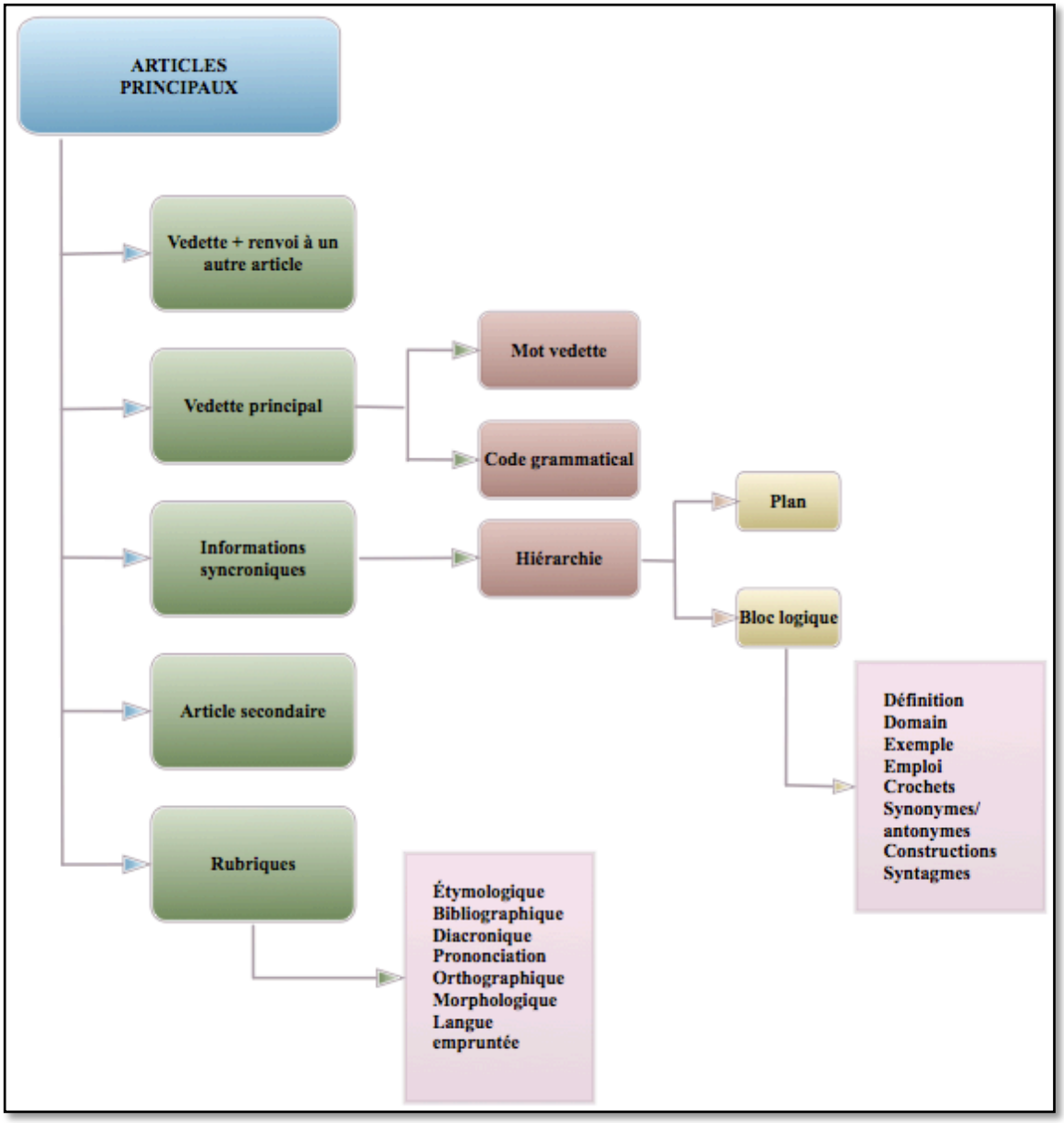


Figure 1.3. Microstructure du *TLFi*

- une définition produite par les rédacteurs du *TLF* ou bien reprise d'une source antérieure ;
- un ou plusieurs exemples, enchaînés (à la suite du texte) ou détachés (retour à la ligne + numérotation) ;
- une indication d'emploi (ex. *P. méton., P. anal., P. métaph.,* etc.) ;
- une indication de domaine général ou technique (ex. *MYTH., HÉRALD., ICONOGR. RELIG., ASTROL., ZOOLOGIE,* etc.) ;

- des informations entre crochets, spécifiant l'usage (ex. *[À propos d'hommes], [Avec une idée de férocité, de puissance déchaînée]*) ;
- un paragraphe décrivant éventuellement plusieurs syntagmes avec possibilité de retours à la ligne (ex. *SYNT. Lion roux; lion de l'Atlas*) ;
- un renvoi à une référence externe au dictionnaire (ex. *cf. BALZAC, Eugénie Grandet, 1834, p.190*), à un autre vocable du dictionnaire (ex. *cf. ABC*), à un autre sens d'une lexie d'un vocable du dictionnaire (ex. *cf. ABAISSER I*) ;
- un ou plusieurs synonymes (ex. *synon. fillette*) ou antonymes (ex. *anton. se tenir*) ;
- un ou plusieurs schémas de construction décrivant des portions grammaticales particulières sous la forme de patrons syntaxiques tels que « qqn donne qqc à qqn » (ex. *A +subst. de l'animé ou de l'inanimé*) ;
- une remarque enchaînée dans le texte (ex. *Rem. Attesté ds les princ. Dict. gén. Du xix^{e} et du xx^{e} siècle*).

4. Des articles secondaires peuvent être inclus dans des articles principaux et sont utilisés pour les vocables construits à l'aide de préfixes, suffixes ou vocables dérivés d'un autre vocable. Par exemple, l'article du vocable PASSE-SINGE se trouve dans l'article du vocable PASSE- et les articles des vocables SINGESQUE, SINGESSE dans celui de SINGE. Environ 50 % des vocables dans le *TLFi* sont secondaires étant cachés dans les articles d'autres vocables, ce qui rendait d'ailleurs la recherche dans le *TLF* papier très difficile. La version informatisée du *TLF* a corrigé ce défaut et permet la recherche dans l'ensemble des entrées du dictionnaire.

   Il convient de noter que le *Trésor de la langue française* possède un nombre non négligeable d'entrées correspondant à des éléments formants tels -PATHIE, -PATHIQUE, -PATHE qui rend difficilement accessibles les nombreux vocables définis sous cette entrée dont certains très utilisés aujourd'hui, telle EMPATHIE, auraient peut-être mérité d'une entrée spécifique (cf. figure 1.4.).

Il existe en fait dans le *TLFi* dix-neuf éléments de rubrique ; nous en présentons quelques-uns ci-dessous :

- indication d'une langue empruntée. Par exemple, PENDULE est un emprunt du latin (ex. *Empr. au lat.*) ;
- rubrique bibliographique (ex. *BBG.- Quem. DDL t.6, 27*) ;
- rubrique étymologique (ex. *Etymol. et Hist.*) ;
- rubrique de prononciation et d'orthographe (ex. *Prononc. et Orth.*) ;
- rubrique diachronique incluant l'histoire et l'étymologie du vocable traité (ex. **pent***hode* ou **pent***ode* (*-ode*, de *(électr)ode*, du gr.).

***-PATHIE**, -PATHIQUE, -PATHE, élém. formants*
Élém. tirés du gr. «ce que l'on éprouve [de mal]» et entrant dans la constr. de mots sav. appartenant notamment au vocab. de la méd.; *-pathie* est empr. au gr. ; les subst. fém. constr. peuvent générer des adj. dér. en *-pathique* et des dér. régressifs en *-pathe,* adj. ou subst. masc.; à noter l'adj. *protopathique* (*infra* I A) qui semble usité sans terme en *-pathie* correspondant.
I. [*-pathie* exprime une manière d'être affecté, de sentir, de ressentir, caractérisée par le 1er élém.; celui-ci est le plus souvent issu du gr.]
A. [Les mots constr. désignent ou expriment un rapport avec des phénomènes sensitifs d'ordre pathol.] :
*hyper***pathie.** Sensibilité exagérée à la douleur, qui s'observe notamment dans les syndromes thalamiques du côté de la lésion cérébrale. La douleur est très pénible et persiste après l'arrêt de l'excitation (*Méd. Biol.* t.2 1971). *Hyper***pathique.** Qui se rapporte à l'hyperpathie *(Méd. Biol.* t.2 1971).
*proto***pathique,** psychol. [P. oppos. à *épicritique (s.v. épi-)*] Qui est apte à percevoir uniquement des stimulations sensitives, tactiles ou thermiques, grossières (*Méd. Biol.* t.3 1972). *Même dans les expériences de Head et Rivers, où les paliers sont nets, au cours de la régénération du nerf, la sensibilité épicritique reparue ne supprime pas entièrement la sensibilité protopathique* (RUYER, *Conscience,* 1937, p.83). […]
B. [Les mots constr. désignent un rapport à autrui]:
***intro*pathie** (*intro-,* v. *intra* rem. gén. 2). Synon. de ***empathie** (infra)* (d'apr. MARCH. 1970). *La connaissance de moi-même est toujours à quelque degré un guide dans le déchiffrement d'autrui, bien qu'autrui soit d'abord et principalement une révélation originale de l'intropathie* (RICOEUR, *Philos. volonté,* 1949, p.14).

Figure 1.4. Exemple complexe d'une entrée cachée sous la définition de l'élément formant -PATHIE dans le *TLFi*

1.3. Modèle informatique des données du *TLFi*

Le modèle informatique des données du *TLFi* est relativement simple. Dans ce modèle, à chaque vocable du *TLFi* correspond un identifiant unique (entryID), l'identifiant de son article (articleID) et l'identifiant de l'article de son père (parentID) dans le cas d'un article secondaire.

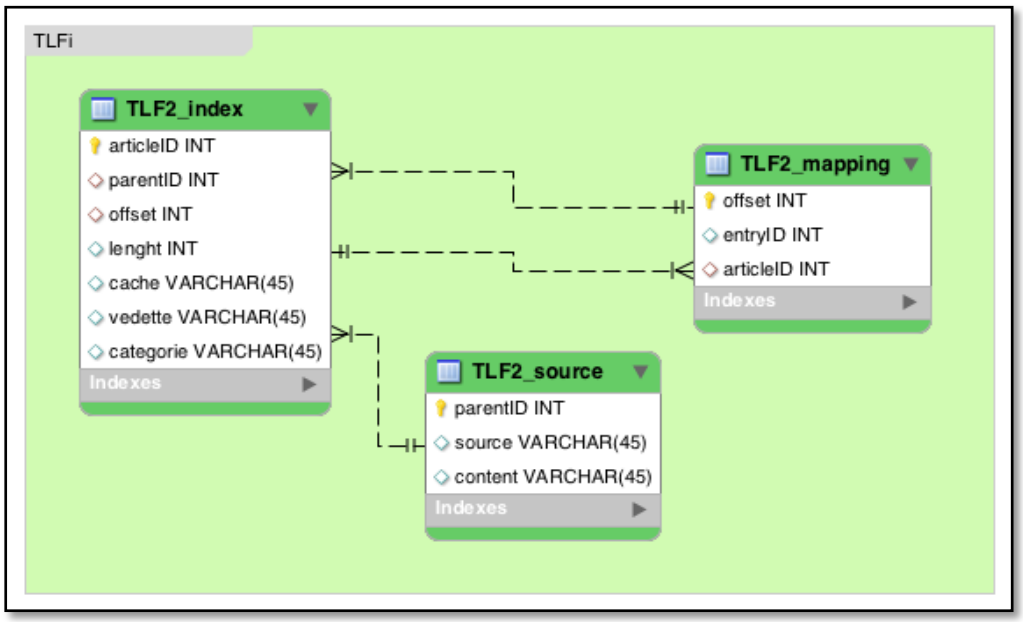


Figure 1.5. Modèle de données du *TLFi*

Nous pouvons constater qu'à partir de ce modèle de données du *TLFi* n'apparaît directement que l'information sur le mot-vedette et sa catégorie grammaticale. Nous pouvons aussi accéder aux articles des vocables, mais leur structure n'est pas de tout modélisée. Ce modèle structuré se limite donc à l'information sur la macrostructure du *TLFi* tandis que l'information sur la microstructure bien quelle soit annotée et exploitable dans le *TLFi* ne permet pas d'exhiber un modèle stable de cette microstructure pour l'ensemble des volumes de dictionnaire. Cela se peut s'expliquer par les évolutions importantes de cette microstructure qui s'est stabilisée au fur et à mesure de la rédaction du *TLFi* qui s'entendit sur plus de vingt-cinq ans. C'est pour cette raison que dans nos recherches nous ne pouvons pas nous contenter de ce modèle de données du *TLFi*. Nous avons besoin d'un modèle de la microstructure du *TLFi* qui nous fournira l'information sur les données de cette dernière, par exemple le nombre de définitions dans un article, le nombre de synonymes correspondants à une lexie d'un vocable, etc. Nous verrons au paragraphe 1.7.1. que le projet SEMEME, dont nous utiliserons les résultats, a eu comme objectif, entre autres, de structurer cette information de microstructure du *TLFi*.

1.4. Les divers accès possibles aux données du *TLFi* en s'appuyant sur sa structure

L'informatisation du *TLF* l'a rendu accessible à un public plus élargi et ses modes de consultation sont devenus plus souples que ceux d'une utilisation manuelle des volumes imprimés. Ainsi, l'interface d'interrogation du *TLFi* permet plusieurs types de recherche à travers le croisement de multiples critères que nous présentons ci-dessous :

1. *Recherche d'un mot-forme.*

Ce type de recherche permet une simple recherche d'un mot-forme dans le *TLFi* en proposant comme résultat les articles des vocables correspondants à celui-ci. Comme l'indique J.-M. Pierrel (2011) « Cette recherche permet un accès à un mot à travers un système de correction et de lemmatisation automatique (forcée ou non) : ainsi, en introduisant la recherche de la forme *etique* (sans accent), on accède aux deux articles correspondant aux mots *étique* ou *éthique* ; de même un accès à partir de la forme *sussiez* permet d'obtenir automatiquement l'article *savoir*. Elle donne aussi la possibilité d'obtention directe des définitions et conditions d'usage d'une unité lexicale ne disposant pas d'un traitement lexicographique autonome (on accède par exemple au substantif masculin *trompette*, traité dans un super-article *trompette* qui englobe tant le masculin que le féminin, à travers la requête « le trompette ») ou d'une expression comme *battre la mesure*, en focalisant la réponse sur l'élément pertinent demandé et en offrant la possibilité, à l'aide d'une sorte de « stabilo boss » électronique, de surligner tel ou tel objet textuel. ».

Il est aussi possible de rechercher le mot-forme soit en le sélectionnant dans les listes défilantes de tranches alphabétiques dans lesquelles il se trouve soit en faisant une recherche phonétique.

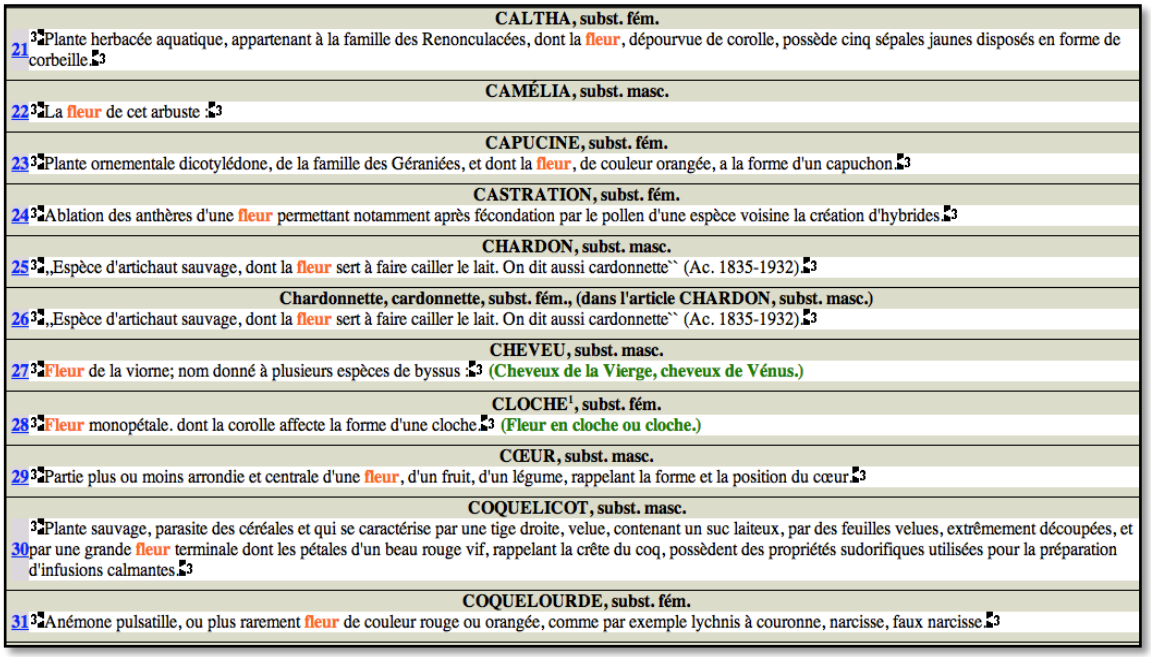
2. *Recherche assistée.*

Cette recherche assistée qui exploite la structure du dictionnaire permet de rechercher dans le *TLFi* une unité lexicale donnée en indiquant sa catégorie grammaticale, en choisissant une discipline parmi les 150 proposées (ex. *arboriculture, boxe, industrie,* etc.) et éventuellement un indicateur d'emploi (ex. *littéraire, familier, péjoratif*, etc.). Elle permet aussi d'indiquer le contenu qui doit ou non être trouvé dans tel ou tel type d'objet (ex. *définition, exemple, syntagme,* etc.) de l'article du vocable.

Une telle recherche assistée permet de retrouver, par exemple, la liste de tous les vocables du *TLFi* qui contiennent dans ses définitions pour le domaine *botanique* le nom *fleur* en formulant la requête suivante :

```
X: code_grammatical(substantif);
   domaine(botanique);
    définition(fleur).
```

La figure 1.6. présente un extrait des résultats obtenus pour cette requête.



CALTHA, subst. fém.

21 Plante herbacée aquatique, appartenant à la famille des Renonculacées, dont la **fleur**, dépourvue de corolle, possède cinq sépales jaunes disposés en forme de corbeille.

CAMÉLIA, subst. masc.

22 La **fleur** de cet arbuste

CAPUCINE, subst. fém.

23 Plante ornementale dicotylédone, de la famille des Géraniées, et dont la **fleur**, de couleur orangée, a la forme d'un capuchon.

CASTRATION, subst. fém.

24 Ablation des anthères d'une **fleur** permettant notamment après fécondation par le pollen d'une espèce voisine la création d'hybrides.

CHARDON, subst. masc.

25 „Espèce d'artichaut sauvage, dont la **fleur** sert à faire cailler le lait. On dit aussi cardonnette`` (Ac. 1835-1932).

Chardonnette, cardonnette, subst. fém., (dans l'article CHARDON, subst. masc.)

26 „Espèce d'artichaut sauvage, dont la **fleur** sert à faire cailler le lait. On dit aussi cardonnette`` (Ac. 1835-1932).

CHEVEU, subst. masc.

27 **Fleur** de la viorne; nom donné à plusieurs espèces de byssus : **(Cheveux de la Vierge, cheveux de Vénus.)**

CLOCHE[1], subst. fém.

28 **Fleur** monopétale. dont la corolle affecte la forme d'une cloche. **(Fleur en cloche ou cloche.)**

CŒUR, subst. masc.

29 Partie plus ou moins arrondie et centrale d'une **fleur**, d'un fruit, d'un légume, rappelant la forme et la position du cœur.

COQUELICOT, subst. masc.

30 Plante sauvage, parasite des céréales et qui se caractérise par une tige droite, velue, contenant un suc laiteux, par des feuilles velues, extrêmement découpées, et par une grande **fleur** terminale dont les pétales d'un beau rouge vif, rappelant la crête du coq, possèdent des propriétés sudorifiques utilisées pour la préparation d'infusions calmantes.

COQUELOURDE, subst. fém.

31 Anémone pulsatille, ou plus rarement **fleur** de couleur rouge ou orangée, comme par exemple lychnis à couronne, narcisse, faux narcisse.

Figure 1.6. Exemple de résultats de recherche dans le *TLFi* avec une recherche assistée du type `X: code_grammatical(substantif) ; domaine(botanique); définition(fleur).`

En analysant les résultats obtenus, nous observons que, dans certaines définitions, *fleur* se trouve en tête de la définition et représente alors le plus souvent l'hyperonyme du lexème défini, ce qui n'est pas le cas lorsque *fleur* apparaît plus loin dans la définition. Ce fait nous permet d'émettre l'hypothèse que les noms situés en début de définitions peuvent être de bons candidats hyperonymes pour les lexèmes définis.

3. *Recherche complexe.*

Ce type de recherche permet de réaliser une recherche plus complexe en indiquant le type de l'objet (ex. *entrée, définition, source,* etc.), la nature des liens entre les objets (ex. *inclus dans l'objet, dépendant de l'objet,* etc.) et le contenu textuel de chaque objet.

Dans ce type de recherche, le contenu textuel peut être spécifié à l'aide d'expressions régulières qui permettent par exemple de trouver toutes les occurrences d'une forme flexionnelle d'un verbe ou seules les définitions qui ne contiennent pas une unité lexicale donnée. Il est aussi possible d'indiquer la position du contenu textuel dans l'objet. Par exemple, nous avons pu trouver la liste de tous les vocables du *TLFi* dont les définitions contiennent en première position le nom *fleur* (cf. figure 1.7.) et le nom *meuble* (cf. figure 1.8.). Nous observons aussi que grâce à ce type de recherche il est possible de trouver une liste de fleurs, d'arbres, de meubles, etc.

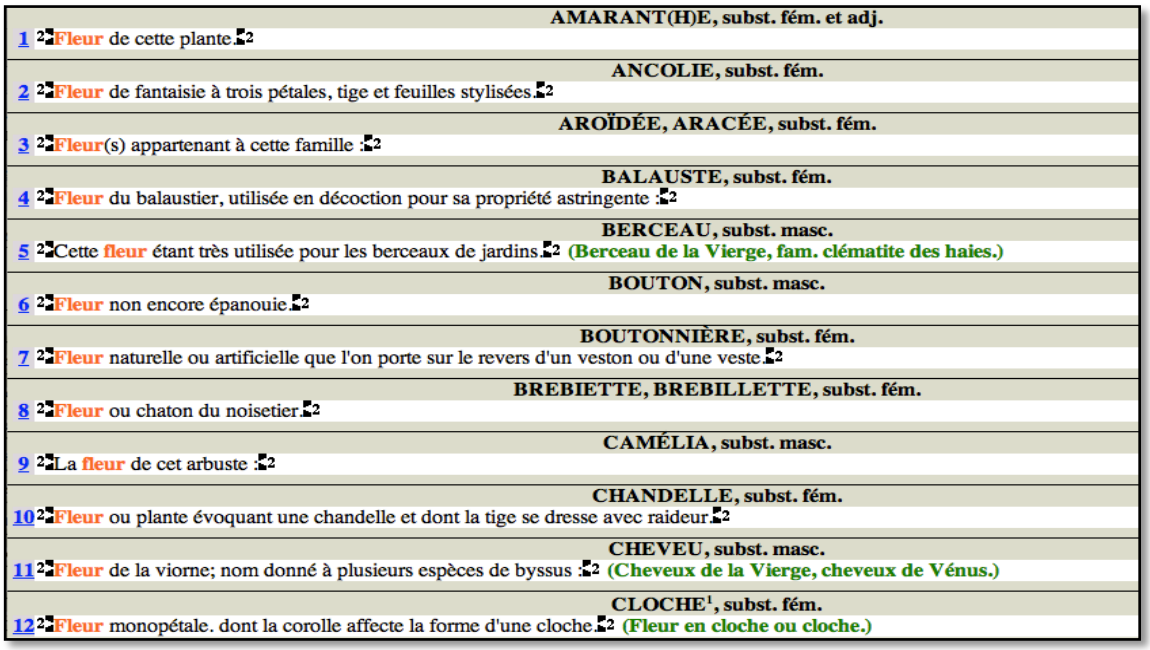


Figure 1.7. Exemple d'une liste de fleurs trouvées dans le *TLFi* avec une recherche complexe du type `X: code_grammatical(substantif); domaine(botanique); définition(fleur(position<2).`

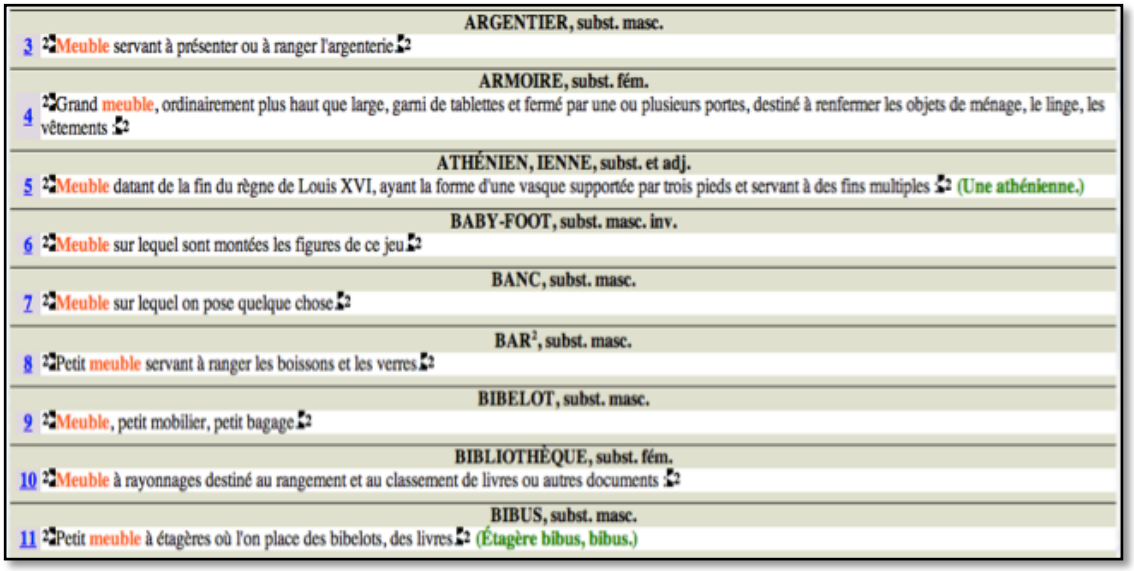


Figure 1.8. Exemple d'une liste de meubles trouvés dans le *TLFi* avec une recherche complexe du type `X: code_grammatical(substantif); définition(meuble(position<2).`

Ces résultats montrent que le nom situé en tête de la définition est classifiant et qu'il représente l'hyperonyme du lexème, ce que renforce notre hypothèse faite lors de la recherche assistée. À ce stade, retenons cette propriété, car dans nos recherches nous sommes intéressée par l'extraction de relations d'hyperonymie à partir des définitions du *TLFi* pour ensuite construire des hiérarchies sémantiques pour chaque vocable donné.

1.5. Les divers types de définitions lexicographiques

L'objectif principal des dictionnaires de langue est d'expliquer les sens des unités lexicales en fournissant une définition. La tâche de rédaction des définitions est attribuée aux lexicographes qui essaient d'expliquer l'unité lexicale par une paraphrase sémantiquement liée à celle-ci. L'énoncé définitionnel rédigé représente une expression paraphrastique et doit permettre au lecteur de construire une signification de l'unité lexicale. Pour ce faire les lexicographes utilisent en fait divers types de définitions.

Nous présentons dans les sous-sections suivantes les principaux types de définitions utilisées dans le *TLFi* en nous basant sur la synthèse présentée par Martin (2001).

1.5.1. Définition hyperonymique (logique)

Selon Choi-Jonin et Delhay (2008, p. 319), la définition hyperonymique est bipartite, constituée d'un *genre prochain* et des *différences spécifiques* : « Héritière de la tradition aristotélicienne, la définition logique ou hyperonymique indique d'une part la classe à laquelle appartient le défini — ou genre prochain — et d'autre part les propriétés qui le particularisent à l'intérieur de cette classe : on les appelle les différences spécifiques. ».

La majorité des définitions des dictionnaires de langue sont de cette catégorie. Par exemple, la définition du *TLFi* du lexème LION est :

> LION. Subst.masc. Mammifère carnivore, de la famille des Félidés, de forte taille, caractérisé par sa face large, sa crinière touffue, son tronc et ses membres trapus, son pelage fauve, et vivant à l'état sauvage surtout en Afrique.

Donc *lion* fait partie de la *famille Félidés*, laquelle représente son genre prochain, et il se distingue du *chat* et du *tigre* par sa *forte taille*, son *pelage fauve*, sa *crinière touffue*, ce qui constitue ses différences spécifiques.

Ce type de définition est souvent appliqué aux noms, mais il convient aussi aux autres catégories grammaticales comme le verbe et l'adjectif. Par exemple, la définition du *TLFi* pour le verbe *courir* est :

> COURIR. V. Se déplacer rapidement par un mouvement successif et accéléré des jambes ou des pattes prenant appui sur le sol.

Ainsi, *courir* est un verbe qui appartient à une classe, *se déplacer*, et se distingue au sein des verbes de cette classe par des propriétés spécifiques qui portent sur le *mouvement successif* et *accéléré des jambes*.

En analysant les définitions hyperonymiques, nous observons que l'hyperonyme qui indique la classe à laquelle appartient le lexème défini est de même catégorie grammaticale que celui-ci (ex. *lion (nom)-mammifère (nom), courir (verbe)-se déplacer (verbe)*). Dans une définition, l'hyperonyme peut être aussi un syntagme (ex. « ramener (un malade, un blessé) à la vie par … » RÉANIMER). Ainsi, il peut être simple (un seul hyperonyme ou syntagme hyperonymique, ex. dans le *TLFi*, ACETAMIDE « amide primaire »), multiple (plusieurs hyperonymes au choix, ex. dans le *TLFi*, ACCUSER

« reprocher, imputer à qqn un défaut »), ou conjonctive (ex. dans le *TLFi*, RATATINÉ « tassé et émacié (par l'âge, la maladie, …) ») (Martin, 2001).

Les définitions logiques sont des définitions par inclusion, car le genre prochain est inclus dans la définition, étant ainsi une partie de cette définition. La difficulté de cette définition réside dans le fait de trouver le genre prochain, c'est-à-dire le concept le plus voisin du lexème à définir (Touratier, 2000).

Il convient de remarquer que les résultats des recherches que nous avons présentés dans la section précédente ne regroupent en fait que des définitions qui toutes relèvent de ce type de définition hyperonymique.

1.5.2. Définition synonymique

Ce type de définition est toujours constitué d'un synonyme définissant un lexème donné. Par exemple, la définition du *TLFi* pour l'adverbe *rapidement* est :

RAPIDEMENT. Adv. Prestement, vivement.

La définition synonymique n'est pas souvent pratiquée, en raison de l'absence dans la langue de deux lexèmes avec un sens parfaitement identique, c'est-à-dire des synonymes stricts, et, d'autre part, à cause du risque de circularité, car un lexème *A* est décrit par le lexème *B* qui peut être lui-même décrit par le lexème *A*.

1.5.3. Définition dérivative

La définition dérivative (translative) définit un lexème en faisant appel aux lexèmes dérivés de celui-ci, appartenant à la même famille morphologique. Nous présentons ci-dessous quelques exemples de définitions dérivatives du *TLFi* pour les lexèmes de différentes catégories grammaticales :

ACCÉS. Subst. masc. Action ou possibilité d'accéder.

ANALYSEUR. Subst. masc. et adj. Celui qui analyse.

DOMESTICITÉ. Subst. fém. État de l'animal qui a été domestiqué, apprivoisé par l'homme.

AGITABLE. Adj. Qui peut être agité.

AFFINER. V. Rendre plus fin.

Dans ce type de définition, le lexème dérivé et le lexème de définition ne sont pas de la même catégorie grammaticale (ex. *accès (nom)-accéder (verbe), affiner (verbe)-fin (adjectif),* etc.). Comme nous avons pu l'observer dans les exemples ci-dessus, les définitions dérivatives sont formées à l'aide d'opérateurs de translation qui varient d'une catégorie grammaticale à l'autre. Par exemple, pour définir un substantif dérivé, on utilise des constructions comme *action de, fait de, celui/celle/ce qui, état/caractère/qualité de ce qui,* etc. Toutefois, ces opérateurs peuvent aussi être rencontrés en dehors de définitions dérivatives, comme dans le cas ci-dessous :

PROGRÈS. Subst. masc. Fait de gagner du terrain.

Ainsi, dans le cas des définitions dérivatives, afin de préciser le sens d'un lexème donné, il convient de se rapporter à l'article du vocable concernant le lexème dérivé utilisé dans la définition (Touratier, 2000).

1.5.4. Définition méréonymique

Ce type de définition s'appuie sur une relation entre un lexème dénotant une partie et un autre dénotant le tout correspondant. Ci-dessous nous présentons un exemple de définition de ce type du *TLFi* :

DOSSIÈRE, subst. fém. Partie du harnais d'un cheval, posée sur le dos et servant à soutenir les brancards.

Les opérateurs méréonymiques peuvent être de plusieurs types, comme *partie de, portion de, morceau de, segment de, chacun de, ensemble de,* etc. Nous présentons ci-dessous quelques exemples de définitions utilisant les opérateurs méréonymiques :

BILLE, Subst. fém. Portion de tronc d'arbre débitée à la scie et non équarrie.

CALE, Subst. fém. Morceau de bois, de fer, etc., qu'on place sous ou contre un objet quelconque, afin de le mettre d'aplomb ou de l'immobiliser.

ÉNONCÉ, Subst. masc. Segment de la chaîne parlée produit par un seul locuteur et situé entre deux silences.

ROUAGE, Subst. masc. Chacun de ces éléments : roue dentée, pignon.

ÉCRITURE, Subst. fém. Ensemble des caractères d'un système de représentation graphique.

1.5.5. Définition approximative

Ce type de définition utilise des opérateurs d'approximation comme *sorte de, espèce de* afin de rapprocher un lexème d'un autre lexème, par exemple :

> TRANSE, Subst. fém. Sorte de sommeil pathologique ou d'altération de la conscience avec indifférence aux événements extérieurs et dont il est difficile de faire sortir le sujet.
>
> ACTÉON, Subst. masc. Espèce de papillon.

Par rapport à la définition hyperonymique, la définition approximative introduit plutôt une relation ontologique entre des objets (genre et espèce), tandis que la relation d'hyperonymie est une relation lexicale et concerne le langage. Ainsi, nous citons l'exemple donné par Gaudin et Guespin (2000) : « On pourra dire que marmonner est un hyponyme de parler, alors que cette relation n'est, habituellement, pas prise en compte si l'on raisonne en termes de genre et d'espèce. ».

1.6. Repérage des modes définitoires dans le *TLFi*

En vue d'un traitement automatique des définitions du *TLFi*, Martin (2001) présente les modalités de repérage des modes définitoires dans le *TLFi*. Selon lui, les opérateurs caractéristiques pour chaque type de définition déterminent finalement la classe à laquelle une définition appartient.

Nous décrivons ci-dessous les modalités de repérage de chaque type de définition présentées par Martin (2001, p. 160) :

1. Dans les définitions méréonymiques, il est important de déterminer l'holonyme du lexème défini. Dans tous les cas, il suit toujours l'opérateur méréonymique (ex. *partie de*). Par exemple, dans la définition « partie de corps humain qui … » du lexème TÊTE, le nom *corps* est un holonyme.
2. Les définitions dérivatives sont plus complexes. Les difficultés viennent du repérage des opérateurs translatifs (ex. *action de, état de ce qui, celui qui*, etc. pour les noms ; *devenir, rendre* + *adj.* pour les verbes, etc.) et surtout des lexèmes de la même famille morphologique que le défini. Pour automatiser le processus, une liste de préfixes et de suffixes peut être utilisée pour regrouper

dans une même famille les lexèmes avec le même suffixe ou préfixe. Toutefois, ce fait peut créer des erreurs à cause des radicaux homonymes, car la famille *rat -age / -er / -é* est différente de celle *rat -on / -ière / -ier*. Ainsi, pour réaliser un regroupement automatiquement, des instructions doivent être formulées pour chaque cas.

3. Les définitions par hyperonymie ou par synonymie sont traitées à la fin en déterminant avec exactitude l'hyperonyme (ou le synonyme qui en tient lieu), soit le syntagme hyperonymique. Les repérages sont différents en fonction de la catégorie grammaticale du lexème. Ainsi, dans les définitions nominales, l'hyperonyme représente d'habitude le premier nom de la définition, précédé dans certains cas par des syntagmes circonstanciels (ex. *dans un magasin, à Rome, au Moyen-Âge,* etc.). Les règles de détermination des hyperonymes dans les définitions verbales se différencient selon la valence verbale[21]. Les définitions adjectivales sont rarement hyperonymiques. Dans ces définitions, le premier adjectif représente plutôt un synonyme pour le lexème défini. Par exemple, le lexème RAUQUE est défini dans le *TLFi* comme « Rude, enroué,… », où les premiers adjectifs représentent des synonymes.

Martin (2001, p. 164) propose aussi des techniques de repérage des contenus spécifiques de définitions en distinguant certains types d'opérateurs, par exemple des opérateurs descriptifs (ex. *comporter, former, représenter, séparé par,* etc.), localisants (ex. *se trouver, entourer, au terme de,* etc.), fonctionnels (ex. *destiné à, servant à, utilisé pour,* etc.), relationnels (ex. *compte tenu de, en fonction de,* etc.). Une classification de tous les opérateurs (marqueurs) en fonction de l'information qu'ils signalent pour des définitions en *botanique* et en *chimie* a été effectuée par Hathout (1996). Ainsi, il a regroupé les marqueurs déterminés dans les définitions de *botanique* dans huit classes : classe de classification (ex. *variété, genre, classe,* etc.), d'odeur (ex. *odeur, odorant,* etc.), de goût (ex. *goût, saveur,* etc.), d'approximation (ex. *rassembler à, proche de, rappeler, etc.*), etc. Les marqueurs de définition de *chimie* ont été rassemblés en huit classes différentes : classe de classification (ex. *de la famille de, du groupe de,* etc.), de composition (ex. *muni de, constitué de, combinaison de,* etc.), d'origine (ex. *dériver de,*

[21] Selon Tesnière (1959), chaque verbe a une valence qui indique le nombre d'actants. Ainsi, il y a des verbes monovalents (ex. *être, tomber, éternuer,* etc.), bivalents (ex. *rassurer, frapper, brûler,* etc.) et trivalents (ex. *dire, donner,* etc.).

obtenu en, tiré de, etc.). Toutefois, afin de repérer tous les opérateurs, une analyse d'un grand nombre de définitions doit être effectuée.

Dans nos recherches nous ne sommes pas intéressée par une classification de définitions ou des opérateurs. Cependant, en réalisant des recherches assistées dans le *TLFi* nous avons vu que le premier nom de la définition n'est pas toujours un hyperonyme, il est souvent précédé par un opérateur comme *espèce*, *part, famille,* etc. qui tient du métalangage du dictionnaire. Ainsi, ces opérateurs peuvent être utilisés pour la détermination des hyperonymes dans les définitions. Nous présentons dans la suite de notre thèse (cf. §4.2.2.2.) une approche que nous avons mise en place pour le traitement de ces opérateurs.

1.7. Le *TLFi* et les projets connexes

L'apparition du *TLFi* sur Internet a été un vrai succès, plusieurs centaines de milliers de requêtes provenant du monde entier sont effectuées chaque jour. Il est utilisé dans le milieu de l'enseignement (les écoles, les lycées, les universités, etc.) avec le même engouement que dans le milieu de la recherche. Ainsi, le *TLFi* devient la base de plusieurs projets de recherche, par exemple le projet Definiens (Barque, Nasr & Polguère, 2010) ou le projet RELIEF[22] (Lux-Pogodalla & Polguère, 2011).

Quels sont donc les facteurs qui déterminent l'utilisation du *TLFi* dans ces projets ? Tout d'abord, le *TLFi* est une ressource normée en XML qui peut servir de support à d'autres projets de recherche. La forme assez normalisée et structurée du *TLFi* permet une extraction de connaissances, par exemple dans le domaine du traitement automatique des langues. L'interrogation du *TLFi* est assez simple et variée. Ainsi, il est possible de l'interroger soit par l'interface accessible sur Internet, soit en local par un langage de requêtes XML permettant l'extraction des différents fragments du *TLFi*, soit à distance via des web services, permettant l'intégration du *TLFi* comme ressource dans des applications distantes (Dedien & Pierrel, 2003).

[22] Ressource Lexicale Informatisée d'Envergure.

1.7.1. Projet SEMEME

SEMEME est une ressource lexicale construite à partir du *TLFi* dans le cadre du projet interne de l'ATILF DIXEME. L'intérêt de SEMEME est qu'il s'attache à décrire l'information de la microstructure du *TLFi*. Ainsi, les définitions du *TLFi* sont tout d'abord lemmatisées et filtrées pour ne garder que les lemmes à sémantisme plein (substantifs, verbes, adjectifs, adverbes)[23]. À chaque définition du *TLFi* sont aussi attribuées des statistiques d'occurrences de chaque lemme, son domaine d'emploi et des informations sur les contraintes structurelles liées à la lexie (lexème ou locution). Dans SEMEME, l'information concernant les lexèmes (vedettes) et les locutions (syntagmes) est représentée séparément (cf. figure 1.9.). Il y a aussi des statistiques sur les occurrences de tous les lemmes d'un article donné. Toutefois, contrairement à l'organisation de l'information dans un article du *TLFi*, SEMEME n'a pas conservé les informations relatives à l'organisation hiérarchique (marques de plan, marques de niveau hiérarchique), l'information lexicographique sur les indicateurs d'emploi, les synonymes et les antonymes. Dans la figure 1.9. nous présentons un extrait de l'information contenue dans SEMEME pour l'article du vocable AVOCAT.

```
<entree TLFID="10745">
    <vedette>
        <formes>
            <forme lemme="avocat" categorie="commonNoun"
                genre="masculine" origine="AVOCAT1, ATE, subst. /
                AVOCAT2 subst. masc." />
            <forme lemme="avocate" categorie="commonNoun"
                genre="feminine" origine="AVOCAT1, ATE, subst." />
        </formes>
        <definitions>
            <definition>
                <source>Personne défendant les intérêts de quelqu'un ou
                    de quelque chose .</source>
                <semes>
                    <seme lemme="défendre" categorie="v"/>
                    <seme lemme="intérêt" categorie="subst"/>
                </semes>
                <stats>
                    <stat lemme="défendre" occurences="1"/>
                    <stat lemme="intérêt" occurences="1"/>
```

[23] Il existe dans SEMEME encore deux autres catégories, Np qui indique majoritairement les noms propres et Ppa les éléments composés.

```
                    </stats>
                </definition>
            </definitions>
            ...
        <syntagmes>
            <syntagme>
                <forme>Avocat conseil .</forme>
                <definitions>
                <definition>
                    <source>,,Avocat ( ... ) attaché à une entreprise et chargé
                     de la défense de ses intérêts `` (Mét. 1955);</source>
                    <semes>
                        <seme lemme="avocat" categorie="subst"/>
                        <seme lemme="attacher" categorie="v"/>
                        <seme lemme="entreprise" categorie="subst"/>
                        <seme lemme="charger" categorie="v"/>
                        <seme lemme="défense" categorie="subst"/>
                        <seme lemme="intérêt" categorie="subst"/>
                    </semes>
                    <stats>
                        <stat lemme="avocat" occurences="1"/>
                        <stat lemme="attacher" occurences="1"/>
                        <stat lemme="entreprise" occurences="1"/>
                        <stat lemme="charger" occurences="1"/>
                        <stat lemme="défense" occurences="1"/>
                        <stat lemme="intérêt" occurences="1"/>
                    </stats>
                </definition>
                </definitions>
            </syntagme>
        </syntagmes>
        ...
        <stats>
            <stat lemme="avocat" occurences="9"/>
            <stat lemme="avoir" occurences="8"/>
            <stat lemme="plaider" occurences="6"/>
            <stat lemme="défendre" occurences="5"/>
            <stat lemme="être" occurences="5"/>
            ...
        </stats>
    </entree>
```

Figure 1.9. Extrait de l'article du vocable AVOCAT dans SEMEME

À partir de différentes informations présentes dans SEMEME, il est possible de synthétiser le schéma d'organisation général de l'information présenté figure 1.10.

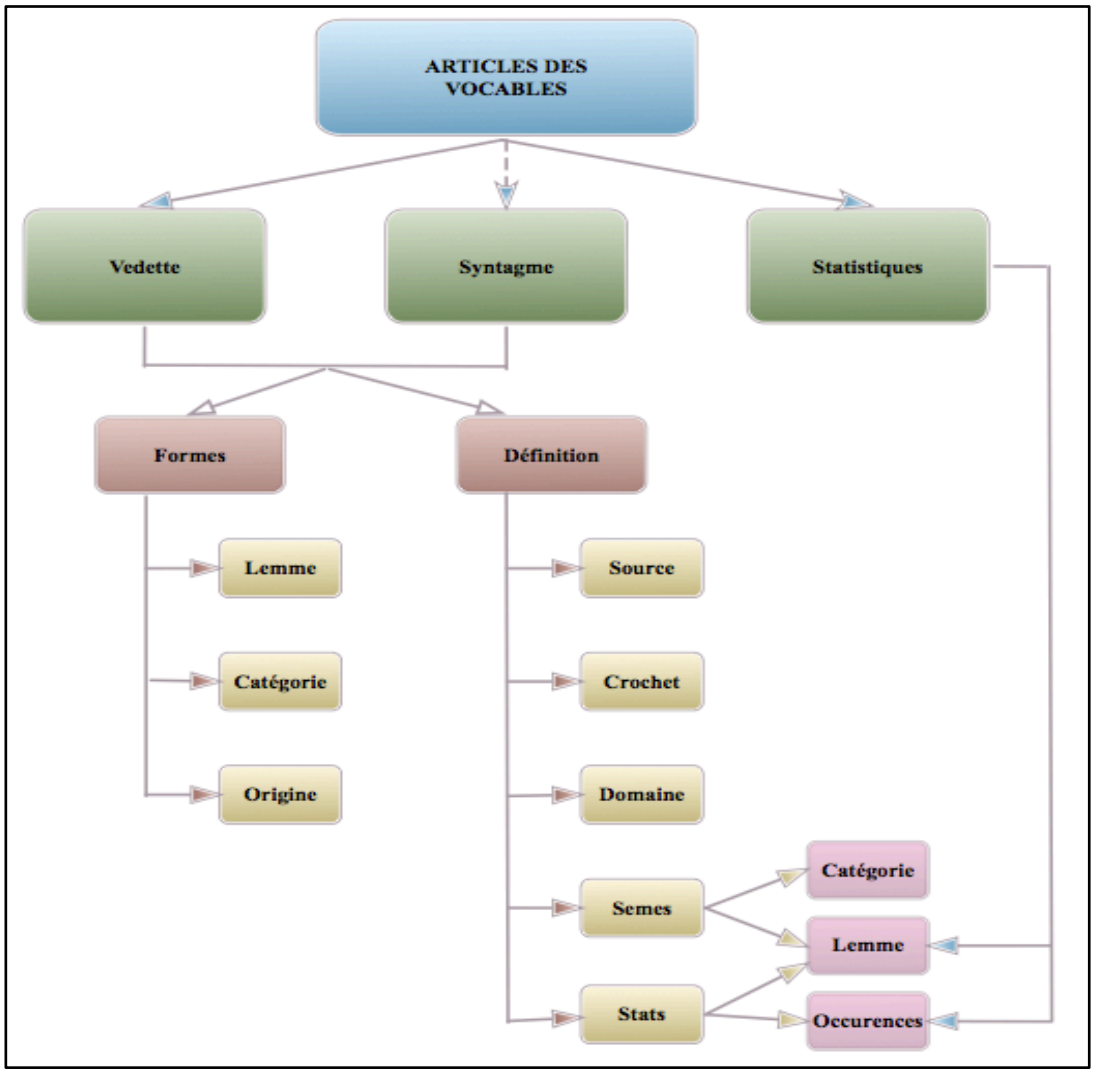


Figure 1.10. Schéma d'organisation de l'information dans SEMEME

Compte tenu du fait que dans SEMEME l'information sur les données de la microstructure du *TLFi* est plus explicitée que dans le modèle de données du *TLFi*, nous allons utiliser cette ressource dans nos recherches comme corpus de travail.

1.7.2. Projet Definiens et ses résultats

Le projet Definiens a été initié à l'Observatoire de linguistique Sens-Texte (OLST) de l'Université de Montréal par Alain Polguère et Lucie Barque. Ce projet a pour but la

structuration des définitions du *TLFi*, afin d'obtenir une nouvelle ressource lexicale de large couverture. Plus précisément, il s'agit de délimiter dans un premier temps dans les définitions du *TLFi* la composante centrale (CC), celle qui caractérise l'unité lexicale et qui représente le genre prochain, et les composantes périphériques (CP) qui correspondent aux différences spécifiques. Dans un deuxième temps, des étiquettes sémantiques[24] sont attribuées à chaque composante définitionnelle déterminée. À la balise <CC> est introduit l'attribut *étiq*, qui spécifie l'étiquette spécifique de la lexie définie et l'attribut *rôle* est introduit à la balise <CP> pour spécifier l'apport informationnel de chaque composante périphérique au contenu de la composante centrale. Pour le moment, c'est seulement la phase de balisage en CC et CP qui est entreprise. L'étiquetage sémantique des CC et l'attribution des rôles aux CP seront réalisés plus tard.

Prenons pour exemple une version structurée de la définition du *TLFi* pour le lexème REMORQUE présentée ci-dessous :

REMORQUE (acception B) =

<DEFI><PARAPH><CC étiq= véhicule>Bateau ou véhicule à roues</CC>
<CP rôle = mode_fonctionnement>dépourvu d'un moyen de propulsion propre</CP>
<CP rôle = fonction >et employé pour le transport des marchandises et/ou des voyageurs
</CP></PARAPH>.</DEFI>

Ainsi, dans cette définition balisée <DEFI> représentée par une paraphrase définitionnelle <PARAPH>, les balises <CC>/</CC> indiquent les composantes centrales et les balises <CP>/</CP> les composantes périphériques.

La détermination de la composante centrale d'une définition représente une tâche assez difficile à accomplir, car l'annotateur doit recourir à son intuition afin de savoir si la dite composante représente une composante classifiante, c'est-à-dire qu'elle est utilisée dans les définitions des autres lexies, ou si elle est ou non trop générale (Barque & Polguère, 2012). Par exemple, dans la définition structurée du lexème AC(C)ON (cf. fig. 1.10.), l'expression tout entière *embarcation à fond plat* a été annotée comme composante centrale et non seulement *embarcation*. La raison principale est le fait que

[24] Une expression linguistique normalisée qui rend compte de la valeur sémantique de la composante centrale (Barque & Polguère, 2009).

cette composante centrale est en même temps une composante classifiante pour 5 lexèmes différents du *TLFi* (AC(C)ON, ALLEGE, PINASSE, PLATE, PRAME).

AC(C)ON =

<DEFI><PARAPH><CC>Embarcation à fond plat</CC>,<CP>servant à divers usages</CP></PARAPH>.</DEFI>

PROSPECTEUR =

<DEFI><PARAPH><CC>Personne</CC>, <CP>qui est chargée par une entreprise de recherche de nouveaux clients</CP></PARAPH>.</DEFI>

ACCIDENT(acception II) =

<DEFI><PARAPH><CC>Événement fortuit</CC>,<CP>sans motif apparent</CP> et <CP>sans lendemain</CP>, <CP>qui affecte une personne ou un groupe de personnes, en interrompant le déroulement normal, probable et attendu des choses</CP> </PARAPH>.</DEFI>

ABSURDE =

<DEFI><PARAPH><CC>Qui est<CP>manifestement et immédiatement</CP> senti </CC><CP>comme contraire à la raison, au sens commun</CP></PARAPH>.</DEFI>

ALPHABET (acception A) =

<DEFI><PARAPH><CC>Ensemble de lettres</CC><CP>figurant les phonèmes d'une langue</CP>et<CP>disposées selon un ordre conventionnel</CP></PARAPH>.</DEFI>

IMAGE (acception B.1) =

<DEFI><PARAPH><CC>Représentation</CC><CP>de la forme ou de l'aspect d'un être ou d'une chose </CP>et<CP>disposées selon un ordre conventionnel</CP></PARAPH>.</DEFI>

a) = <DEFI><PARAPH><CC></CC><CP> par le dessin, la peinture</CP></PARAPH></DEFI>

b) = <DEFI><PARAPH><CC></CC><CP>par des procédés d'enregistrements photographiques </CP></PARAPH></DEFI>

c) = <DEFI><PARAPH><CC></CC><CP>par les arts plastiques</CP></PARAPH>. </DEFI>

CAMOUFLAGE =

<DEFI><PARAPH><CC>Action de camoufler</CC></PARAPH> ;

<PARAPH><CC>résultat de cette action</CC></PARAPH>.</DEFI>

Figure 1.11. Exemple des définitions structurées du *TLFi* dans le projet Definiens

Toutefois, dans la définition du lexème PROSPECTEUR, la composante centrale *personne* a été considérée comme suffisamment classifiante, tandis que la composante centrale *événement* du lexème ACCIDENT inclut aussi l'adjectif *fortuit*, car *événement* seul n'a pas un sens suffisamment spécifique. Les composantes centrales des définitions adjectivales incluent les séquences de type *qui est, qui a, que, dont* qui indiquent le fait qu'il s'agit d'une définition d'un adjectif et une composante périphérique CP enchâssée dans la composante centrale. Par exemple, la composante centrale de la définition du lexème ABSURDE est *qui est senti* et la composante périphérique est *manifestement et immédiatement*. De même, les séquences *sorte de, genre de, famille de, ensemble de,* etc. sont aussi incluses dans la CC. Il peut aussi arriver que, dans Definiens, certaines CC soient vides. C'est le cas lorsqu'une définition est construite dans le prolongement d'une autre (ex. lexème IMAGE, a) de la figure 1.10.). Dans certains cas, une CC peut contenir aussi un élément anaphorique renvoyant à une autre paraphrase (ex. lexème CAMOUFLAGE, cf. figure 1.11.).

Ce projet est toujours en cours de réalisation vu le grand travail d'analyse métalexicographique à effectuer. Le résultat de son achèvement sera une base de données sémantiques dérivées du *TLFi*, exploitable informatiquement, et constituant aussi une ressource pour la recherche en sémantique lexicale, TAL, etc.

Néanmoins, compte tenu de l'avancé de ce projet et fort de la validation linguistique des définitions structurées en CC et CP du projet Definiens, nous utiliserons dans nos recherches ces définitions structurées comme ressource de référence, pour valider certaines de nos approches proposées, en particulier celles permettant d'extraire automatique les informations les plus pertinentes au sein des définitions du *TLFi* (cf. chapitre 4).

1.7.3. Utilisation du *TLFi* pour l'enrichissement d'une ontologie

Eckard, Barque, Nasr et Sagot (2012) extraient des éléments des structures à partir du *TLFi* pour enrichir une ontologie existante WOLF[25] (Sagot & Fišer, 2012). Plus précisément, il s'agit d'attribuer à chaque élément du synset de l'ontologie la définition du *TLFi* dénotant son sens. Pour ce faire, dans un premier temps, ils ont mis en place un processus d'indentification automatique dans les définitions du *TLFi*, annotées préalablement à l'aide du logiciel MACAON (Nasr et al., 2011), des composantes centrales au moyen d'un ensemble de cinquante patrons lexico-syntaxiques constitués manuellement. Dans un deuxième temps, afin de désambiguïser chaque unité lexicale d'un couple hyperonyme-hyponyme de WOLF en lui assignant sa définition spécifique du *TLFi*, ils ont implémenté une heuristique (nommée *hypernymic ascent*) inspirée de celle proposée par Navigli (2009) qui consiste à lier deux unités lexicales avec une relation d'hyperonymie et de stocker les sens à travers lesquels la connexion se produit avec succès. À chaque étape, une définition est associée à un candidat hyperonyme, celui-ci étant généralement la tête du genre prochain de la définition du *TLFi* obtenu lors de l'étape précédente. Ainsi, les sens de chaque unité lexicale sont explorés de manière récursive à l'aide d'un algorithme de parcours en largeur[26] jusqu'à ce que le but est atteint. Par exemple, la figure 1.12. présente les sens par lesquels le couple *abordage-action* est lié.

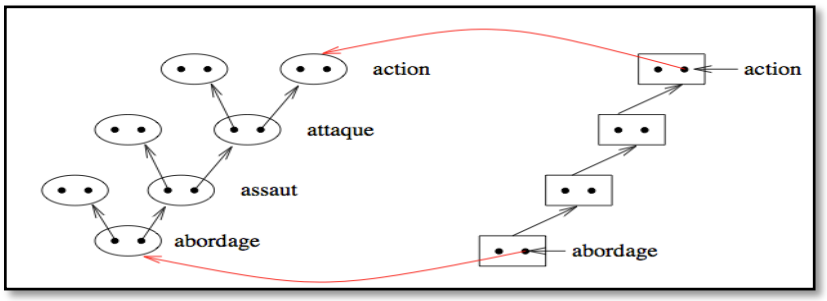


Figure 1.12. Exemple d'une structure ambigüe du *TLFi* (à gauche) et structure de WOLF (à droite) (Eckard, Barque, Nasr, & Sagot, 2012)

[25] WOLF (WOrdnet Libre du Français) est une ontologie pour le français construite automatiquement à partir de WordNet et de diverses ressources multilingues.
[26] L'algorithme de parcours en largeur (ou BFS, pour Breadth First Search) est un algorithme de recherche dans un graphe qui commence avec un nœud sommet et explore tous les nœuds voisins.

L'évaluation de la méthode a été effectuée manuellement sur cent couples hyperonymes-hyponymes en obtenant une précision de 45 %. Malgré les résultats obtenus, le travail réalisé montre la possibilité d'extraction automatique d'une structure ontologique à partir d'un dictionnaire — possibilité qui avait été rejetée dans les travaux d'Ide et Véronis (1993).

1.7.4. Projet RELIEF

Le projet RELIEF (Lux-Pogodalla & Polguère, 2011) est un projet de recherche et de développement, qui est réalisé dans le cadre d'une collaboration entre l'ATILF et une entreprise lorraine spécialisée dans la publication électronique MVS SA. Il s'agit d'un projet lexicographique qui vise la construction d'une nouvelle ressource lexicale du français à large couverture, appelée le *Réseau Lexical du Français (RLF),* en s'appuyant sur les acquis du *TLFi*. L'architecture de la nouvelle ressource est un système lexical (Polguère, 2009) — un grand graphe d'entités lexicales fortement connectées par des liens variés. La mise en place d'une telle architecture s'appuie sur le système des fonctions lexicales standard issues de la lexicologie explicative et combinatoire (Mel'cuk, Clas & Polguère, 1995). Par rapport à d'autres ressources lexicographiques, l'accès au RLF sera possible par champs sémantiques et non plus exclusivement par entrées. Les définitions utilisées pour le projet RLF sont semi-formalisées selon les mêmes principes que ceux appliqués dans le cadre du projet Definiens. Contrairement au *TLF*, le RLF se concentre seulement sur le français contemporain. Ce projet a commencé en 2011 et s'étend sur une période plus de trois ans.

1.8. Conclusion

Dans ce chapitre, nous avons présenté de manière approfondie le *TLFi*, la principale ressource lexicale utilisée dans nos recherches. Nous avons analysé la macro- et microstructure du dictionnaire ainsi que la typologie des définitions. Les divers accès possibles aux données du *TLFi* en s'appuyant sur sa structure nous ont permis d'émettre l'hypothèse que les noms situés en tête des définitions représentent des hyperonymes possibles de lexèmes définis. La relation d'hyperonymie/homonymie étant transitive[27]

[27] Si X est l'hyperonyme de Y et Y est l'hyperonyme de Z, alors X est aussi un hyperonyme de Z.

elle permet de hiérarchiser les lexèmes généralement en trois niveaux fondamentaux, du plus générique au plus spécifique. Étant donné que notre objectif est de construire des hiérarchies sémantiques à partir du *TLFi*, cette relation sémantique hiérarchique est fondamentale dans notre travail. En même temps l'analyse de différents projets qui utilisent ou qui ont utilisé le *TLFi*, nous a permis de choisir :

1. Comme corpus de travail, la ressource lexicale SEMEME construite à partir du *TLFi* car elle contient l'information formalisée de la microstructure du dictionnaire ;
2. Comme ressource de référence, les définitions structurées du projet Definiens. Elles nous permettront d'évaluer et valider certaines des nos approches proposées, en particulier celles permettant d'extraire automatique les informations les plus pertinentes au sein des définitions du *TLFi*.

Avant de présenter au chapitre 4 la méthode d'extraction automatique d'informations pertinentes au sein des définitions du *TLFi* que nous proposons pour construire ensuite automatiquement des hiérarchies sémantiques (cf. chapitre 5) et dont nous testerons la pertinence de leur exploitation pour l'indexation et la recherche d'images (cf. chapitre 6), nous allons rappeler dans le chapitre suivant les fondements de la recherche d'information et faire, dans le chapitre 3, un rapide état de l'art sur les ressources lexicales utilisées pour l'indexation et la recherche d'images. Plus particulièrement, compte tenu des orientations actuelles de Xilopix, nous nous situerons dans le cas de grandes bases de données images décrites uniquement par une suite restreinte de mots-clés, contexte qui rend inopérantes la plupart des méthodes fondées sur les techniques de recherche d'informations textuelles classiques qui nécessitent un contexte textuel beaucoup plus large qu'une suite restreinte de mots-clés.

CHAPITRE 2

Fondements de la recherche d'information et spécificités de la recherche d'images

Sommaire

2.1. Architecture d'un SRI 54
2.1.1. Phase d'indexation 54
2.1.2. Phase de recherche et aperçu de principaux modèles classiques de RI .. 58
2.2. Les systèmes d'indexation et de recherche d'images 60
2.2.1. Indexation textuelle manuelle d'images 61
2.2.2. Indexation textuelle automatique d'images 63
2.2.3. Recherche textuelle d'images 65
2.2.4. Indexation et recherche d'images par le contenu 68
2.3. Évaluation des SRI 71
2.3.1. Mesures de la qualité de l'indexation 71
2.3.2. Mesures de la qualité de la recherche 73
2.3.3. Problèmes spécifiques à l'évaluation des systèmes de recherche d'images 75
2.4. Conclusion 77

Chapitre 2. Fondements de la recherche d'information et spécificités de la recherche d'images

Un système d'indexation et de recherche d'information a pour but de permettre à ses utilisateurs de retrouver, à partir d'une collection de documents, les documents qui correspondent à leurs besoins d'information exprimés d'habitude sous forme de requêtes textuelles. Pour qu'un document soit retrouvé, il doit être tout d'abord indexé. Le processus d'indexation consiste à décrire un document par un ensemble de mots-clés ou de descripteurs en vue de représenter son contenu. La liste des mots-clés retenue par ce processus d'indexation constitue l'index du document.

Pendant longtemps, les seuls contenus indexés étaient ceux de documents textuels. En effet, les livres ou les documents papier sont indexés manuellement (classement par auteur ou par mots-clés par exemple), de sorte que le bibliothécaire ou le documentaliste puisse retrouver facilement et rapidement un ouvrage.

Le fait d'indexer pour faciliter la recherche de documents représente actuellement le principe fondamental des systèmes de recherche d'information (SRI), quelle que soit la nature du document à indexer : texte, image, vidéo, audio, etc.

Compte tenu des finalités de cette thèse qui consiste à étudier l'exploitation possible d'un dictionnaire de langue, le *TLFi,* pour améliorer l'indexation et la recherche d'images au sein du moteur de Xilopix, ce chapitre vise à décrire les spécificités de systèmes de recherche d'images existants, et à fournir des liens avec les approches utilisées en recherche d'information (RI) textuelle classique. Nous commençons par un aperçu de principaux fondements de la recherche d'information et ensuite nous décrivons les bases des systèmes d'indexation et de recherche d'images. En particulier, nous détaillons les techniques développées, les problèmes liés à chaque type d'indexation et de recherche et leurs limitations. Enfin, une dernière section présente les principales mesures utilisées pour l'évaluation des SRI ainsi que les problèmes spécifiques à l'évaluation des systèmes de recherche d'images.

2.1. Architecture d'un SRI

Le but fondamental d'un SRI est de trouver les documents qui répondent au mieux aux besoins d'information des utilisateurs. Un SRI repose donc sur deux phases qui forment le processus clé du système : l'indexation et la recherche.

- L'indexation est une étape qui consiste à analyser les documents, à extraire et à modéliser leur contenu textuel sous forme d'index qui soient ensuite exploitables par le SRI. Pour notre part nous souhaitons proposer des outils exploitant le *TLFi* pour améliorer l'indexation d'images à partir de la liste des mots-clés attachés à chaque image ;
- La recherche correspond à la seconde étape qui permet d'effectuer une comparaison entre la base d'index et la représentation de la requête de l'utilisateur en utilisant une mesure de correspondance (fonction d'appariement) du modèle de RI. Le résultat présenté à l'utilisateur en réponse à sa requête est un ensemble de documents dont les termes d'indexation sont les plus proches de ceux de sa requête. Dans ce cadre nous souhaitons pour notre part déterminer comment les informations extraites du *TLFi*, en particulier les hiérarchies sémantiques dont nous proposerons la construction au chapitre 5, pourraient permettre d'enrichir le thésaurus utilisé dans cette phase et d'améliorer la recherche elle-même.

Certaines SRI contiennent éventuellement une phase de « bouclage de pertinence » (Rocchio, 1971) qui prend en compte les jugements des utilisateurs sur les documents trouvés afin d'améliorer les résultats de la recherche. L'architecture globale d'un SRI est présentée dans la figure 2.1.

Dans les sections suivantes, nous expliquons les deux phases fondamentales qui permettent de mettre en œuvre le processus de recherche d'information.

2.1.1. Phase d'indexation

La tâche principale du processus d'indexation est de transformer le contenu des documents en un ensemble d'index ou de descripteurs capables de représenter leur contenu. Le problème à résoudre lors de ce processus est donc, tout d'abord, de choisir les termes à extraire les plus représentatifs du contenu du document. Ainsi, à la base

d'un système d'indexation textuelle résident deux processus : celui du choix des termes d'indexation et celui de la pondération des descripteurs.

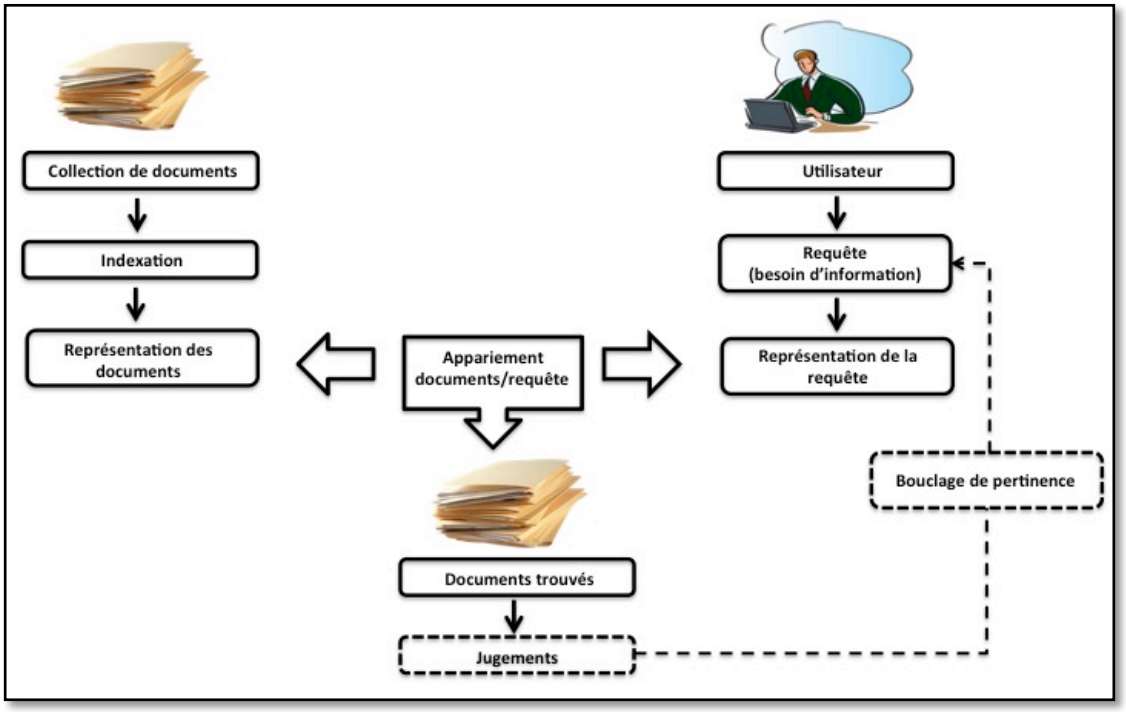


Figure 2.1. Schéma d'un système classique de recherche d'information

Le choix des termes d'indexation contribue à la structuration de l'espace dans lequel les documents seront représentés. Ce choix se réalise en plusieurs étapes comme suit :

1. La première étape est la *tokenisation*, qui correspond au processus de transformation du texte en chaîne de mots-formes[28] ou *tokens*. Le problème principal de ce processus est le fait de décider quels sont les délimiteurs des mots-formes. D'habitude, les mots-formes sont séparés par des espaces ou des signes de ponctuation, mais certains cas posent des problèmes comme dans l'exemple suivant illustré par Claveau, Moreau et Sébillot (2007, p.165) : « c'est-à-dire qu'aujourd'hui, les pommes de terre des U.-S.-A. sont cultivées in vitro ». Faut-il considérer *pommes de terre* comme un mot-forme ou comme un groupement de trois mots-formes ?

[28] Un mot-forme au sens lexicographique peut représenter soit une instance d'un lexème (« pommes » est une instance d'un lexème POMME), soit une instance d'un élément d'une locution (« pommes » est une instance de l'élément *pommes* dans la locution POMMES DE TERRE).

2. La deuxième étape est la *lemmatisation*. Ce processus regroupe tous les mots-formes d'un lexème afin qu'ils puissent ensuite être analysés comme une seule entité appelée *lemme*. Ainsi, le lemme correspondant à tous les mots-formes *fais, faisons, font,* etc. est *faire.* Dans certains cas, cette étape de lemmatisation peut conduire à des ambiguïtés d'interprétation, comme dans le cas de la phrase suivante : *la belle ferme le voile* où par exemple le mot-forme *ferme* peut être lemmatisé en *ferme, substantif féminin, singulier* ou *ferme, verbe indicatif présent du verbe fermer, 3^{e} personne du singulier*.
3. La troisième étape consiste à éliminer, par l'emploi de listes de *mots vides*, les mots grammaticaux qui participent peu à l'élaboration du contenu sémantique du document (auxiliaires, articles, prépositions, etc.).
4. La dernière étape cherche à déterminer des termes d'indexation parmi les lemmes. Différentes techniques statistiques sont utilisées en s'appuyant sur les fréquences des mots-formes dans le document et dans les collections afin de déterminer les termes ayant un pouvoir discriminant. Salton, Yang et Yu (1975) considèrent comme les meilleurs termes d'indexation pour une collection de *n* documents les termes ayant une fréquence d'apparition comprise entre *n/100* et *n/10*.

Le processus de *pondération des descripteurs* permet de déterminer quel terme est plus important que l'autre pour décrire un document. À chaque terme est attribué un poids dans le document où il apparaît, qui est calculé automatiquement à partir de trois critères : l'importance du terme dans le document, l'importance du terme dans la collection des documents et la taille du document. Conformément aux facteurs de pondération de Salton et Buckley (1988), ces trois critères correspondent à deux facteurs de pondération, les pondérations locales et globales, et un facteur de normalisation.

1. Pondération locale.

La pondération locale mesure l'importance d'un terme dans un document. Elle prend en compte uniquement les informations concernant le document donné. Cette pondération, notée *TF (term frequency),* est la fonction de la fréquence des occurrences d'un terme dans un document. Plus un terme est fréquent dans un document, plus il est

considéré comme pertinent pour décrire celui-ci. Soit le document *d* et le terme *t*, la fréquence *TF(t, d)* du terme *t* dans le document *d* est alors souvent utilisée directement ou exprimée selon la formule suivante :

$$TF = \frac{freq(t,d)}{\sum_i freq(t_i,d)} \tag{2.1.}$$

où

$freq(t,d)$ est la fréquence d'un terme *t* dans le document *d,*

$\sum_i freq(t_i,d)$ est la somme des fréquences de tous les termes dans le document.

2. Pondération globale.

La pondération globale, contrairement à la pondération locale qui a tendance à favoriser les termes très présents, et par suite le rappel (Salton & Buckley, 1988), privilégie plutôt les termes qui apparaissent dans peu de documents pour avoir une bonne précision. La mesure utilisée pour la pondération globale est la fréquence documentaire inverse, notée *IDF (inverse document frequency)*. Cette mesure détermine l'importance d'un terme *t* dans un ensemble de documents D_t de la collection dans lesquels il apparaît. Elle consiste à calculer le logarithme de l'inverse de la proportion de documents du corpus qui contiennent le terme. Cette mesure est exprimée de la façon suivante :

$$IDF = log\frac{|N|}{n} \tag{2.2.}$$

où

n est le nombre de documents contenant le terme *t,*

N est le nombre total de documents dans la collection.

Ainsi la fonction de pondération de la forme TF-IDF, la plus souvent utilisée, consiste à multiplier ces deux mesures de la manière suivante :

$$TF - IDF = TF(t,d) * IDF(t,D) \tag{2.3.}$$

La combinaison de ces deux mesures permet de trouver les termes qui sont en même temps fréquents dans le document et très discriminants.

3. Normalisation.

La normalisation permet d'obtenir des poids de termes peu sensibles à la taille des documents, afin de ne pas favoriser les documents longs. En effet, un document court contiendra peu de mots-formes, et par la suite peu de termes de poids importants. À l'opposé, un document long contiendra des occurrences plus nombreuses d'un terme et sera privilégié dans le cas où une requête porte sur l'un de ces termes. Plusieurs schémas de normalisation sont utilisés, nous présentons ci-dessous la normalisation par cosinus de Salton et Buckley (1988) :

$$W_i = \frac{l_t * g_t}{\sqrt{\sum_{t=1}^{n}(l_t * g_t)^2}} \qquad (2.4.)$$

où

W_i indique la pondération associé au terme *t* dans le document *i,*

$l_t$et g_t représentent respectivement la pondération locale et globale du *t* terme.

Nous allons revenir sur les notions de pondération dans le chapitre 4.

2.1.2. Phase de recherche et aperçu de principaux modèles classiques de RI

La recherche de documents est un processus d'interaction entre l'utilisateur et le SRI illustré dans la figure 2.1. Elle permet d'associer à une requête un ensemble de documents pertinents à restituer. Ce processus implique les étapes suivantes :

1. L'acquisition par le système de l'expression du besoin d'information de l'utilisateur sous forme d'une requête ;
2. L'interprétation de la requête selon le langage d'indexation défini ;
3. L'évaluation par le système, au travers d'une fonction d'appariement, de la pertinence des documents par rapport à cette requête. Les scores de pertinence entre la requête *Q* et les descripteurs du document *D* sont calculés selon une formule de similarité appelée *Retrieval Status Value*, notée *RSV (Q, D).*

Comme résultat de la recherche, le système renvoie à l'utilisateur une liste de documents généralement triés par ordre de pertinence de valeur du score. L'utilisateur

choisit à son tour parmi les documents renvoyés ceux qui correspondent le mieux à son besoin d'information et au contexte dans lequel la recherche a été effectuée.

Pour améliorer les résultats de la recherche, certains systèmes prennent en considération le point de vue de l'utilisateur en lui proposant d'indiquer des exemples de documents pertinents et non pertinents pour sa requête, afin de générer une nouvelle requête et poursuivre la recherche (Rocchio, 1971 ; Lv & Zhai, 2009 ; Xu, Jones, & Wang, 2009). Dans ce cas, le processus de recherche représente une boucle fermée où les résultats antérieurs sont pris en compte lors d'une nouvelle recherche.

Pendant l'indexation, comme nous l'avons vu en section §2.1.1., le système sélectionne un ensemble de termes utilisés pour constituer une représentation du contenu du document ou de la requête. L'organisation de ces termes en une représentation dépend du modèle de RI utilisé. Ainsi, dans différents modèles, le même ensemble de termes pourra avoir une signification différente. Le modèle tient un rôle important dans un SRI. Il permet de donner une représentation interne à un document ou à une requête basée sur ces termes. Le modèle RI définit aussi la méthode de comparaison entre le document et la requête de l'utilisateur afin de déterminer leur degré de correspondance (ou similarité). Les modèles fondamentaux de RI sont :

1. *Modèle booléen.*

Le modèle booléen (Salton G., 1969) est basé sur la théorie des ensembles. Dans ce modèle un document ou une requête sont représentés comme une conjonction logique de termes non pondérés qui constituent l'index.

Ce type de modèle pose plusieurs difficultés. Tout d'abord, compte tenu de la représentation binaire {1,0} de la correspondance entre un document et une requête et du non-ordonnancement des réponses obtenues, il n'est, par exemple, pas possible de dire quel document correspond mieux qu'un autre à la requête. De plus, dans un document ou une requête, les termes sont pondérés de la même façon {1,0} et il est donc difficile de juger si un terme est plus important qu'un autre dans une représentation.

2. *Modèle vectoriel.*

Le modèle vectoriel introduit par Salton (1989) est un modèle utilisé dans de très nombreux SRI. Ce modèle représente un document ainsi qu'une requête par un vecteur

de poids, où chaque poids désigne l'importance d'un terme correspondant dans ce document ou dans la requête.

Le modèle vectoriel repose sur la théorie d'indépendance des termes, ce qui rend impossible la représentation des phrases, des mots composés ou des termes multi-mots. Enfin, un tel modèle vectoriel ne permet pas la prise en compte de certains phénomènes linguistiques telle la synonymie des mots — phénomène dont on sait qu'il est fortement contextuel et lié à l'ambiguïté lexicale. Même si le modèle vectoriel est critiqué (Raghavan & Wong, 1986), il demeure néanmoins un des modèles les plus utilisés.

3. Modèle probabiliste.

Le modèle probabiliste utilise un modèle mathématique fondé sur la théorie des probabilités (Maron & Kuhns, 1960 ; Robertson & Spärck Jones, 1976 ; Salton & McGill, 1983). Le principe du modèle probabiliste consiste à retourner pour chaque requête soumise par l'utilisateur une liste de documents qui ont à la fois une forte probabilité d'être pertinents, et une faible probabilité d'être non pertinents. Le modèle probabiliste cherche à estimer la probabilité qu'un document *d* soit pertinent pour la requête *q* de l'utilisateur. Même s'il est plus efficace qu'un modèle booléen, l'un des inconvénients de ces modèles est qu'il est impossible d'estimer les probabilités utilisées pour évaluer la pertinence des résultats si de vastes corpus d'entraînement ne sont pas disponibles.

2.2. Les systèmes d'indexation et de recherche d'images

La documentation a été le premier domaine à proposer des méthodes pour la recherche de documents. Dans les sections précédentes, nous avons décrit les principes de fonctionnement d'un SRI textuel. Ces méthodes constituent également le fondement de la recherche d'information multimédia-image. La recherche multimédia-image, domaine de spécialité de Xilopix, est donc un domaine apparenté à la recherche textuelle, mais s'en différencie sur un nombre de points importants. Dans les sections suivantes nous avons choisi de nous focaliser sur le cas d'indexation et de recherche d'images en montrant les types possibles d'accès à ces collections, car notre recherche s'insère dans le cadre du projet Xilopix d'élaboration de moteur de recherche d'images.

2.2.1. Indexation textuelle manuelle d'images

L'indexation textuelle manuelle d'images est effectuée par une personne qui attribue à chaque image un ensemble de mots-clés, souvent en utilisant un vocabulaire contrôlé, afin d'indexer les images. Ce type d'indexation a été utilisé bien avant que les images puissent être numérisées, en particulier par les bibliothécaires, les conservateurs, et les archivistes. Actuellement, elle est souvent utilisée pour l'indexation des bases de données d'images personnalisées, par exemple les albums photo digitalisés tels que Picasa[29] ou Flickr[30].

Toutefois, la représentation textuelle des images pose des problèmes, car une image peut transmettre des informations sur ce qui est représenté dans l'image ainsi que sur ce que l'image signifie. Par exemple, une image peut représenter *un couple devant le prêtre dans une église*, mais porter sur *le mariage*. Cette problématique a été étudiée par Panofsky (1955) qui propose une approche d'analyse des niveaux iconographiques de sens dans des images. Les travaux de Panofsky ont servi de source d'inspiration pour Shatford (1986), qui les appliquait au processus d'indexation des images. Shatford a classé les sujets des images en trois types : *générique (generic of)*, *spécifique (specific of)* et *à propos (about)*. Au niveau générique, des objets généraux et des actions sont décrits. Par exemple, une femme, une voiture. Le niveau spécifique décrit les objets nommés individuellement et les événements, tels que la *Tour Eiffel* ou la *chute du mur de Berlin*. Le dernier niveau correspond aux émotions, aux abstractions et aux symboles, comme le *bonheur*, la *justice* ou le *rideau de fer*. De plus, à chaque niveau sont ajoutées quatre facettes : l'objet (qui), les activités ou les événements (quoi), l'endroit (où) et le temps (quand). Ainsi, le modèle Panofsky/Shatford est devenu un modèle très répandu pour la classification des descriptions d'images et a été utilisé par plusieurs chercheurs. Mais ce type d'indexation souffre souvent du faible accord entre les indexeurs sur le choix des termes ainsi qu'entre les indexeurs et les requêtes des utilisateurs (Enser & McGregor, 1993).

L'indexation manuelle d'images est en effet une tâche liée à la subjectivité de la perception humaine qui s'appuie sur le savoir et l'érudition. C'est pour cette raison que le W3C (World-Wide Web Consortium) met à disposition des vocabulaires spécifiques

[29] http://picasa.google.com/
[30] http://www.flickr.com/

pour décrire les propriétés et le contenu d'une image à indexer. Par exemple, les propriétés d'une image comme la résolution, la date de capture peuvent être décrites en utilisant les termes du vocabulaire VRA Core (*Visual Resource Association Core Categories)*[31]. Pour la description du contenu de l'image, en fonction du domaine auquel elle appartient, une variété de vocabulaires et d'ontologies est proposée.

Toutefois, l'utilisation des vocabulaires contrôlés ne permet pas de décrire tous les types d'images, car la plupart d'entre eux ne sont pas assez riches pour fournir de l'information descriptive adéquate. En outre, un autre inconvénient de l'indexation à l'aide de vocabulaires contrôlés est que les termes d'indexation proposés par les vocabulaires contrôlés ne coïncident pas toujours avec les termes utilisés par les utilisateurs lors de la formulation de leurs requêtes de recherche. L'étude de Furnas, Landauert, Gomez et Dumais (1987) sur le problème de vocabulaires montre qu'un nouvel utilisateur du système ne trouve du premier coup les informations qui l'intéressent que dans seulement 10 % à 20 % des cas. Ce phénomène est essentiellement dû à la différence des vocabulaires utilisés : un même terme peut être représenté par son synonyme ou son hyperonyme.

Comme alternative à l'utilisation des vocabulaires contrôlés s'offre l'indexation collaborative, dont la principale caractéristique est l'utilisation d'un vocabulaire libre, qui contient les mots courants de la langue (Macgregor & McCulloch, 2006). L'indexation collaborative est utilisée par plusieurs services de partage d'images sur le Web (Flickr en est un exemple) où les utilisateurs attribuent eux-mêmes des mots-clés (ou tags) aux images. Les mots-clés sont ensuite utilisés pour des recherches ciblées, ce qui peut augmenter le rappel, car le plus souvent l'utilisateur constitue la requête de recherche en faisant aussi appel au vocabulaire libre. Cependant, l'inconvénient majeur de ce type d'indexation est que les indexeurs ne sont pas toujours conscients de l'importance que constitue l'attribution des termes d'indexation et par conséquent les images sont souvent indexées par des mots-clés qui ne décrivent pas véritablement leur contenu. La vérification des mots-clés des images est faite par les autres membres de la communauté.

[31] Un exemple d'annotation d'image en utilisant VRA Core : http://www.vraweb.org/projects/vracore3/examples.html

Même si l'indexation manuelle d'images est une méthode efficace, elle se révèle coûteuse en termes d'argent et de temps. Ainsi, elle ne peut pas s'appliquer aux grandes bases de données, qu'elles soient générales ou spécialisées, telles les grandes bases médicales ou astronomiques.

2.2.2. Indexation textuelle automatique d'images

L'indexation textuelle automatique d'images est un processus d'attribution des informations aux images, sans aucune intervention humaine. L'indexation textuelle automatique peut se réaliser en exploitant soit le texte associé à l'image, soit le contenu visuel de l'image.

La première approche s'effectue d'habitude à partir de données textuelles associées aux images par les annotateurs : noms, légendes, mots-clés, rubriques ou textes en langage naturel (Chen & Rasmussen, 1999). Prenons par exemple une image de *lion* de la base Flickr qui a été annotée par son auteur (cf. figure 2.2.).

Figure 2.2. Image de *lion* dans Flickr

Cette image a été annotée seulement avec une courte description *le roi des animaux* et avec une liste de mots-clés comme *Afrique du Sud, tourisme, safari, animaux sauvages,* etc. Pour l'indexer, la plupart des systèmes réalisent une analyse lexicale (tokenisation, lemmatisation) de la description textuelle afin de choisir les termes d'indexation *roi*, *animal* par lesquels l'image sera ensuite indexée. Ainsi, l'image est indexée par tous les mots-clés de sa description textuelle. Dans Flickr, tous les mots-clés de l'image sont considérés à parts égales comme des termes d'indexation. Le

désavantage de cette technique d'indexation est qu'aucune analyse sémantique de la description textuelle de l'image n'est effectuée, ce qui, par conséquent, ne permet pas, dans cet exemple, l'indexation de l'image par le terme *lion*[32]. Pour résoudre ce type de problèmes, des ressources linguistiques peuvent être utilisées. Nous étudierons leur apport dans le chapitre 3 de cette thèse.

Sur le Web, les images peuvent aussi être indexées à partir du titre de la page, de l'URL ou du texte proche de l'image. Cette technique d'indexation est utilisée par les différents moteurs de recherche (Google, Yahoo, Bing, etc.). La difficulté qui se pose ici est la sélection des mots-clés correspondants à l'image à partir des textes qui l'entourent. Il est évident que tous les mots-clés d'une page web n'ont pas tous de liaison avec l'image de la même page. De plus, deux images d'une même page web ne devraient pas toujours être indexées avec les mêmes mots-clés. La détection des mots-clés qui ne sont pas liés au contenu de l'image introduit alors beaucoup de bruit lors de la recherche d'images. Par exemple, un utilisateur qui cherche une image de *lion* aura, parmi les réponses de la recherche, une image de *chat*, tout simplement parce que cette image a été indexée avec le mot-clé *lion* qui se trouvait dans le texte entourant l'image. Par ailleurs, de nombreux travaux ont tenté d'exploiter la relation entre les textes et les images de la même page web pour identifier les images similaires sur le Web. Dans ce but, Coelho, Calado, Souza, Ribeiro-Neto et Muntz (2004) ont combiné les différentes sources de descriptions textuelles associées aux images, en obtenant de meilleures performances qu'en utilisant une seule source de descriptions textuelles. Wang, Ma, Xue et Li (2004) considéraient les images du Web et les textes environnants comme deux types d'objets différents. En construisant la structure des liens et en exploitant le renforcement mutuel entre eux, ils ont combiné la description textuelle et le contenu visuel pour déterminer les images web connexes ou similaires.

En l'absence d'information textuelle, la deuxième approche s'effectue en exploitant le contenu visuel de l'image. Ce sont des méthodes d'apprentissage supervisé. Elles utilisent un ensemble de données d'apprentissage, à savoir des images déjà annotées avec des mots-clés afin de prédire des mots-clés d'une nouvelle image. Ainsi, les mots-clés peuvent être associés à l'image dans son ensemble *: coucher de soleil* (annotation globale), ou aux différentes régions d'image : *soleil, mer, personne* (annotation

[32] C'est ce qui se produit quand le mot-forme *lion* est absent de la liste de mots-clés descriptifs de l'image.

globale), une segmentation préalable de l'image en régions étant nécessaire. Plusieurs travaux ont été proposés dans ce sens (Qi & Han, 2007 ; Wong & Leung, 2008). L'inconvénient majeur de cette approche est qu'elle nécessite des bases d'images entièrement annotées, qui sont très difficiles à obtenir, car elles requièrent un travail coûteux d'annotation manuelle de la base. Pour obtenir des résultats satisfaisants, ces méthodes peuvent être appliquées sur les bases de données de petite taille avec un vocabulaire réduit.

Dans notre thèse, nous adopterons la première approche qui consiste à indexer les images en exploitant leurs descriptions textuelles. Toutefois, comme nous l'avons déjà mentionné, l'inconvénient de cette approche est qu'en général aucune analyse sémantique de la description textuelle d'images n'est réalisée et par conséquent les images sont indexées par tous les mots-clés associés. Afin d'éviter l'indexation d'images par tous les mots-clés nous allons utiliser des hiérarchies sémantiques construites à partir du *TLFi* pour déterminer les termes d'indexation d'images ainsi que leurs domaines d'utilisation.

2.2.3. Recherche textuelle d'images

Aujourd'hui, les moteurs de recherche d'images les plus utilisés (Google, Picsearch, etc.) sont basés sur une recherche par mots-clés. Ce procédé d'accès aux données photographiques fonctionne ainsi : l'utilisateur formule une requête composée de termes langagiers et le système lui propose en réponse des images qu'il considère comme proches des termes de sa requête. Pour ce faire, le système, le plus souvent à l'aide de calculs statistiques, détermine la similarité entre les termes de la requête et les termes avec lesquels les images ont été indexées. Les résultats de la recherche sont affichés sous forme d'une mosaïque d'images. Par rapport, aux systèmes de recherche textuelle, les utilisateurs peuvent juger de la pertinence des résultats immédiatement, sans accéder à la page source de l'image trouvée.

Toutefois, les images proposées à l'utilisateur ne semblent pas toujours correspondre à leur requête initiale, car les systèmes ne réalisent en fait aucune analyse du contenu de la requête pour déterminer sa signification. L'ambiguïté des mots-clés, les phénomènes de synonymie et d'hyperonymie sont les problèmes principaux de la recherche d'images par mots-clés. Pour donner un exemple, interrogeons Google Images avec la requête

avocat. Comme résultat de la recherche (cf. figure 2.3.), le système affiche des images qui représentent l'*avocat* comme un *fruit* et comme un *métier*. Vu l'ambiguïté du mot-clé *avocat*, Google Images propose en même temps certaines suggestions par l'ajout des nouveaux termes au mot-clé de la requête initiale (cf. figure 2.3. : *avocat* métier, *avocat* fruit). Cela permet d'affiner la recherche et de trouver les images pertinentes. Cependant, les suggestions proposées sont basées sur des recherches populaires. C'est pourquoi, parmi les suggestions proposées pour la requête *lion,* on ne trouve pas de suggestions de type *lion* monnaie, *lion* constellation, etc. (cf. figure 2.4.).

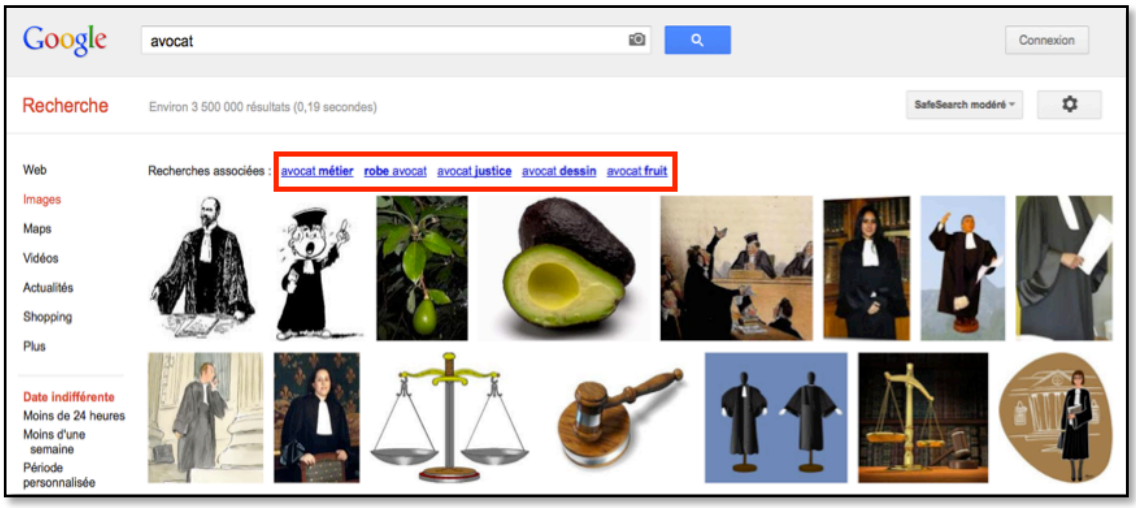


Figure 2.3. Résultats de la recherche pour la requête *avocat* dans Google Images (octobre 2012)

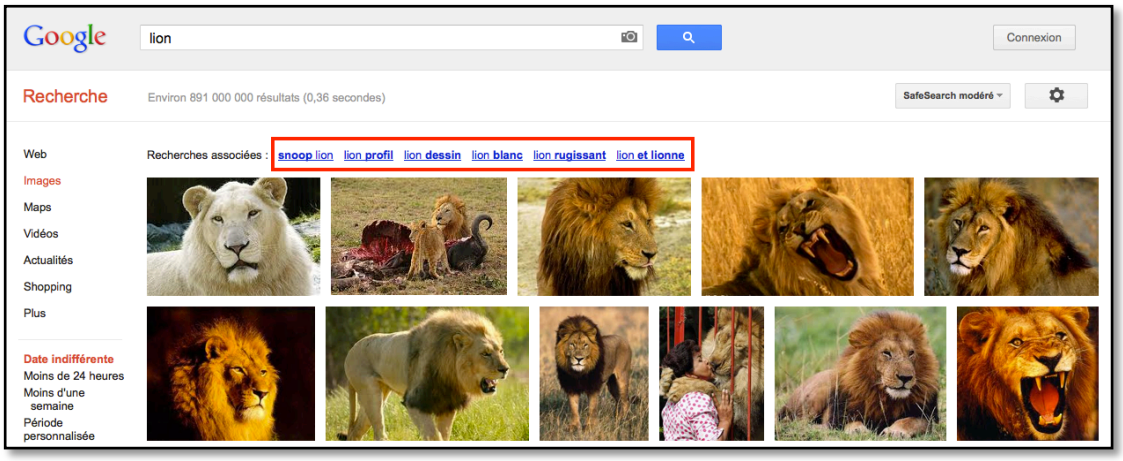


Figure 2.4. Résultats de la recherche pour la requête *lion* dans Google Images (octobre 2012)

Si l'on interroge Getty Images avec la même requête *lion,* le système nous propose de préciser la recherche en proposant deux suggestions : *lion* (grand félin) et *signe du lion* (signe de feu) (cf. figure 2.5.). Les suggestions proposées sont faites sur base du thésaurus et, même si elles sont meilleures que celles proposées par Google Images, elles ne sont toutefois pas complètes et ne représentent pas tous les sens du mot-clé *lion*.

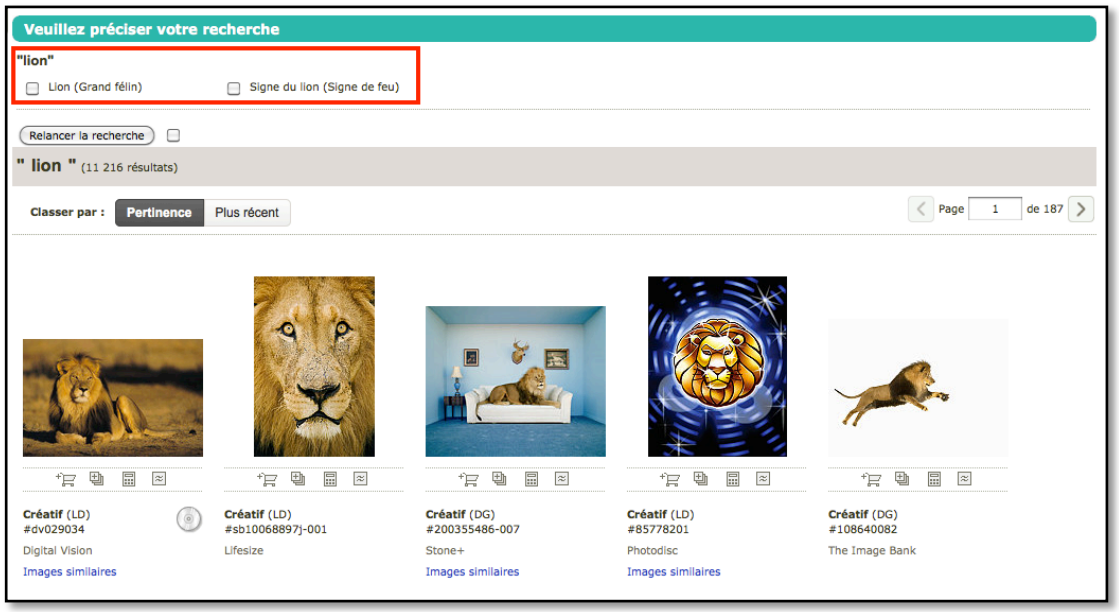


Figure 2.5. Résultats de la recherche pour la requête *lion* dans Getty Images (octobre 2012)

Un autre problème de ces moteurs de recherche est le fait qu'ils ne donnent pas immédiatement un aperçu de la diversité d'images trouvées. Ainsi, dans Google Images, l'utilisateur intéressé par les images du *lion* comme *signe astrologique* ne satisfera sa requête qu'à partir de la page 8 (cf. figure 2.6.)[33]. Il doit donc soit parcourir huit pages d'images avant de trouver celles qui l'intéressent, ce qui est fastidieux, soit reformuler sa requête, ce qui n'est pas évident quand l'utilisateur n'a pas une vision claire de ce qu'il recherche.

[33] Entre temps, Googles Images (pour l'anglais) vient de mettre en place une nouvelle technique d'affichage des images.

Figure 2.6. Exemple des images trouvées dans Google Images pour le concept *lion* (octobre 2012)

Certains systèmes de recherche textuelle d'images utilisent un bouclage de pertinence (Zhang, Chai & Jin, 2005 ; Chai, Zhang & Jin, 2007). Toutefois, les résultats obtenus ne sont pas satisfaisants car les mots-clés utilisés par les utilisateurs ne sont pas toujours pertinents.

À vrai dire, pour résoudre les problèmes liés à la recherche d'images par mots-clés et rendre à l'utilisateur des résultats répondant mieux à sa requête, il conviendrait d'utiliser des ressources sémantiques[34] ; c'est ce que nous allons présenter dans la suite de cette thèse (cf. chapitre 3). En ce que concerne notre recherche, nous proposons d'exploiter dans le processus de recherche d'images des hiérarchies sémantiques construites automatiquement à partir du *TLFi,* cela afin de proposer une nouvelle structuration des images trouvées en exploitant les domaines des définitions (au sens des domaines du *TLFi*) auquel les mots-clés de la requête peuvent faire référence (cf. chapitre 6).

2.2.4. Indexation et recherche d'images par le contenu

Une image peut être représentée non seulement via un langage à base de symboles, mais aussi par des descripteurs de bas niveau comme la couleur, la texture, la forme.

L'indexation d'images par le contenu consiste à extraire automatiquement les caractéristiques physiques de l'image en utilisant des méthodes d'analyse d'images. Les caractéristiques visuelles extraites d'une image sont décrites par un vecteur numérique nommé *signature* de l'image. Deux images sont considérées similaires si leurs

[34] Les ressources sémantiques peuvent être représentées sous forme de bases de connaissances structurées comme les taxonomies, les thésaurus, les ontologies.

signatures sont similaires. L'indexation d'images par le contenu visuel peut être effectuée en utilisant des informations globales ou locales. Ainsi, les descripteurs visuels peuvent être extraits soit pour différentes régions de l'image, soit pour l'image dans son ensemble. Pour utiliser l'information locale, un découpage préalable de l'image en régions est alors nécessaire. Les descripteurs sont définis pour capturer une certaine propriété visuelle de l'image. Les descripteurs visuels les plus utilisés correspondent à la couleur (Solli & Lenz, 2010), la texture (Murala, Gonde, & Maheshwari, 2009), la forme (Gevers & Smeulders, 2000 ; Kekre, et al., 2010).

La recherche d'images par le contenu visuel s'effectue d'habitude via une image requête. L'utilisateur est invité à formuler sa requête à partir d'images exemples, de couleurs sélectionnées dans un spectre, de modèles de texture ou à partir de son propre croquis dessiné. Ce type de recherche est très orienté système et son usage s'avère être difficile pour un simple utilisateur. La difficulté à utiliser un tel système vient du fait que l'utilisateur n'est pas habitué à exprimer ses besoins d'information en paramètres visuels. Par exemple, si un utilisateur est intéressé par les images de la mer, se pose la question suivante : « Quelles couleurs ou quelles textures doit-il choisir pour formuler sa requête ? » Nous avons essayé de répondre à cette question en cherchant les images de la mer dans la galerie digitale du musée de l'Ermitage[35], qui utilise le système QBIC (Flickner, et al., 1995). L'interface propose deux types de recherche par la couleur et la texture. Pour trouver les images qui représentent la mer, nous avons exprimé notre requête en choisissant trois nuances de la couleur bleue dans le spectre proposé, comme le montre la figure 2.7.a). D'une part, les résultats proposés par le système ne correspondaient pas tout à fait à notre besoin d'information, car seulement une image représentait la mer (image numéro 2). Pourtant, les images proposées étaient bien en correspondance avec notre requête : la couleur bleue était prédominante dans toutes les images.

[35] http://www.hermitagemuseum.org/fcgi-bin/db2www/qbicSearch.mac/qbic?selLang=English

a) Formulation de la requête b) Les résultats de la recherche

Figure 2.7. Exemple d'utilisation du système QBIC

La non pertinence des résultats de recherche aux besoins d'information de l'utilisateur provient du fait que, lors de la recherche, les caractéristiques de la requête sont utilisées pour définir les descripteurs globaux des images et non de certaines régions. Bien que l'analyse de l'image s'effectue en s'appuyant sur les caractéristiques des régions comme dans le cas du système Netra (Ma & Manjunath, 1999), il y a un grand risque de surreprésentation de régions qui ne présentent aucun intérêt pour la recherche. Afin de faciliter la recherche d'images par le contenu, certains systèmes comme IKONA (Boujemaa, et al., 2001) ont permis à l'utilisateur de formuler sa requête sans indiquer aucune image ou région exemple. Ce système propose la représentation des images par un ensemble de termes visuels correspondants à des catégories. Les catégories représentent des regroupements de régions similaires de la base et appartiennent à un thésaurus photométrique. Ainsi, dans le scénario de recherche, l'utilisateur doit choisir dans le thésaurus des catégories descriptives des images.

La recherche d'images par le contenu symbolique ne peut pas satisfaire complètement les besoins des utilisateurs. La cause principale en est le *fossé sémantique* entre les descripteurs de bas niveau et les descripteurs de haut niveau (Wang, Zhang & Zhang, 2008 ; Enser, Sandom, Lewis & Hare, 2006). Ainsi, dans la figure 2.7.b), les quatre images ont des histogrammes de couleurs très similaires, mais elles sont sémantiquement très différentes. Plusieurs méthodes ont été proposées afin de réduire ce fossé sémantique comme l'utilisation des ontologies visuelles (Mezaris, Kompatsiaris & Strintzis, 2003), des boucles de pertinence (Ferecatu, Crucianu &

Boujemaa, 2005) ou l'utilisation à la fois de la description textuelle associée à l'image et du contenu visuel (Wang, Ma & Li, 2004 ; Ferecatu, Boujemaa & Crucianu, 2008). Un état de l'art de toutes ces méthodes est représenté dans l'article de Liu, Zhang, Lu et Ma (2007).

Si l'indexation textuelle d'images vise à modéliser le contenu sémantique de l'image, l'indexation par le contenu d'images vise à décrire les caractéristiques de l'image. Ainsi, pour un utilisateur, il est plus naturel d'exprimer sa requête en mots-clés (ex. *mer, montagne, soleil,* etc.) que par les termes de valeurs symboliques (ex. *zone bleue, zone jaune*). C'est pour cette raison que les systèmes de recherche d'images par le contenu symbolique ne sont utilisés que dans les domaines spécifiques (la médecine, l'astronomie, le design, la publicité) et ne sont pas orientés vers tous les types d'utilisateurs. Dans nos recherches, comme nous utilisons seulement des bases d'images généralistes (images personnelles, du web, etc.), nous n'avons pas retenu ces approches dans notre étude.

2.3. Évaluation des SRI

L'évaluation est un procédé essentiel pour comprendre si un SRI fonctionne ou non de manière efficace. Un SRI est jugé performant s'il propose des réponses pertinentes à la requête de recherche de l'utilisateur, tout en étant rapide. Généralement, l'évaluation d'un SRI se concentre plutôt sur la pertinence des résultats, car un moteur de recherche rapide n'est d'aucune utilité s'il ne donne pas de bons résultats. Dans les sous-sections suivantes, nous allons détailler les mesures de performance pour chaque phase d'un SRI.

2.3.1. Mesures de la qualité de l'indexation

Les résultats de la recherche dépendent en grande partie de la qualité de l'indexation. Un document ne sera pas retrouvé s'il n'a pas été indexé. Toutefois, plusieurs travaux effectuent une évaluation globale des systèmes de recherche par rapport à la pertinence des résultats, ce que ne permet pas une évaluation de la qualité de l'indexation. La nécessité de mesurer spécifiquement la qualité de l'indexation a été soulevée dans les travaux de Soergel (1994), Kefi, Berrut et Gaussier (2005). Selon Kefi, Berrut et Gaussier (2005), la qualité de l'indexation peut être représentée par *l'exactitude* et *la*

consistance. Ainsi, les mesures de la qualité de l'indexation, en lien avec les points de vue document ou terme (cf. tableau 2.1.), doivent permettre de répondre aux questions suivantes :

- les documents sont-ils indexés correctement ?
- les termes d'indexation sont-ils attribués correctement aux documents ?

Exactitude L'exactitude de l'indexation est liée à la présence ou à l'absence de termes pertinents d'indexation. Ainsi, un document peut être indexé par un terme incorrect (erreur d'excédent) ou peut ne pas être indexé par un terme pertinent (erreur d'omission). Dans le tableau 2.1. nous présentons les mesures qui permettent de caractériser l'exactitude de l'indexation.

	Mesures de l'exactitude	
	Point de vue document D	Point de vue terme t
Complétude	$\frac{\text{nb } t \text{ correctement affectés à } D}{\text{nb } t \text{ qui devraient être affectés à } D}$	$\frac{\text{nb } D \text{ correctements indexés par } t}{\text{nb } D \text{ qui devraient être indexés par } t}$
Pureté	$\frac{\text{nb } t \text{ correctement rejetés pour } D}{\text{nb } t \text{ qui devraient être rejetés pour } D}$	$\frac{\text{nb } D \text{ correctements non indexés par } t}{\text{nb } D \text{ qui ne devraient pas être indexés par } t}$
Justesse	$\frac{\text{nb } t \text{ correctement affectés à } D}{\text{nb total de } t \text{ affectés à } D}$	$\frac{\text{nb } D \text{ correctement indexés par } t}{\text{nb total de } D \text{ indexés par } t}$

Tableau 2.1. Formules de calcul des mesures de l'exactitude de l'indexation

Complétude La complétude de l'indexation du point de vue document permet de savoir si tous les documents qui devraient être indexés par le terme ont réellement été indexés. Du point de vue terme, la complétude répond à la question : « Parmi les termes qui devraient être affectés au document, combien ont été réellement affectés ? »

Pureté La pureté de l'indexation est liée à l'absence d'erreurs d'excédent et d'omission. Elle peut aussi être considérée par rapport à un document ou à un terme d'indexation.

Justesse La justesse de l'indexation permet de savoir, du point de vue document, si tous les termes d'indexation associés à un document sont corrects. Du

point de vue terme, elle détermine combien de documents indexés par le terme sont corrects.

Consistance Dans l'indexation des documents, il est très important qu'un accord soit fait entre les indexeurs sur le choix des termes d'indexation. Ainsi, la consistance de l'indexation permet d'assurer que les documents similaires sont indexés par les mêmes termes d'indexation indifféremment des annotateurs. Elle peut être mesurée en comparant les réponses entre deux indexeurs différents (consistance inter-indexeurs), ou entre deux indexations différentes du même document effectuées par le même indexeur lors des sessions différentes (consistance intra-indexeur).

	Mesure de la consistance	
	Point de vue document *D*	Point de vue terme *t*
Consistance inter-indexeurs	$\frac{\text{nb } t \text{ affectés à } D \text{ par les indexeurs } A \text{ et } B}{\text{nb } t \text{ affectés à } D \text{ par les indexeurs } A \text{ ou } B}$	$\frac{\text{nb } D \text{ indexés par } t \text{ par les indexeurs } A \text{ et } B}{\text{nb D indexés par } t \text{ par les indexeurs } A \text{ ou } B}$

Tableau 2.2. Formules de calcul des mesures de la consistance de l'indexation

Les mesures de la qualité de l'indexation aident à améliorer le processus d'indexation, ce qui assure par conséquent la performance de la recherche.

Une synthèse de mesures existantes est présentée par Kefi, Berrut et Gaussier (2005).

2.3.2. Mesures de la qualité de la recherche

Les mesures classiques utilisées pour évaluer la qualité de la recherche sont le rappel et la précision. Pour mieux comprendre comment sont calculées ces deux mesures, nous nous appuyons sur le scénario suivant : un système de recherche d'information pour la requête *ville de Nancy* a retourné 120 documents. Sur 150 documents qui traitaient *Nancy* comme *ville,* seulement 90 ont été trouvés, mais les 30 autres documents portaient sur *Nancy Sinatra.*

Rappel Le rappel est la proportion de documents pertinents trouvés par rapport au nombre total des documents pertinents. Le rappel consiste à se poser la question :

« Parmi tous les documents pertinents existants combien ont été retrouvés? » Le rappel *R* d'un SRI est calculé selon la formule suivante :

$$R = \frac{nb\ D\ pertinents\ trouvés}{nb\ total\ de\ D\ pertinents} \quad (2.5.)$$

Dans le cas de notre scénario, le rappel est de 60 % (90/150). Toutefois, 40 % de documents pertinents n'ont pas été retrouvés. Dans ce cas, on parle de *silence (1-R)*.

Précision La précision est la proportion de documents pertinents trouvés par rapport au nombre total de documents trouvés. Elle permet de répondre à la question suivante : « Parmi tous les documents retrouvés combien sont pertinents ? » La précision *P* d'un SRI est calculée selon la formule suivante :

$$P = \frac{nb\ D\ pertinents\ trouvés}{nb\ total\ de\ D\ trouvés} \quad (2.6.)$$

Dans ce cas précis, la précision est de 75 % (90/120). Toutefois, 25 % de documents non pertinents ont été trouvés. Dans ce cas, on parle de *bruit (1-P)*.

Un système de recherche idéal devrait avoir un rappel et une précision égale à 1, ce qui signifie qu'il trouve tous les documents pertinents existants sans aucune erreur. En pratique, il est très difficile d'avoir un tel système, car les deux mesures ne sont pas indépendantes. Ainsi, d'une part, pour avoir un système avec un rappel de 100 % il suffit de trouver tous les documents existants, mais dans ce cas, c'est la précision qui va être très faible et, d'une autre part, pour avoir un système avec une précision de 100 % il suffit de trouver un nombre minimal de documents pertinents, mais dans ce cas, c'est le rappel qui en pâtit. Dans les deux cas, le système de recherche ne peut pas satisfaire tous les besoins d'information de l'utilisateur[36] : un compromis entre ces deux mesures est donc nécessaire.

F-mesure Ainsi, pour trouver un meilleur compromis entre le rappel et la précision, une *F-mesure* a été proposée, permettant de pondérer de la même façon les deux mesures. Elle est définie comme une moyenne pondérée harmonique du rappel et de la précision, qui est :

[36] En fonction des buts suivis, l'une des deux mesures peut être plus importante que l'autre.

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} \qquad (2.7.)$$

où P est la précision, R est le rappel et le paramètre α détermine la pondération de la précision et du rappel. Si α=1, alors la précision et le rappel ont la même importance.

Courbe rappel-précision Les courbes rappel-précision sont utilisées pour analyser le comportement d'un système de recherche. Pour les obtenir, les valeurs de précision sont calculées pour une valeur de rappel donné. Généralement, une courbe rappel-précision est représentée sur onze points de rappel, de 0 % à 100 % avec un pas de 10 %. Un exemple de courbe rappel-précision est présenté dans la figure 2.8.

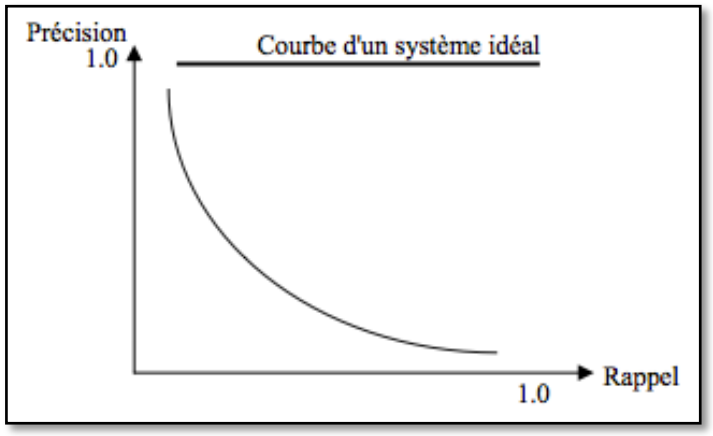


Figure 2.8. Exemple de courbe rappel-précision

Parce que les moteurs de recherche ne sont pas parfaits et récupèrent presque toujours des documents non pertinents, la précision tend à diminuer avec l'augmentation du rappel.

Une étude d'un ensemble de mesures de performance utilisées pour l'évaluation des SRI peut être trouvée dans l'article de Baccini, Dejean, Kompaore et Mothe (2010).

2.3.3. Problèmes spécifiques à l'évaluation des systèmes de recherche d'images

Les systèmes de recherche d'images peuvent être évalués en utilisant les mêmes mesures de performance que pour les SRI. Toutefois, l'évaluation des systèmes de recherche d'images est plus complexe que celle de SRI, à cause de la subjectivité de

perception des images : une image peut avoir une signification différente pour chaque personne. Ainsi, dans le cas des systèmes de recherche d'images, l'évaluation de la pertinence des résultats n'est pas toujours suffisante. En lien avec l'application, nous pouvons évaluer la qualité esthétique des images (Marchesotti et al., 2012) pour ensuite choisir les images de la meilleure qualité ou évaluer aussi les termes d'indexation de l'image en fonction du niveau qu'ils caractérisent (générique, spécifique, etc.). Certains travaux (Cardaci, Gesu, Petrou & Tabacchi, 2006 ; Perreira Da Silva & Courboulay, 2011) s'intéressent aussi à l'évaluation de la complexité visuelle de l'image qui peut être utile pour l'étude psychologique des phénomènes de perception ainsi que pour les applications informatiques qui relèvent de la compression de l'image ou de la théorie de l'information. Néanmoins, l'évaluation des systèmes de recherche d'images est souvent plus rapide que celle de SRI, car la pertinence des résultats peut être jugée immédiatement.

Afin d'évaluer les performances des systèmes d'indexation et de recherche, plusieurs campagnes d'évaluation sont organisées comme MediaEval[37], TREC[38]. L'une des campagnes d'évaluation de l'annotation et de la recherche d'images est ImageCLEF[39]. Pendant cette campagne plusieurs tâches d'évaluation sont proposées comme l'annotation automatique d'images avec les concepts, la recherche d'images en combinant les caractéristiques textuelles et visuelles, la recherche multilingue d'images, etc. La campagne a débuté en 2003 et se tient chaque année. Pendant la campagne un corpus d'essai (benchmark) est fourni qui permet aux chercheurs ou industriels de mesurer les performances de leur système pour le comparer à d'autres. Dans ce type des campagnes d'évaluation, les corpus de référence contiennent d'habitude les images avec leurs descriptions textuelles en anglais.

Pour ce qui nous concerne, nous n'avons pas pu au cours de notre thèse nous confronté à une telle campagne d'évaluation.

De plus dans nos recherches nous tenons à bien distinguer les phases d'indexation et de recherche d'images dans les processus d'évaluation afin de montrer les apports de hiérarchies sémantiques construites à partir du *TLFi*. Pour ce faire nous allons utiliser les mesures classiques que nous avons présentées dans les sections précédentes (cf.

[37] http://www.multimediaeval.org/
[38] http://trec.nist.gov/
[39] http://www.imageclef.org

§2.3.1., §2.3.2.). En complément à ces mesures classiques, nous présentons diverses évaluations réalisées dans l'objectif de montrer l'apport de nos propositions.

Nos évaluations se présentent comme suite :

- Évaluation automatique des procédures de sélection des hyperonymes candidats par l'exploitation des définitions du *TLFi*. Cette évaluation est en fait double (cf. chapitre 4) : (i) comparaisons de notre mesure de pondération des noms dans les définitions du *TLFi* par comparaison avec la mesure classique TF-IDF ; (ii) comparaison de nos résultats avec les travaux issus du projet Definiens (Barque, Polguère, & Nasr, 2010) (Barque & Polguère, 2009) (cf. §1.7.2.) dont les résultats correspondent à une structuration des définitions du *TLFi* en composantes centrales (CC) et composantes périphériques (CP) ;
- Évaluation externe par des documentalistes de Xilopix des liaisons hyperonymiques déterminées automatiquement (cf. §5.3.) ;
- Évaluation des hiérarchies sémantiques déterminées automatiquement en comparaison avec celles du thésaurus Xilopix (cf. §5.4.) ;
- Évaluation des performances du prototype simple d'indexation en utilisant les hiérarchies sémantiques construites à partir du *TLFi* (cf. §6.1.) ;
- Évaluation des résultats du prototype simple de recherche en exploitant ces hiérarchies sémantiques et le thésaurus de domaines du *TLFi* (cf. §6.3.).

2.4. Conclusion

Dans nos recherches, nous nous basons sur l'exploitation d'informations textuelles associées aux images. C'est pour cela que dans ce chapitre nous avons décrit aussi les techniques de la RI textuelle, qui peuvent être ensuite adaptées pour les systèmes de recherche textuelle d'images. Nous avons également donné un aperçu des systèmes de recherche d'images par le contenu, afin de montrer les difficultés liées à leur utilisation. Bien que les moteurs de recherche d'images par mots-clés soient beaucoup plus répandus, ils rencontrent à leur tour des problèmes liés à l'ambiguïté et la polysémie des mots-clés.

Un autre problème est lié au fait que la plupart des systèmes de recherche utilisent des modèles classiques de représentation des documents et des requêtes. Ils se fondent sur l'hypothèse de *l'indépendance des mots* où les mots (lexèmes) sont représentés

comme dépourvus de sens (c'est ce qui correspond à la notion de « sac des mots »). Même si cette hypothèse facilite beaucoup les calculs, elle ne représente pas la réalité : les lexèmes sont en réalité reliés entre eux par des relations sémantiques. Ainsi, ce type de modèle ne permet pas aux SRI de saisir des phénomènes sémantiques fortement contextuels tels la synonymie et l'hyperonymie. Par conséquent, d'un côté, les images sont indexées par tous les mots-clés de leurs descriptions textuelles associés et d'un autre côté, lors de la recherche sont trouvées toutes les images qui contiennent les termes de la requête sans prendre en compte la signification des mots-clés. Nous avons aussi remarqué qu'une des faiblesses des moteurs de recherche d'images actuels est qu'ils ne donnent pas une visibilité immédiate sur la diversité d'images trouvées. Nous pensons qu'une meilleure exploitation des domaines de définitions des termes d'indexation devrait pouvoir remédier à cette faiblesse et améliorer notablement la présentation des résultats de la recherche.

Dans la suite de ce manuscrit, nos recherches vont donc s'orienter vers l'indexation et la recherche d'images à partir de données textuelles spécifiques : les mots-clés associés aux images, d'une part, et les définitions lexicographiques des lexèmes issues du *TLFi*, d'une autre part. Par ailleurs, d'un côté, nous tenons à proposer des approches permettant de réaliser une indexation automatique d'images à partir les métadonnées textuelles associées en déterminant les descripteurs pertinents et d'un autre côté, nous sommes intéressée par l'enrichissement automatique du thésaurus Xilopix. Nos travaux ont donc pour objectif de montrer que l'utilisation d'un dictionnaire de langue, tel le *TLFi*, peut contribuer à ces deux points au travers de processus entièrement automatique. Afin de valider le processus d'indexation et de montrer l'intérêt d'exploitation de connaissances dictionnairiques, nous allons mettre en place un prototype de moteur de recherche simple.

Ainsi, nous commençons le chapitre suivant, en montrant comment les ressources lexicales peuvent venir en aide aux systèmes d'indexation et à la recherche textuelle d'images.

CHAPITRE 3

Construction de ressources lexicales et leur utilisation pour l'indexation et la recherche d'images

Sommaire

3.1. Lexiques informatiques 81
3.1.1. WordNet 81
3.1.2. ConceptNet 86
3.1.3. DBpedia 87
3.2. Thésaurus 88
3.2.1. Thésaurus — outil d'indexation et de recherche 89
3.2.2. Les relations dans le thésaurus 92
3.2.3. Les modes de représentation des thésaurus 93
3.3. Ontologies 96
3.3.1. Structures et types d'ontologies 97
3.3.2. Utilisation d'ontologies dans le domaine d'indexation et de recherche d'images 100
3.4. Construction automatique de hiérarchies sémantiques 102
3.4.1. Les études de construction de hiérarchies sémabtiques à partir de textes 102
3.4.2. Les études de construction de hiérarchies sémantique à partir de définitions lexicographiques 105
3.4.3. Questions posées par l'extraction de connaissances à partir de dictionnaires. 108
3.5. Conclusion 113

Chapitre 3. Construction de ressources lexicales et leur utilisation pour l'indexation et la recherche d'images

Le présent chapitre a pour objectif de décrire les principales ressources lexicales utilisées pour l'indexation et la recherche d'images ainsi que les techniques de construction de nouvelles ressources susceptibles d'être utilisées dans ce domaine. Nous commençons par la description des principaux types de ressources lexicales exploitables dans le domaine de l'indexation et de la recherche d'images. Par la suite, nous nous intéressons aux techniques de construction automatique de hiérarchies sémantiques à partir de textes et plus particulièrement à partir de dictionnaires. Nous concluons ce chapitre en présentant les principales questions posées par l'extraction de connaissances à partir de dictionnaires.

3.1. Lexiques informatiques

Les lexiques informatiques peuvent être envisagés comme des versions informatisées des ressources sémantiques. Ainsi, ils peuvent regrouper des taxonomies, des glossaires, des bases de données terminologiques et des ontologies (Litkowski, 2005). À partir des années 1980, plusieurs lexiques informatiques ont été développés de manière manuelle ou automatique. Parmi les lexiques informatiques constitués manuellement, citons WordNet (Miller, 1995 ; Fellbaum, 1998), Cyc (Guha & Lenat, 1990) et ceux constitués automatiquement comme MindNet (Dolan, Vanderwende, & Richardson, 2000), YAGO (Suchanek, Kasneci & Weikum, 2007 ; Suchanek, Kasneci & Weikum, 2008), DBpedia (Auer, Bizer, Kobilarov, Lehmann & Ives, 2007).

Dans les sections suivantes, nous allons présenter les lexiques informatiques qui sont les plus utilisés dans le domaine de la recherche d'images.

3.1.1. WordNet

WordNet (Miller, 1995 ; Fellbaum, 1998) est une base de données lexicales conçue manuellement par les linguistes et les lexicographes pour une utilisation dans des

applications informatiques. Dans WordNet, les catégories grammaticales comme *nom*, *verbe*, *adjectif*, *adverbe* sont regroupées en ensemble de synonymes cognitifs (synsets), chacun exprimant un concept distinct[40]. La version anglaise courante de WordNet 3.0. contient 117 798 des noms regroupés dans 82 115 synsets. Les synsets sont reliés entre eux par les relations sémantiques suivantes :

- *Synonymie* : c'est la relation fondamentale, car dans WordNet les synsets sont utilisés pour exprimer le sens des lexèmes par un ensemble de synonymes ;
- *Antonymie* : ce type de relation est spécialement utilisé pour l'organisation des adjectifs et des adverbes, car les pairs des antonymes comme *young-old* (jeune-vieux) reflètent le contraste sémantique fort entre les lexèmes ;
- *Hyperonymie/Hyponymie* : ce type de relation permet l'organisation hiérarchique des noms, par exemple *bed-furniture* (lit-mobilier) ;
- *Méronymie* : c'est une relation de partie à tout, elle signifie qu'un lexème est une sous-partie d'un autre lexème, par exemple *chair-backrest* (fauteuil-dossier) ;
- *Troponymie* : c'est une relation entre deux verbes converses qui permet d'exprimer d'une manière plus spécifique l'action de l'autre verbe, par exemple *buy-pay* (acheter-payer).

La figure 3.1. présente un exemple des différents sens du lexème LION dans WordNet[41].

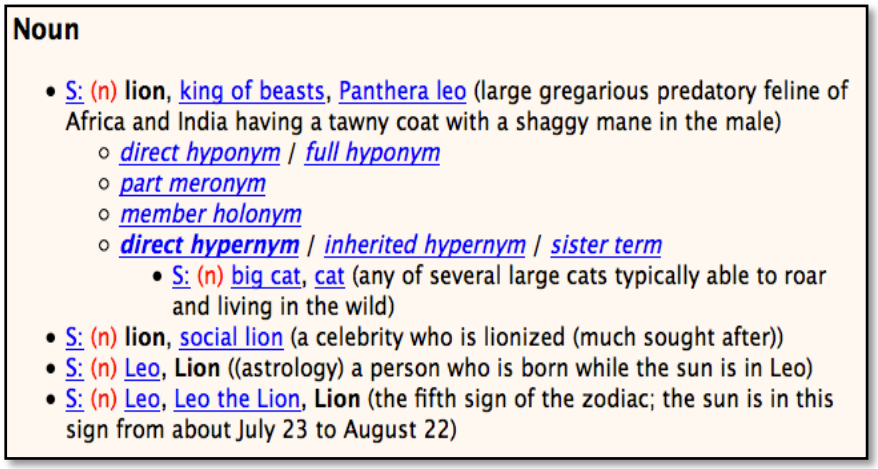


Figure 3.1. Exemple des différents sens du LION dans le WordNet anglais

[40] http://wordnet.princeton.edu/
[41] http://wordnetweb.princeton.edu/perl/webwn

On peut remarquer que WordNet donne une définition du lexème, qui est plus courte que celle donnée par des dictionnaires classiques et que le domaine d'utilisation (ou de définition) n'est pas précisé. Toutefois, pour résoudre ce problème, une ressource complémentaire a été créée WordNet Domains[42] qui a permis d'annoter les synsets de WordNet avec au moins un domaine sémantique choisi parmi les deux cents existants dans la hiérarchie de domaines. En même temps, les relations sémantiques entre les lexèmes et les concepts dans WordNet sont explicitées et étiquetées, ce qui permet aux utilisateurs une navigation dans l'espace conceptuel.

Plusieurs travaux ont porté sur l'exploitation des connaissances de WordNet pour l'indexation et la recherche d'images. Ainsi, Hollink, Schreiber et Wielinga (2007) réalisent une étude sur l'utilisation des relations sémantiques de WordNet dans l'extension des requêtes pour la recherche d'images. Dans leurs expériences, ils utilisent 202 peintures de 25 peintres différents qui ont été annotées par 12 annotateurs. Ensuite, pour retrouver les peintures annotées, 15 requêtes ont été formulées avec des concepts génériques de type *mountain* (montagne), *apple* (pomme), *cloud* (nuage), *tree* (arbre), etc. Chaque requête a été formulée de trois façons différentes afin que le système retrouve les peintures contenant le concept de la requête, les concepts liés par la relation d'hyponymie ou par d'autres relations sémantiques. Ils montrent que l'extension des requêtes par les relations d'hyperonymie améliore la précision de 85 %, tandis que les autres relations sémantiques, comme la méronymie ou l'holonymie, améliorent le rappel de la recherche.

Popescu, Grefenstette et Moëllic (2007) ont développé un système qui utilise la représentation OWL (Web Ontology Language) de WordNet pour l'extension des requêtes. Ainsi, pour une requête donnée, le système retourne les images dont les concepts sont associés aux concepts de la requête par les relations d'hyponymie. Dans le cas d'ambiguïté, ils proposent également d'utiliser WordNet pour regrouper les images suivant les différents sens du lexème ambigu. Par exemple, les images pour le mot-clé *angora* peuvent être regroupées selon les trois sens de ce lexème dans WordNet : lapin, chat et chèvre. Ainsi, des ensembles d'images structurés pour chaque sens du mot-clé de la requête sont proposés à l'utilisateur au lieu d'une simple liste d'images. Pour ce faire, les images sont collectées tout d'abord sur le web pour chaque

[42] http://wndomains.fbk.eu/index.html

terme du synset d'une entrée de WordNet en utilisant un système de recherche classique. Avant que les images trouvées ne soient attachées à une entrée ontologique, *grizzly* par exemple, l'algorithme k-SNN de clusterisation est utilisé pour regrouper les images en ensembles visuellement cohérents. Cela leur permet ensuite de déterminer seulement deux clusters pour *grizzly* (cf. figure 3.2.) en utilisant les caractéristiques visuelles des images retrouvées. Pour évaluer le système, Popescu et ses collègues l'ont testé avec quinze concepts familiers de type *car* (voiture), *butterfly* (papillon), *apple* (pomme). Les évaluations effectuées ont montré qu'avec l'extension des requêtes, la pertinence des images retrouvées s'améliore. Ils obtiennent une précision moyenne des images trouvées de 80 %, tandis que la même mesure est de 60 % pour Yahoo. Toutefois, l'inconvénient majeur de leur système est qu'il ne peut pas être interrogé avec des requêtes complexes[43]. Dans ce cas, les auteurs suggèrent l'utilisation d'une autre ressource comme ConceptNet (Liu & Singh, 2004), qui fournit des connaissances sur les sens communs des termes en anglais.

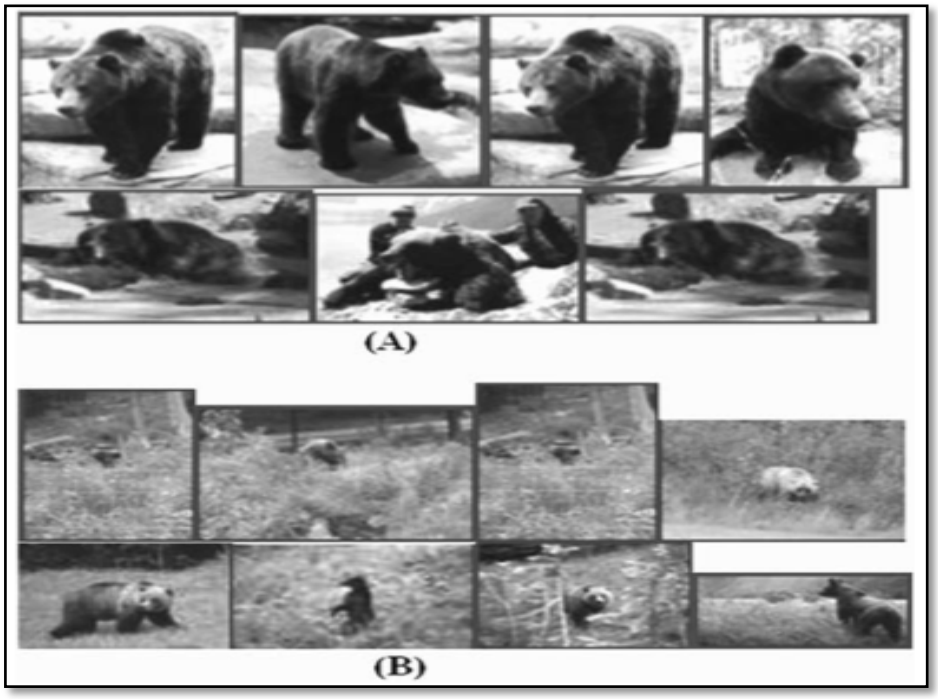
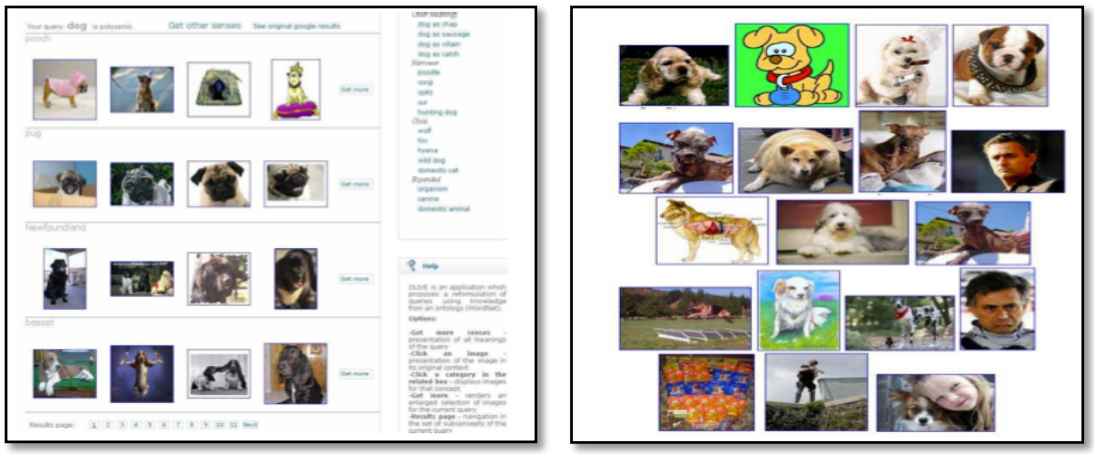


Figure 3.2. Exemple de deux clusters déterminés pour *grizzly* (Popescu, Grefenstette, & Moëllic, 2007)

[43] Requêtes contenant plus d'un concept.

Popescu et Grefenstette (2008), en utilisant leur système de recherche OLIVE décrit ci-dessus, réalisent une autre série d'expérimentations afin de valider leur méthodologie. Dans un premier temps, les expériences sont menées sur un corpus plus grand pour évaluer les performances de leur système par rapport à Google Images. Pour 40 concepts évalués, Google Images n'a proposé de meilleurs résultats qu'OLIVE que pour 2 concepts *lake* (lac) et *computer* (ordinateur), dans les autres cas, OLIVE retrouvait des images plus pertinentes avec une précision de 84,9 % par rapport à 68,2 % de précision obtenue par Google Images. Les résultats obtenus sont cohérents avec ceux obtenus auparavant (Popescu, Grefenstette & Moëllic, 2007) et montrent une fois encore les performances des systèmes utilisant WordNet.



(a) (b)

Figure 3.3. Exemple des résultats obtenus pour *dog (chien)* dans Olive (a) et Google Images (b) (Popescu & Grefenstette, 2008)

Dans un deuxième temps, ils ont réalisé un nouveau test afin d'évaluer l'interaction des utilisateurs avec le système. Ils ont demandé aux évaluateurs d'utiliser le système et de répondre ensuite à dix questions sur les caractéristiques du processus d'interaction. Neuf évaluateurs sur dix ont trouvé que la présentation structurée des résultats est préférable à celle de Google Images. En outre, les évaluateurs ont trouvé que les résultats proposés par OLIVE étaient plus pertinents que ceux proposés par Google Images. Les résultats de ces travaux montrent que l'utilisation de structures sémantiques dans le processus de recherche d'images permet de mieux répondre aux besoins d'information des utilisateurs et d'introduire un certain niveau d'interactivité avec

l'utilisateur. C'est pour cette raison que dans notre travail, nous allons nous inspirer de l'approche de structuration de résultats de recherches proposée par Popescu et Grefenstette (2008) en utilisant à la place de WordNet des hiérarchies sémantiques construites automatiquement à partir du *TLFi*.

En effet, même si WordNet est utilisé avec succès dans plusieurs travaux de recherche, l'ontologie qui en résulte, avec les concepts et les relations hiérarchiques correspondantes, est assez restreinte et ne permet pas de décrire toute l'information sur les images ou de répondre aux requêtes complexes. La richesse des informations contenues dans un grand dictionnaire de langue tel que le *TLFi* devrait remédier à cette faiblesse.

3.1.2. ConceptNet

ConceptNet (Liu & Singh, 2004) est un réseau sémantique construit automatiquement à partir des 700 000 phrases écrites par 14 000 web-contributeurs du projet *Open Mind Common Sense* (Singh & Barry, 2003). Ainsi, la base de connaissances du ConceptNet contient plus de 300 000 concepts et 1,6 million d'assertions. Parce qu'il a été créé automatiquement, ConceptNet présente des avantages et des inconvénients. Les principaux avantages sont que, par rapport à WordNet, il contient vingt relations sémantiques différentes telles que des relations de causalité, de location, etc., et des concepts composés d'ordre supérieur (concepts avec arguments) qui décrivent par exemple des verbes d'action avec un ou deux arguments directs ou indirects (ex. *buy food*, *drive to store*).

Dans la figure 3.4., qui présente un extrait du réseau sémantique du ConceptNet, nous voyons que *read newspaper* (lire le magazine) est un sous-événement *du eat breakfast* (prendre le petit déjeuner), même si dans le schéma on peut juste s'interroger sur l'orientation des relations.

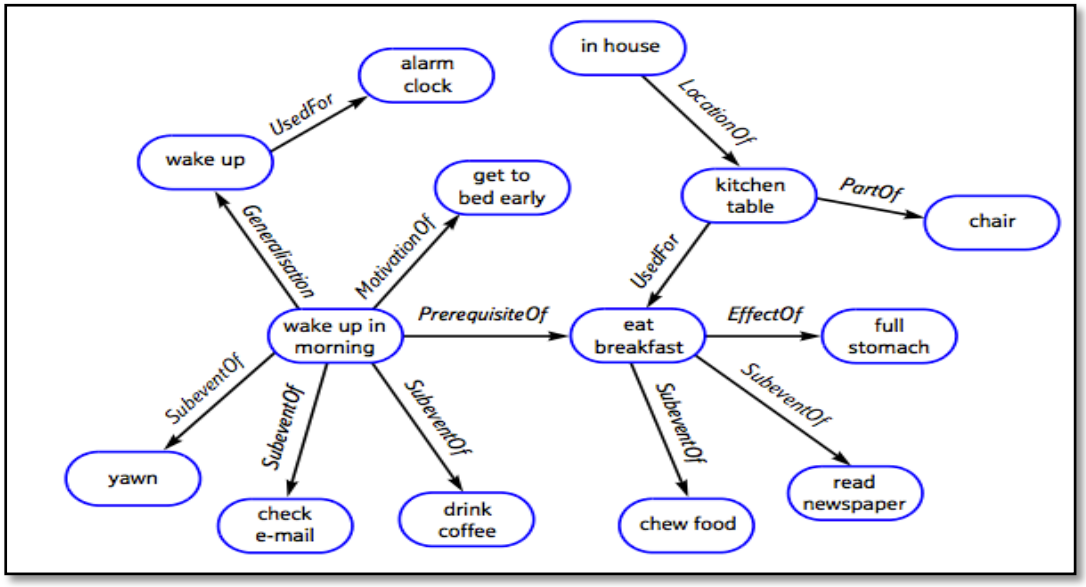


Figure 3.4. Un exemple d'un extrait du réseau sémantique du ConceptNet (Liu & Singh, 2004)

L'inconvénient du ConceptNet est qu'il ne distingue pas les différents sens d'un lexème. Toutefois, une désambiguïsation contextuelle peut être effectuée pour déterminer le sens le plus probable d'un terme ayant de multiples significations. Plusieurs autres travaux utilisent ConceptNet dans la recherche d'images. Ainsi, Hsu et Chen (2006) utilisent les relations spatiales de ConcepNet pour trouver les concepts qui sont spatialement liés aux concepts de la description textuelle de l'image. Ils obtiennent une amélioration de la précision de seulement 2,83 % dans le cas d'exploitation du titre et de la légende de l'image. Une amélioration plus significative de la précision peut être obtenue en combinant Wordnet et ConcepNet (Hsu, Tsai, & Chen, 2008).

3.1.3. DBpedia

DBpedia (Auer, Bizer, Kobilarov, Lehmann & Ives, 2007) est une ressource qui a été construite automatiquement grâce à un effort communautaire à partir des pages de Wikipédia[44] pour structurer les informations et les rendre disponibles sur le web. Elle offre de plus de nouveaux moyens d'accès à ces données. Ainsi, en utilisant DBpedia il

[44] Wikipédia (www.wikipédia.org) est une encyclopédie collective établie sur Internet, donc le contenu est publié par des contributeurs qui sont de simples utilisateurs ou internautes.

est possible de construire des requêtes complexes (ex. *Quelles sont les communes d'Ile-de-France ?*) afin d'interroger les données de Wikipédia.

Les connaissances contenues dans DBpedia sont disponibles en cent onze langues. La version anglaise de DBpedia décrit actuellement 3,77 millions d'objets, dont 2,35 millions sont classés dans une ontologie cohérente, en incluant 764 000 personnes, 573 000 places (y compris 387 000 lieux habités), 333 000 œuvres de création (films, jeux vidéo), 192 000 organisations (entreprises, établissement d'enseignement), 202 000 espèces (animales et florales) et 5 500 maladies[45]. Ainsi, par rapport aux autres bases de connaissances, DBpedia couvre de nombreux domaines et elle est véritablement multilingue.

Aspura, Khalid, Noah et Abdullah (2011) utilisent DBpedia comme ontologie de domaine pour construire une ontologie multimodale qui fournit des interprétations sémantiques du contenu de l'image. L'information sémantique des images se fonde sur trois ontologies principales (ontologie de domaine, ontologie de descriptions textuelles, ontologie de descriptions visuelles) créées à partir des descriptions textuelles et des caractéristiques visuelles des images. Une approche similaire est proposée par Khalid et Noah (2011), qui utilisent aussi DBpedia comme ontologie de domaine du sport.

3.2. Thésaurus

Un thésaurus est une liste de termes normalisés dans un domaine de connaissance. Le plus souvent les termes du thésaurus sont reliés par des relations hiérarchiques, synonymiques ou associatives.

La définition du lexème *terme* est ici ambigüe. Le sens traditionnel avec lequel le lexème *terme* est souvent utilisé selon Felber (1987) est celui du « représentant linguistique d'un concept dans un domaine de connaissance ». Mais en indexation la notion de *terme* est parfois confondue avec la notion de *descripteur* (unité d'indexation – unité lexicale). Dans le *Vocabulaire de la documentation* (Boulogne, 2004, p.73), le *descripteur* est défini comme « un terme retenu dans un thésaurus pour représenter sans ambiguïté une notion contenue dans un document ou dans une demande de recherche documentaire ». Les descripteurs peuvent être un nom commun ou un nom propre, une locution, un mot composé ou un groupe de lexèmes retenus après un processus

[45] http://wiki.dbpedia.org/About

d'analyse terminologique. Un lexème sera considéré comme terme par décision humaine, selon son utilisation dans le domaine. Dans le thésaurus un terme se situe entre les concepts et les descripteurs. Par exemple, le concept de *loisir* peut s'exprimer par les termes *Loisir* ou *Activité*, eux-mêmes s'exprimant par les descripteurs différents.

Comme l'un de nos objectifs est de pouvoir s'intégrer à terme dans le système de Xilopix qui utilise un thésaurus, nous allons expliciter un peu plus l'usage de thésaurus en indexation et recherche d'images.

3.2.1. Thésaurus — outil d'indexation et de recherche

Le thésaurus représente un outil d'indexation et de recherche qui détermine l'environnement sémantique d'un terme. Avoir un thésaurus bien structuré et normalisé est important car cela facilite l'accès et la recherche d'informations. Le rôle du thésaurus est de hiérarchiser des termes de manière à permettre un accès facile, une harmonisation pour contrôler le vocabulaire, une reformulation pour élargir ou focaliser la recherche.

D'une part, le thésaurus est un outil documentaire d'indexation. En utilisant un thésaurus pertinent, il est possible de représenter le contenu d'un document par un ensemble de lexèmes précis, extraits rigoureusement. Cela permet ensuite d'assurer l'indexation de documents dans une banque de données bibliographiques par exemple. D'autre part, le thésaurus est aussi un instrument de recherche. Par exemple, dans le catalogue d'un centre de documentation disposant de vocabulaires et de règles d'indexation, l'utilisateur peut optimiser ses requêtes ou même explorer le thésaurus par une navigation intuitive. C'est pourquoi le thésaurus est utilisé dans de nombreux travaux (Voorhees, 1993 ; Sparck Jones, 1986) comme outil d'indexation et de recherche d'informations.

En vue de leur double utilisation, les termes du thésaurus sont de deux types :

1. les descripteurs utilisés pour indexer un document ;
2. les non-descripteurs[46] utilisés lors de la recherche pour diriger les utilisateurs vers un ou plusieurs descripteurs.

[46] Terme du thésaurus non retenu pour représenter une notion. Il s'agit de synonymes, quasi-synonymes, abréviations ou variantes orthographiques du concept retenu comme descripteur.

Parfois l'utilisateur peut formuler des requêtes à l'aide de descripteurs ou de non-descripteurs (dans ce cas, le système le renvoie au descripteur), ce qui peut, dans une certaine mesure, limiter le bruit ou le silence documentaire.

Le thésaurus utilisé comme outil d'indexation et de recherche facilite donc l'accès et la recherche documentaire en augmentant l'efficacité grâce à des relations sémantiques contrôlées. Il permet également un repérage exhaustif en reliant les concepts et les termes proches, et réduit l'impact des problèmes liés à la synonymie présente dans le langage naturel.

Dans le domaine de l'indexation d'images, plusieurs thésaurus sont utilisés. Ils sont construits manuellement ou automatiquement, pour un ou plusieurs domaines compte tenu du type d'informations traitées et se différencient de ceux utilisés dans la RI textuelle. Ainsi, le thésaurus utilisé pour l'indexation d'images doit permettre de décrire toutes les représentations figurées : tableaux, peintures, figures, images, vitraux, etc. en prenant en compte les différences du contenu représenté. Il doit fournir tous les descripteurs qui permettront de répondre aux questions que le documentaliste se pose avant d'indexer une image : « Quelle est la signification principale de l'image ? Comment est-elle exprimée ? Quels sont les personnages ? Quelles sont leurs caractéristiques ? Connaît-on leurs noms ? Quels éléments concourent à l'expression de la signification principale ? Quelles sont les relations importantes, signifiantes et signifiées ? Quel est le lieu ? Peut-on l'identifier géographiquement ? Convient-il de le décrire ? La compréhension de l'image nécessite-t-elle une référence à un contexte historique ? Pose-t-elle des problèmes de datation ? » (Baryla, 1985). La construction d'un tel thésaurus nécessite beaucoup de temps et de connaissances approfondies du contexte de description de la part du documentaliste.

Un exemple de thésaurus utilisé pour l'indexation manuelle d'images est le thésaurus iconographique Garnier utilisé dans les musées. C'est un thésaurus qui traite les sujets représentés dans les œuvres d'art, comme le thème, la nature, le corps et la vie matérielle, la vie psychologique et morale, la société et la vie sociale, la vie politique et administrative, l'armement et la vie militaire, l'agriculture, la chasse et la pêche, l'énergie, l'artisanat, l'industrie et le commerce, les services, le transport et les communications, la vie intellectuelle et scientifique, les arts et les spectacles, la vie religieuse, l'être imaginaire, l'ornement, le sujet géographique, le sujet biblique, le sujet

bouddhique, le sujet taoïste, la mythologie, les noms de groupes (nationalité, culture), la périodisation et le personnage imaginaire.

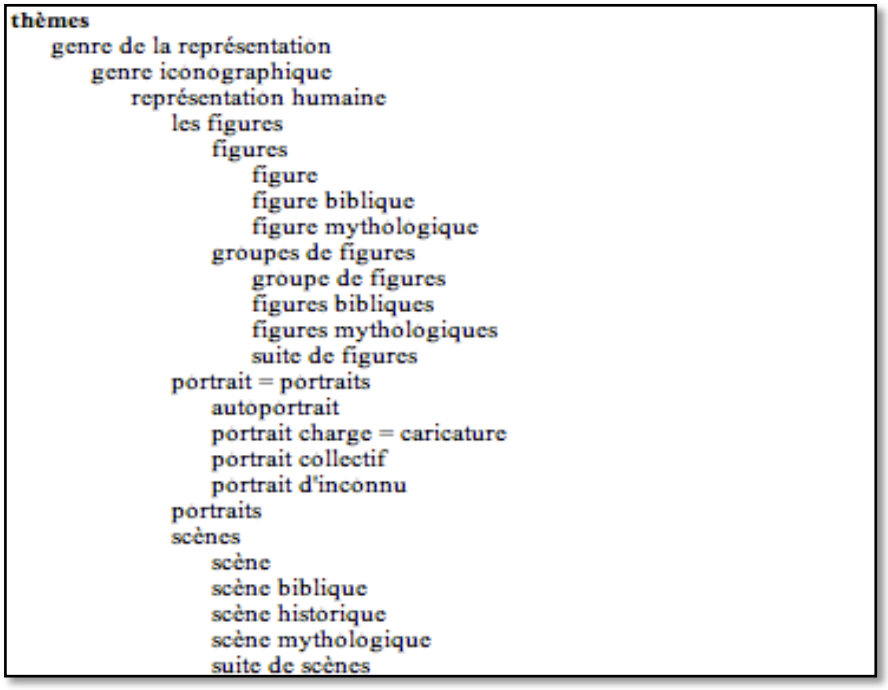


Figure 3.5. Exemple d'un extrait de thésaurus Garnier sur le sujet *thèmes*

La figure 3.5. ci-dessus présente un extrait du thésaurus Garnier sur le sujet *thèmes*.

Le thésaurus Garnier est utilisé dans sa version informatisée dans le portail des collections des musées de France Joconde[47]. L'exemple du thésaurus Garnier montre que l'indexation des images se fait sur la base d'un thésaurus spécifique au domaine.

Les systèmes de recherche d'images en ligne comme Getty Images[48] utilisent aussi des thésaurus qui permettent aux documentalistes d'annoter les images et ensuite de les indexer. Ainsi, chaque thésaurus est construit en fonction du domaine d'utilisation.

L'inconvénient majeur de ces thésaurus est qu'ils ne permettent pas de décrire les relations entre les objets de l'image ou les caractéristiques visuelles. Un autre problème réside dans les différences entre les termes de la requête de recherche et ceux du thésaurus. Ainsi, ces derniers temps, de nombreux travaux utilisent des ontologies pour l'annotation et la recherche d'images (cf. §3.3.).

[47] http://www.culture.gouv.fr/documentation/joconde/fr/pres.htm
[48] http://www.gettyimages.fr/

3.2.2. Les relations dans le thésaurus

Les termes d'un thésaurus sont organisés hiérarchiquement. Tout thésaurus comporte au moins trois catégories de termes : les termes génériques et les termes spécifiques qui doivent être utilisés comme descripteurs, et les termes équivalents considérés comme non-descripteurs :

- Les *termes génériques* **(TG)** désignent les entités ou concepts principaux en référence aux autres termes et au domaine considéré ;
- Les *termes spécifiques* **(TS)** précisent et identifient les entités ou concepts particuliers à l'intérieur du champ sémantique d'un terme générique donné ;
- Les *termes équivalents* **(EM)** sont des variantes des termes spécifiques (synonymes ou quasi-synonymes). Ils sont donc équivalents dans le langage courant, mais donnés comme complémentaires dans l'emploi du thésaurus.

Généralement, dans un thésaurus, on peut trouver aussi des *termes associés* **(TA)** (relation d'association : causalité, localisation, etc.). Ces termes sont aussi des descripteurs et permettent à l'utilisateur de modifier sa requête ou de l'élargir sans faire appel aux termes génériques.

Dans un thésaurus, les relations entre les termes sont donc de trois types :

- *relation hiérarchique* (entre descripteurs). Elle représente un lien vers un concept de sens plus large ou plus précis (exemple : TG vertébrés, TS mammifères, oiseaux). Il existe plusieurs types de relations hiérarchiques génériques, partitives, organisation par thèmes ou facettes. Les termes sont regroupés par champs sémantiques d'environ cinquante lexèmes, eux-mêmes hiérarchisés.
- *relation d'équivalence* (entre descripteurs et non-descripteurs). C'est une relation entre termes représentant un même concept. Elle recouvre les fonctions de synonymie (exemple : aspirine, acide acétylsalicylique) et de quasi-synonymie qui s'appliquent selon le contexte (exemple : lunette, télescope).
- *relation d'association* (entre descripteurs). Elle représente un lien vers les termes liés par le sens uniquement, sans liaison hiérarchique (exemple : coiffure, séchoir à cheveux), et peut aussi être utilisée pour l'enrichissement sémantique.

La figure 3.6. présente les relations et les termes contenus dans un thésaurus de genres cinématographiques.

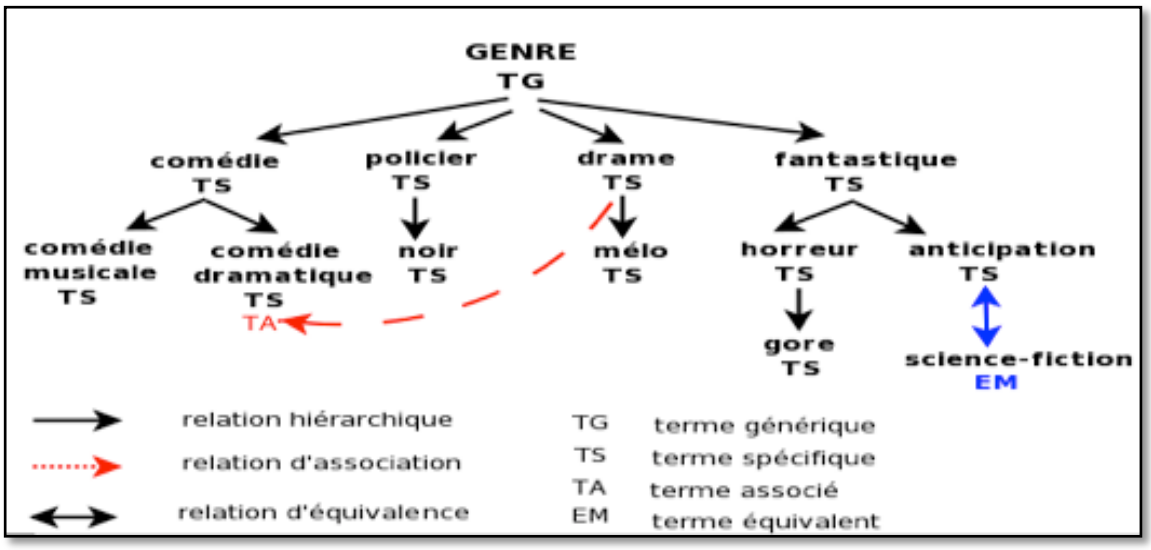


Figure 3.6. Représentation des relations et des termes dans un thésaurus de genres cinématographiques

3.2.3. Les modes de représentation des thésaurus

La présentation d'un thésaurus pour l'utilisation ou l'exploration peut se faire selon plusieurs modes :

- liste alphabétique de termes, pour une approche globale ou la recherche d'un terme particulier ;
- liste hiérarchique de termes, pour l'approfondissement d'une notion ;
- liste d'occurrences, pour la vérification de la pertinence d'un élément d'une expression utilisée comme descripteur.

Les thésaurus peuvent être représentés sous forme graphique. Ce type de représentation facilite l'accès aux informations et regroupe les concepts qui ont des sens proches. Aitchison et Gilchrist (1992) différencient trois types de représentation graphique des thésaurus :

1. *Structure arborescente.*

Dans ce type de représentation, les termes sont hiérarchisés : le terme le plus général figure en haut de la hiérarchie et les termes spécifiques sont situés aux niveaux inférieurs. Les relations entre les termes sont indiquées par les liens verticaux qui les relient.

En général, dans une telle structure nous distinguons deux types d'arborescences :

- Le thésaurus hiérarchique à arborescence précise

  Dans ce type de thésaurus, les fils du nœud père représentent des partitions de celui-ci et sont des éléments distincts sans aucune relation entre eux (cf. figure 3.7.).

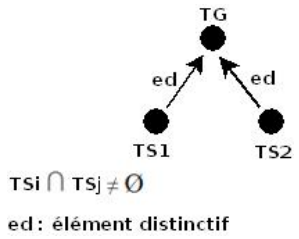


Figure 3.7. Modèle de thésaurus hiérarchique à arborescence précise

- Le thésaurus hiérarchique à arborescence simple

  Dans ce type de thésaurus, les fils du nœud père représentent des éléments de celui-ci liés par la relation *est-un* (cf. figure 3.8.).

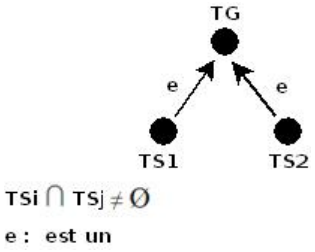


Figure 3.8. Modèle de thésaurus hiérarchique à arborescence simple

2. *Schémas fléchés.*

Ce sont des représentations du thésaurus sous forme de tableaux pour un certain nombre de domaines spécifiques (champs sémantiques) et de sous-domaines. Les tableaux peuvent être présentés aussi sous forme de grilles. Dans ces tableaux, les termes génériques sont habituellement situés au milieu et les termes spécifiques sont concentrés autour des termes génériques. Les niveaux hiérarchiques sont représentés à l'aide de flèches qui les relient depuis le plus haut degré de spécificité jusqu'au plus bas et les relations associatives avec les descripteurs d'autres domaines sont représentées par les traits non fléchés à l'extérieur du cadre. Ainsi, dans la figure 3.9., tous les descripteurs du domaine *Traitement* du thésaurus *Vieillissement* (Baddas & Labarde, 2003) sont représentés en utilisant le schéma fléché.

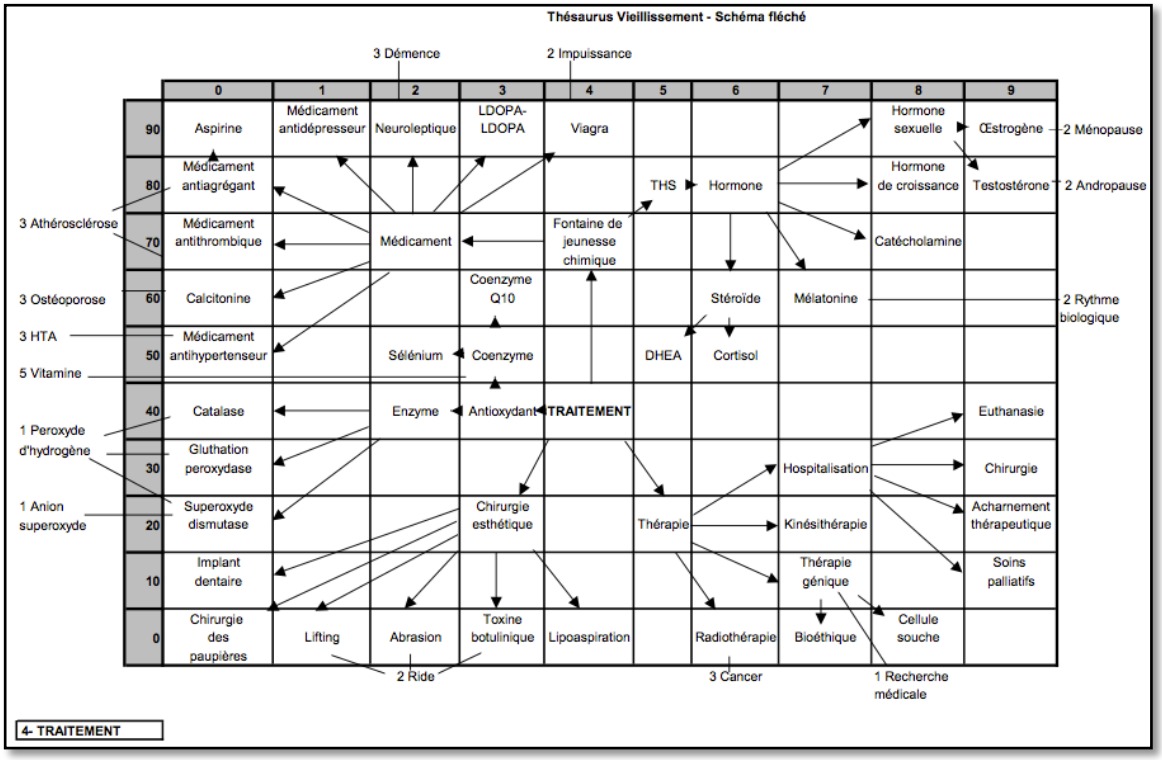


Figure 3.9. Exemple de schéma fléché du domaine *Traitement* du thésaurus *Vieillissement* (Baddas & Labarde, 2003)

3. Terminogrammes.

Cette représentation du thésaurus est appelée aussi « schéma compartimenté ». Elle permet de découper le vocabulaire du thésaurus en domaines spécifiques (champs sémantiques), où chacun appartient à un tableau, lequel contient des descripteurs caractéristiques et des termes équivalents identifiés. Dans la hiérarchie de chaque domaine spécifique, les descripteurs spécifiques sont décalés vers la droite par rapport aux descripteurs génériques. À l'extérieur de chaque tableau, des relations associatives avec les descripteurs d'autres domaines sont indiquées. La figure 3.10. présente un terminogramme pour le domaine *Traitement* du thésaurus *Vieillissement* (Baddas & Labarde, 2003).

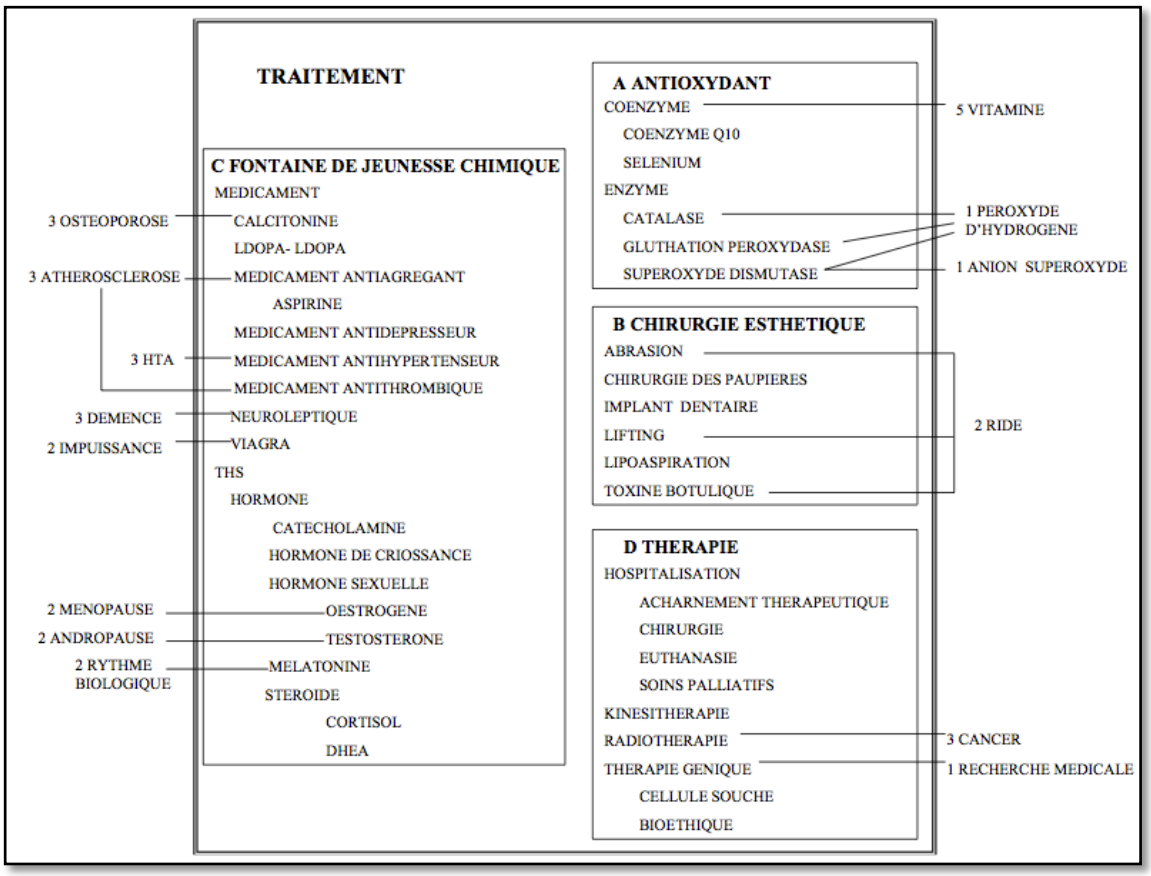


Figure 3.10. Exemple de terminogramme pour le domaine *Traitement* du thésaurus *Vieillissement* (Baddas & Labarde, 2003)

En vue de réaliser un enrichissement du thésaurus existant chez Xilopix qui a une structure arborescente, nous sommes intéressée par la construction de hiérarchies sémantiques à partir du *TLFi* sous forme d'arbres hiérarchiques à arborescence simple.

3.3. Ontologies

Quand on parle d'« ontologies », on utilise souvent la notion de *concept* qui, selon le *TLFi,* est « une représentation mentale abstraite et générale, objective, stable, munie d'un support verbal ». C'est pourquoi Gruber (1993) considère une ontologie comme une spécification d'une conceptualisation, c'est-à-dire d'une vue simplifiée du monde que nous voulons représenter pour un certain but (Gruber, 1995). Ainsi, nous pouvons affirmer que l'ontologie organise les connaissances pour un domaine donné en précisant les différentes relations qui peuvent exister entre ces connaissances. Toutefois, elle ne représente pas simplement un vocabulaire spécialisé pour un domaine, mais plutôt une

conceptualisation des termes d'un vocabulaire pour un domaine donné. Chandrasekaran, Josephson et Benjamins (1999) mentionnent aussi que l'identification du vocabulaire et des conceptualisations sous-jacentes exige une analyse minutieuse des types d'objets et des relations qui peuvent exister dans le domaine. La conceptualisation des termes peut poser aussi un problème car le sens de tous les concepts n'est pas formulable linguistiquement. Ainsi, Bachimont (2000) donne l'exemple du concept *timbre* du domaine de la *musique* qui est défini comme « ce qui n'est ni la hauteur, ni l'intensité du son », ce qui en réalité ne permet pas de déterminer le contenu de ce concept.

3.3.1. Structures et types d'ontologies

Dans les ontologies la structuration hyperonymique/hyponymique est aussi présente, les relations entre les concepts étant formalisées explicitement. Ainsi, une ontologie représente une hiérarchie de concepts pour décrire des connaissances plus spécifiques dans un domaine particulier. Elle est composée des éléments suivants :

- *Classes* : elles constituent le centre d'intérêt de l'ontologie et décrivent les concepts dans le domaine. Par exemple, une classe *Fromage* représente tous les fromages. Les fromages spécifiques comme *Emmental*, *Compté* sont des instances de cette classe.
- *Attributs* : ils décrivent les propriétés des classes et des instances. Par exemple, le fromage grec *Feta* est un produit du lait de brebis. Les autres attributs décrivant les instances de la classe *Fromage* peuvent être le producteur, la valeur énergétique, le niveau de lipides, etc.
- *Facettes* : elles correspondent à des restrictions sur les attributs. Par exemple, une facette peut spécifier que la valeur de l'attribut producteur est de type chaîne de caractère.

Ce qui distingue les différentes approches des ontologies est la précision des sens des termes (Uschold & Gruninger, 2004). Ainsi, selon les auteurs, il existe deux types d'ontologies :

- *Ontologies légères (lightweight ontologies)* : ce sont des ontologies qui ne s'appuient pas sur un formalisme comme les taxonomies ou les dictionnaires (cf. figure 3.11.). Dans ce type d'ontologies, les concepts sont liés par l'intermédiaire d'associations dont le type de relation n'est souvent pas précisé.

Dans le domaine de la recherche d'informations, les ontologies légères sont utilisées essentiellement pour l'extension des requêtes ou la catégorisation des documents.

- *Ontologies lourdes (heavyweight ontologies) :* ce sont des ontologies formelles (cf. figure 3.11.) qui fournissent une représentation formelle des concepts, par exemple à l'aide de logiques de description. Les ontologies lourdes sont souvent utilisées pour l'extraction d'information à partir de textes, car dans ce type d'ontologies il est possible de définir des règles de type comme *Si X Alors Y*.

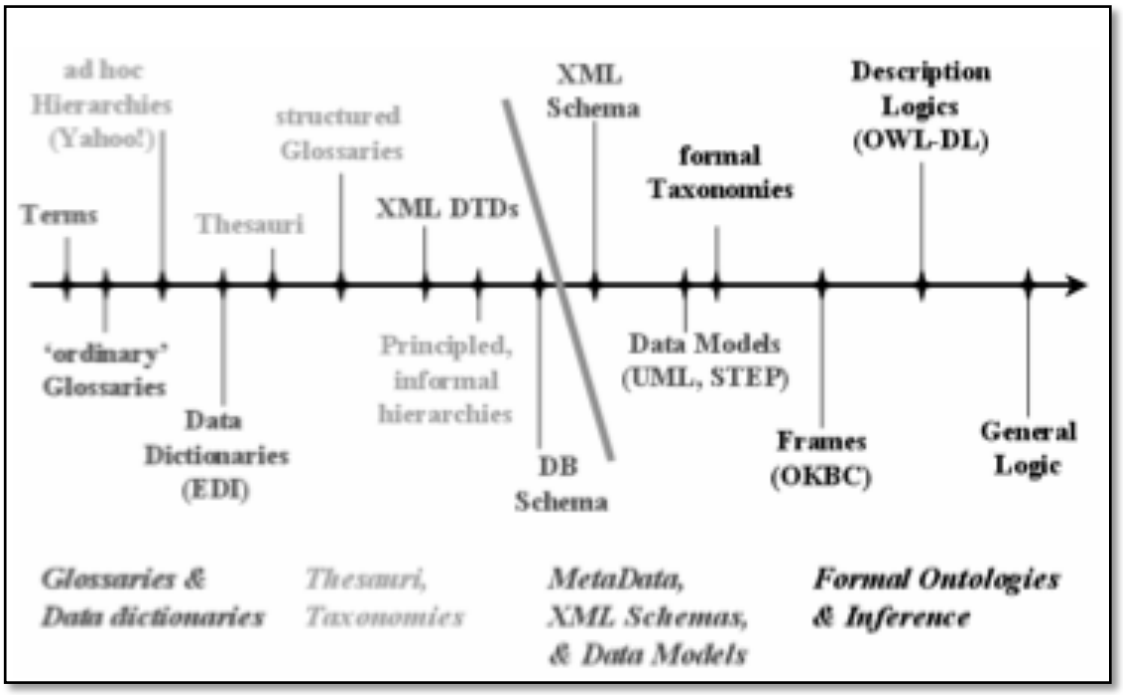


Figure 3.11. Exemple de types d'ontologies (Uschold & Gruninger, 2004)

Toutefois, selon le niveau de généralité (d'abstraction), les ontologies peuvent être de quatre types (Guarino, 1998) :

1. *Ontologies de haut niveau (top level ontology*) : elles décrivent des concepts très génériques comme le temps, l'espace, les objets, les événements, les actions, etc. et sont indépendantes d'un domaine particulier.
2. *Ontologies de domaine (domain ontology)* : elles décrivent le vocabulaire lié à un domaine générique comme le domaine de la médecine ou de l'automobile.
3. *Ontologies de tâche (task ontology)* : elles décrivent le vocabulaire lié à une tâche ou une activité générique comme le diagnostic ou la vente.

4. *Ontologies d'application (application ontology*) : elles décrivent les concepts en fonction à la fois du domaine particulier et de la tâche en représentant souvent la spécialisation de ces deux ontologies connexes. Elles sont utilisées pour répondre à des utilisations spécifiques, par exemple la radiologie dentaire ou la radiologie pulmonaire dans le domaine de l'imagerie médicale.

La figure 3.12. montre les types d'ontologies selon leur niveau de dépendance par rapport à une tâche particulière ou un point de vue : les flèches représentent ici des relations de spécialisation.

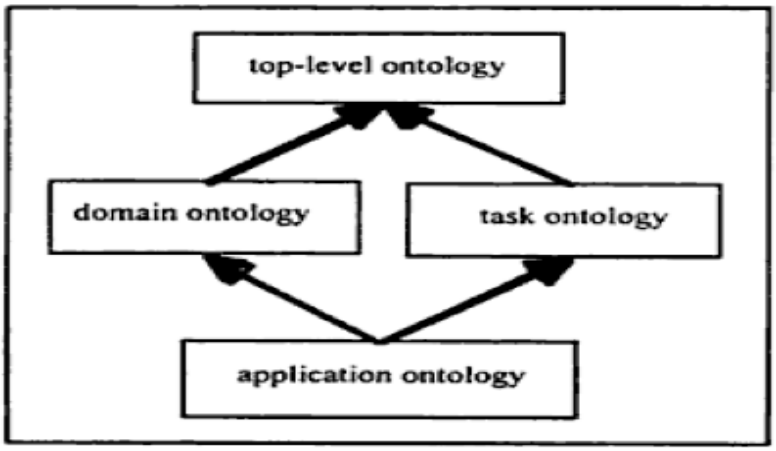


Figure 3.12. Différents types d'ontologies selon leur niveau d'abstraction

(Guarino, 1998)

Ainsi, les ontologies sont souvent organisées sous forme d'un arbre hiérarchique de conceptualisation, allant d'un niveau très générique vers un niveau très spécifique. Toutefois, en fonction des ontologies, les concepts génériques peuvent avoir différents sous-types. En effet, lors de la construction de l'ontologie, différents critères de catégorisation peuvent être pris en compte. Ainsi, Chandrasekaran, Josephson et Benjamins (1999) donnent un exemple de catégorisation d'un même concept générique *thing* (chose) dans différentes ontologies (cf. figure 3.13.). De plus, les auteurs mentionnent que la construction des ontologies dépend beaucoup de leur utilisation. C'est pourquoi plusieurs ontologies peuvent coexister pour un même domaine, mais encodent des informations différentes en fonction de la tâche à accomplir. Ainsi, une ontologie de domaine *apiculture* peut décrire l'aspect de *l'élevage des abeilles* ou de *la production du miel.*

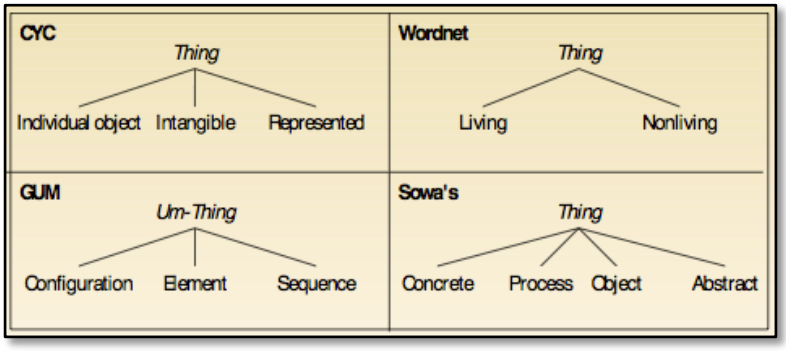


Figure 3.13. Exemple de catégorisation d'un concept générique dans différentes ontologies (Chandrasekaran, Josephson, & Benjamins, 1999)

3.3.2. Utilisation d'ontologies dans le domaine d'indexation et de recherche d'images

Plusieurs travaux exploitent les ressources existantes pour construire automatiquement des ontologies qui sont ensuite utilisées pour l'indexation et la recherche d'images. Toutefois, ces approches conviennent plus pour l'indexation d'images du Web. L'indexation d'images de domaines particuliers nécessite beaucoup plus de précision et par conséquent ce sont les ontologies construites manuellement qui conviennent le mieux pour cette tâche.

Hyvönen, Styrman et Saarela (2002), Hyvönen, Saarela et Viljanen (2004) utilisent une ontologie pour annoter les images de *Helsinki University Museum*. La même ontologie est ensuite utilisée pour la recherche d'images, ce qui permet aux utilisateurs d'user de la même terminologie lors de la formulation des requêtes de recherche.

Selon Hayes et Warren (2010) les vocabulaires existants pour la description des œuvres d'art, tels que *Visual Resources Association*[49] (VRA), ne permettent pas de décrire toute l'information sur les images. En analysant le vocabulaire utilisé par des photographes ou des peintres pour la description de leurs œuvres, ils proposent donc une ontologie adaptée pour la description d'images.

Dans le domaine de la médecine, Iakovidis, Schober, Boeker et Schulz (2009) proposent une nouvelle approche ontologique qui formalise les concepts et les relations concernant les représentations d'images pour l'extraction d'images médicales.

[49] http:// www.vraweb.org/projects/vracore4/

Différents descripteurs sont proposés pour décrire les régions de l'image, les pixels, les caractéristiques et les relations spatiales.

Selon Kong, Hwang, Na et Kim (2005) et Kong, Hwang et Kim (2006), l'utilisation d'ontologies construites manuellement, qui considèrent les caractéristiques du contenu d'images, permet de résoudre le problème de vocabulaire entre les termes de l'image et ceux de la requête. Ils proposent la création d'une ontologie personnalisée directement par les utilisateurs lors de l'annotation de l'image (cf. figure 3.14.). Afin d'avoir plusieurs descriptions sémantiques sur des images, ils adoptent aussi une ontologie spatiale qui représente les relations spatiales entre les objets de l'image. Une telle approche est très importante car elle permet ensuite de répondre aux requêtes complexes du genre *Ronaldo shoot the ball* (Ronaldo lance le ballon). Lors de l'évaluation de leur système, ils montrent que l'utilisation d'ontologies personnalisées et la prise en compte des caractéristiques visuelles des images permettent d'obtenir une meilleure précision que l'utilisation des ontologies issues de WordNet.

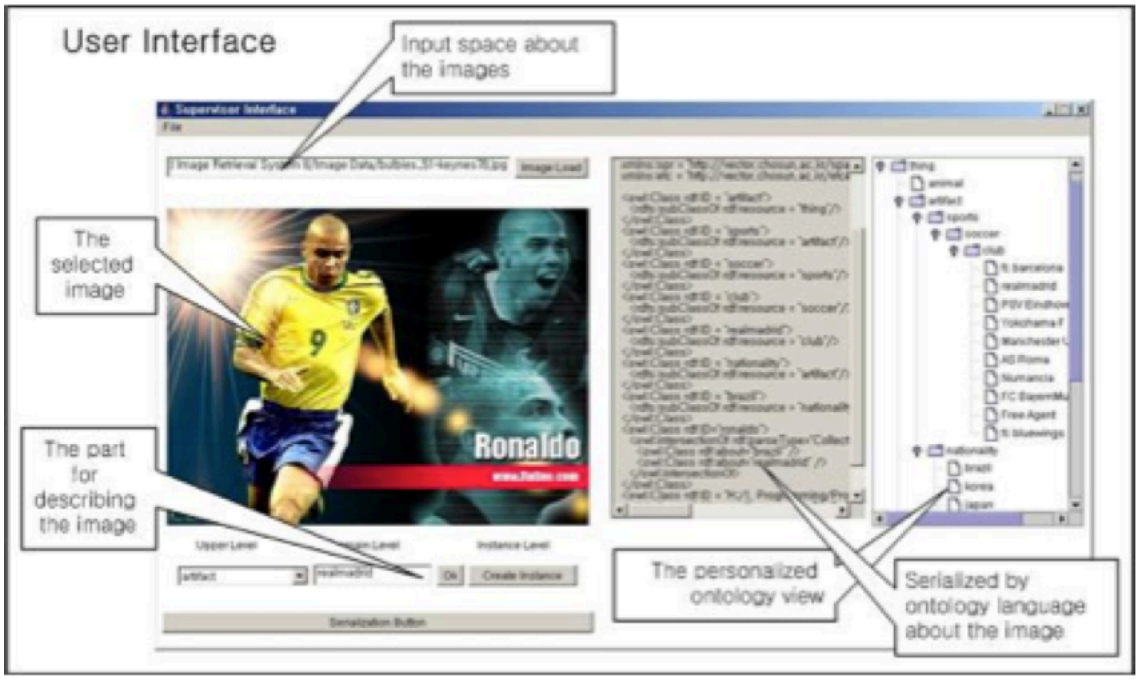


Figure 3.14. Exemple de l'interface de l'utilisateur pour l'annotation de l'image (Kong, Hwang, Na, & Kim, 2005)

Même si les ontologies permettent de mieux décrire le contenu des images que les thésaurus, le principal problème qui découle de leur utilisation est que la construction de telles ontologies reste très coûteuse en temps.

3.4. Construction automatique de hiérarchies sémantiques

Dans les sections précédentes, nous avons présenté les principales ressources lexicales utilisées dans le domaine d'indexation et de recherche d'images. Nous avons vu que les principaux travaux dans ce domaine soit exploitent les ressources lexicales existantes, soit construisent manuellement de nouvelles ressources (thésaurus, ontologies). Dans cette section nous nous intéressons aux méthodes de construction automatique de hiérarchies sémantiques à partir de textes et plus particulièrement à celles à partir de définitions lexicographiques.

3.4.1. Les études de construction de hiérarchies sémantiques à partir de textes

Les thésaurus et les taxonomies sémantiques de type WordNet représentent une source essentielle de connaissances pour les applications de traitement du langage naturel en fournissant ainsi des informations sur les relations entre les lexèmes de manière explicite. Toutefois, la construction de ces taxonomies est un processus extrêmement lent et laborieux. En outre, les taxonomies sémantiques sont toujours limitées en termes de portée et de domaine. Enfin le coût élevé de leur personnalisation ou de leur extension pour une application donnée réduit souvent leur utilité. C'est ainsi que beaucoup de travaux se sont orientés vers la recherche de méthodes pour apprendre automatiquement des relations taxonomiques et construire des hiérarchies sémantiques à partir du texte.

Les approches les plus utilisées pour la construction automatique de taxonomies ou d'ontologies à partir de textes en exploitant le contexte d'apparition des termes reposent principalement sur deux méthodes basées soit sur des patrons lexico-syntaxiques, soit sur une analyse distributionnelle.

Ainsi, les patrons lexico-syntaxiques sont définis à l'aide d'analyse du contexte d'apparition des termes dans le corpus afin de déterminer des relations sémantiques. Une grande partie des travaux antérieurs sur la classification sémantique automatique des lexèmes ont été basés sur l'idée proposée par Hearst (1992), que la présence de certains « patrons lexico-syntaxiques » peut indiquer une relation sémantique particulière entre deux noms. Hearst a remarqué que, par exemple, la liaison entre deux

syntagmes nominaux (NP) via les constructions « Such NP_Y as NP_X » ou « NP_X or/and other NP_Y » implique souvent que NP_X est un hyponyme de NP_Y. Depuis, cette technique a été adaptée dans plusieurs travaux. Ainsi, Cederberg et Widdows (2003) ont défini manuellement un petit nombre (généralement moins de dix) de patrons lexico-syntaxiques afin d'étiqueter automatiquement les relations sémantiques. Bien que ces patrons aient réussi à identifier des exemples de relations comme hyperonymiques, cette méthode est fastidieuse et sévèrement limitée par le petit nombre de patrons habituellement utilisés.

Afin de remplacer l'identification manuelle des patrons lexico-syntaxiques, plusieurs travaux utilisent des relations de dépendance syntaxique entre les termes pour déterminer par l'apprentissage supervisé de nouveaux patrons (Wu, et al., 2011 ; Ritter, Soderland & Etzioni, 2009 ; Pantel et Ravichandran, 2004 ; Agirre, et al., 2000). L'une des méthodes de plus haute couverture est proposée par Snow, Jurafsky et Ng (2005) qui utilisent des exemples de paires d'hyperonymes déjà connus à partir de WordNet pour identifier automatiquement de nouveaux patrons lexico-syntaxiques. Ensuite, ils combinent ces patrons avec un algorithme d'apprentissage supervisé pour obtenir un classificateur d'hyperonymes de haute précision. Ils ont montré que les patrons de Hearst sont parmi les patrons de plus haute précision des 70 000 patrons qu'ils ont induits pour l'identification de la relation hyperonyme-hyponyme. En raison de leur simplicité et leur efficacité, les méthodes basées sur les patrons de Hearst ont été appliquées à de nombreux systèmes d'extraction d'informations comme KnowItAll (Etzioni, et al., 2004), TextRunner (Banko, et al., 2007) et les travaux de Pasca (2008).

Les méthodes basées sur l'exploitation de la distribution contextuelle des termes visent à regrouper des termes qui partagent le même contexte syntaxique (Harris, 1990). Cette méthode a été mise en œuvre dans les logiciels UPERY (Bourigault, 2002), (Bourigault, Aussenac-Gilles, & Charlet, 2004), LEXICLASS (Assadi & Bourigault, 2000), LEXTER (Bourigault, 1994). Pour ce faire, les dépendances syntaxiques entre les lexèmes sont tout d'abord collectées à partir du corpus et ensuite une analyse de leur distribution est réalisée. Ainsi, dans leurs expérimentations Baneyx, Malaisé, Charlet, Zweigenbaum et Bachimont (2005) utilisent pour calculer les dépendances syntaxiques, l'analyseur syntaxique SYNTEX (Bourigault & Fabre, 2000). L'analyse de leurs distributions étant effectuée par le module UPERY. Le principal avantage de cette

approche est qu'elle permet de retrouver de relations qui n'apparaissent pas forcement dans le texte.

Certains travaux combinent les deux approches Caraballo (1999), (Yang & Callan, 2009). Ainsi, les termes qui partagent le même contexte sont tout d'abord regroupés et ensuite de patrons lexico-syntaxiques sont appliqués pour déterminer des relations d'hyperonymie.

Récemment, plusieurs travaux s'intéressent à l'extraction de connaissances sémantiques à partir de Wikipédia. Ainsi, Ponzetto et Strube (2007, 2011) dérivent une taxonomie à grande échelle contenant de relations de subsomption basées sur le système de catégories de Wikipédia. Ils considèrent les catégories de Wikipédia comme des concepts d'un réseau sémantique et annotent les relations entre les concepts comme étant *is-a* ou *not-is-a* en utilisant les méthodes basées sur la connectivité du réseau et en appliquant des patrons lexico-syntaxiques sur un très grand corpus. Leur taxonomie dérivée de Wikipédia s'est avérée être compétitive avec les ontologies existantes les plus importantes comme Cyc[50] (Guha & Lenat, 1990 ; Lenat, 1995) et WordNet. Afin de désambiguïser les catégories de Wikipédia et de restructurer la taxonomie déjà générée, Ponzetto et Navigli (2009) utilisent les synsets de WordNet. Garcia, Rensing et Steinmetz (2011), en adaptant des heuristiques existantes, ont montré qu'il est possible d'extraire de grands ensembles de relations d'hyponymie à partir de multiples versions de Wikipédia avec peu d'informations sur chaque langue possible.

Les travaux de Velardi, Faralli et Navigli (2011, 2013) s'intéressent également à l'induction automatique de taxonomies sémantiques à partir d'un corpus donné. Le procédé consiste à obtenir tout d'abord une terminologie du corpus, qui est ensuite utilisée de manière itérative pour obtenir des relations taxonomiques et de nouveaux termes potentiels à partir des textes. Les relations taxonomiques sont extraites à l'aide d'algorithme supervisé Word-Class Lattices[51] (WCLs) (Navigli & Velardi, 2010) qui détermine pour chaque terme de la terminologie extraite une définition avec son hyperonyme correspondant. Ensuite une technique de filtrage est appliquée pour filtrer

[50] Cyc est une base de connaissances qui représente sous une forme formalisée une grande quantité de connaissances fondamentales comme des faits, les heuristiques sur les objets et les événements de la vie quotidienne, à l'aide d'un langage spécialement conçu.

[51] Cette méthode permet de déterminer dans un corpus donné des définitions textuelles en annotant le genre prochain et les différences spécifiques de celles-ci ainsi que les hyperonymes contenus dans le genre prochain des définitions.

les définitions qui n'appartiennent pas au domaine d'utilisation de la terminologie extraite. Dans une dernière étape, le graphe est découpé en fonction des informations de connectivité et des restrictions imposées par les relations taxonomiques.

Ainsi, ces approches visent à construire plutôt des taxonomies que des ontologies à part entière comme l'ontologie YAGO construite automatiquement par Suchanek, Kasneci et Weikum (2008). Cette ontologie décrit 1,7 million d'entités, des personnes, des organisations, des villes et qui contiennent 15 millions de faits sur ces entités. Les données d'YAGO proviennent de Wikipédia et ont été structurées à l'aide de relations taxonomiques de WordNet. En évaluant leur ontologie, ils obtiennent une précision de 95 %, à comparer aux 87 % obtenus par Ponzetto et Strube (2007). Weber et Buitelaar (2006), quant à eux, ont proposé un système nommé Information System for Ontology Learning and Domain Exploration (ISOLDE) pour dériver des ontologies de domaine à l'aide d'un corpus de domaine annoté, un tagger d'usage général pour les entités nommées et les ressources du Web comme Wikipédia et Wiktionnaire.

Un état de l'art plus complet sur les techniques d'extraction et de structuration automatique des connaissances d'un domaine à partir du texte est proposé par (Clark, et al., 2012), (Manser, 2012), (Auger & Barrière, 2008).

Pour notre part, notre objectif n'est pas de structurer les connaissances à partir de textes, mais celles des définitions lexicographiques, qui sont de nature différente, dans un double objectif d'enrichissement des mots-clés des images et du thésaurus Xilopix. Nous allons donc dans la suite étudier plus particulièrement comment sont extraites les relations sémantiques à partir de définitions lexicographiques.

3.4.2. Les études de construction de hiérarchies sémantiques à partir de définitions lexicographiques

Les dictionnaires informatisés représentent une source d'information sémantique très importante pour les applications de traitement automatique des langues. Toutefois, la structure des dictionnaires informatisés ne permet pas leur utilisation immédiate par les systèmes puisque initialement les dictionnaires ont été conçus pour les humains et non pour les machines.

Ainsi, au début des années 80, plusieurs travaux se sont intéressés à l'exploitation des dictionnaires informatisés afin d'extraire automatiquement des bases de connaissances (Amsler, 1980 ; Michiels & Noël, 1982 ; Calzolari, 1984 ; Chodorow, Byrd & Heidorn, 1985 ; Byrd, Calzolari, Chodorow, Klavans, Neff & Rizk, 1987 ; Alshawi, 1987 ; Nakamura & Nagao, 1988 ; Copestake, 1990). L'élément clé de ces recherches est que les définitions dans les dictionnaires sont constituées d'un *genre prochain* qui représente la classe à laquelle appartient le lexème défini et des *caractéristiques spécifiques* qui le caractérisent au sein de cette classe. Ils se sont appuyés sur le fait que, dans ce type de définitions, la tête du genre prochain représente l'hyperonyme pour un lexème défini. À partir de cette hypothèse, plusieurs méthodes d'extraction de taxonomie hyperonymique ont été proposées. La faisabilité de la génération des hiérarchies sémantiques a été tout d'abord prouvée par Amsler (1980) qui a extrait et désambiguïsé manuellement les genres prochains du dictionnaire de poche *Merriam-Webster*. Toutefois, le but des autres travaux a été d'automatiser ce processus. Ainsi, la méthodologie de Chodorow, Byrd et Heidorn (1985) a consisté à trouver le genre prochain dans les définitions des verbes en extrayant les verbes précédés dans les définitions par le lexème *to.* Par exemple, dans la définition « to pass the winter » du verbe *winter* (passer l'hiver) le verbe *pass* (passer) représente le genre prochain. Toutefois, l'heuristique de détermination des genres prochains dans les définitions nominales est plus complexe à cause des différentes structures des définitions nominales. Elle consiste à extraire les lexèmes précédés par des premiers pronoms relatifs, des prépositions non suivies par une conjonction, des participes présents suite à un nom. Dans le cas où le lexème appartient aux expressions métalangagières comme *one* (un), *class* (classe), *manner* (manière), etc., le lexèmes suivi par *of* (de) est considéré comme hyperonyme. Ainsi, les genres prochains pour les définitions nominales ont été déterminés avec une précision de 98 %. Afin d'élargir l'arbre sémantique construit, pour chaque hyperonyme extrait est déterminé son genre prochain dans les définitions du dictionnaire. Par exemple, si *vehicle* (véhicule) représente l'hyperonyme du nom *ambulance* (ambulance), pour le nom *vehicle* (véhicule) à son tour sont déterminés ses hyperonymes comme *agent* (agent), *equipement* (équipement), *means* (moyens), *medium* (moyenne) (cf. figure 3.15.).

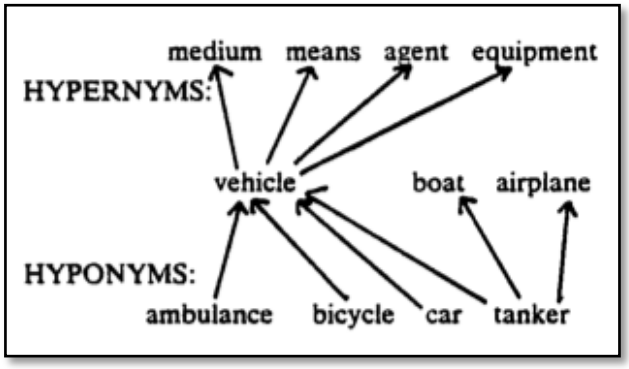


Figure 3.15. Exemple de structure hyperonymique

(Chodorow, Byrd, & Heidorn, 1985)

Toutefois, lors de la construction de l'arbre hiérarchique, l'utilisateur est consulté pour chaque nouveau lexème. Ce fait permet de désambiguïser certains lexèmes en leur attribuant seulement un sens mais pas de façon automatique.

Si les travaux présentés ci-dessus sont basés sur la reconnaissance des occurrences de mots-formes des lexèmes ainsi que leurs cooccurrences dans les définitions, les travaux suivants dans ce domaine de Jensen et Binot (1987), Ravin (1990), Montemagni (1992), Montemagni et Vanderwende (1992) se sont orientés vers l'extraction de l'information sémantique à partir de dictionnaires en ligne à l'aide d'analyseurs syntaxiques. Une analyse syntaxique des définitions dans ce cas est tout d'abord réalisée, l'information syntaxique étant ensuite utilisée pour améliorer la précision de l'identification des patrons lexico-syntaxiques.

Ainsi, les deux approches d'extraction d'information sémantique à partir des définitions des dictionnaires ont été appliquées dans la plupart des travaux de recherche dans ce domaine et semblaient donner des résultats prometteurs. Néanmoins, aucun travail présenté ci-dessus ne mentionne si les arbres hiérarchiques extraits ont été exploités par les applications de traitement automatique des langues.

Comme nous pouvons le constater, les travaux d'extraction des connaissances à partir de dictionnaires ont été très « à la mode » pendant les années 80-90 car les dictionnaires représentaient à l'époque une source d'information très importante et facilement accessible. Avec l'arrivée de l'Internet, les travaux se sont orientés plutôt vers l'extraction d'information à partir du Web qui donne accès à une riche source d'informations linguistiques. C'est pour cette raison qu'au cours des dernières années il

y a peu de travaux qui s'intéressent à la construction des hiérarchies sémantiques à partir de dictionnaires de langue. Les dictionnaires sont surtout utilisés comme source d'information pour désambiguïsation de sens (Navigli, 2009 ; Castillo, Real, Asterias, & Rigau, 2004) ou pour l'enrichissement des ontologies (Eckard, Barque, Nasr et Sagot, 2012 ; Navigli et Velardi, 2008). Par exemple, Navigli et Velardi (2008) proposent une méthodologie pour l'enrichissement automatique d'ontologies et l'annotation des documents avec des concepts et des relations ontologiques de domaine. Afin d'identifier les relations sémantiques ainsi que celles spécifiques à un domaine, ils traitent les définitions de glossaires disponibles dans un domaine donné en utilisant des expressions régulières définies à partir des propriétés de l'ontologie existante CIDOC (Doerr, 2003). Ils ont évalué les performances de leur système en extrayant les relations taxonomiques et non taxonomiques. Ainsi, ils déterminent que le système extrait les relations d'hyperonymie à partir les définitions du thésaurus d'Art et d'Architecture (AAT) avec une précision de 94,8 %.

Dans la section suivante, nous nous intéressons aux problématiques d'extraction de connaissances à partir de dictionnaires.

3.4.3. Questions posées par l'extraction de connaissances à partir de dictionnaires

L'avantage majeur de ressources lexicales existantes utilisées dans l'indexation et la recherche d'images comme WordNet, ConceptNet, etc. est le fait que l'information sémantique qu'ils contiennent est déjà explicite, ce qui n'est pas le cas d'un dictionnaire de langue. Pour notre part, l'une des difficultés rencontrées est d'expliciter les informations disponibles dans le dictionnaire *TLFi* afin de la rendre utilisable par les applications du TAL et, comme nous allons le voir, ce n'est pas une tâche simple.

Ainsi, Ide et Véronis (1994) montrent, dans une certaine mesure, les difficultés d'extraction de relations d'hyperonymie à partir des définitions lexicographiques. Pour cela, ils implémentent l'heuristique de Chodorow, Byrd et Heidorn (1985) pour cinq grands dictionnaires d'anglais. L'évaluation des arbres hiérarchiques extraits automatiquement en les comparant avec des arbres construits manuellement montre que 55-70 % de l'information extraite est en quelque sorte confuse. Ide et Véronis réalisent une analyse des problèmes rencontrés lors de la construction des arbres hiérarchiques à

partir des définitions des dictionnaires. Ils montrent que les hiérarchies extraites en utilisant des approches comme celle de Chodorow, Byrd et Heidorn (1985) ont de sérieuses lacunes et sont donc inutilisables dans des systèmes de TAL.

Nous présentons ci-dessous une synthèse des problèmes posés par les définitions lors de la construction automatique d'arbres hiérarchiques :

1. *L'information incomplète.*

Dans le dictionnaire, l'information est incomplète car le dictionnaire est le produit du travail de plusieurs lexicographes pendant de nombreuses années. Ainsi, il existe des incohérences au niveau des critères selon lesquels les hyperonymes donnés dans les définitions sont choisis. De plus, les restrictions sur la constitution des définitions (espace, lisibilité, etc.) font que certaines informations ne sont pas spécifiées dans les définitions ou sont laissées sous-entendues par d'autres parties de celles-ci.

2. *L'attachement est très élevé.*

Le problème le plus répandu dans les hiérarchies extraites automatiquement à partir du dictionnaire est que certains termes sont attachés à un niveau très haut de la hiérarchie. Par exemple, dans le dictionnaire *CED bottle* (bouteille) et *pan* (casserole) sont des *vessels* (récipients), tandis que *cup* (tasse) et *bowl* (bol) sont des *containers* (conteneurs). En outre, dans les autres dictionnaires *vessel* (récipient) est l'hyperonyme pour *cup* (tasse) et *bowl* (bol). Par conséquent, l'attachement de *cup* (tasse) et *bowl* (bol) au terme de niveau supérieure de la hiérarchie montre une incohérence au sein du dictionnaire *CED* (cf. figure 3.16.).

Toutefois, le problème de l'attachement aux termes de niveau supérieur de la hiérarchie résulte aussi d'un manque de termes à partir desquels choisir. Par exemple, dans le dictionnaire *LDOCE ladle* (louche) et *dipper* (louche) sont attachés au terme *spoon* (cuillère) or dans le dictionnaire *CED ladle* (louche) est attaché au *dipper* (louche) et celui-ci au terme *spoon* (cuillère) (cf. figure 3.16.). Ainsi, il n'est pas possible de déduire que *dipper* (louche) soit défini comme *ladle* (louche) puisque celui-ci n'est pas dans la définition de *dipper* (louche) dans le dictionnaire *LDOCE*. Ce phénomène contribue au fait que la hiérarchie extraite ne soit pas développée en profondeur mais représente plutôt une hiérarchie plate.

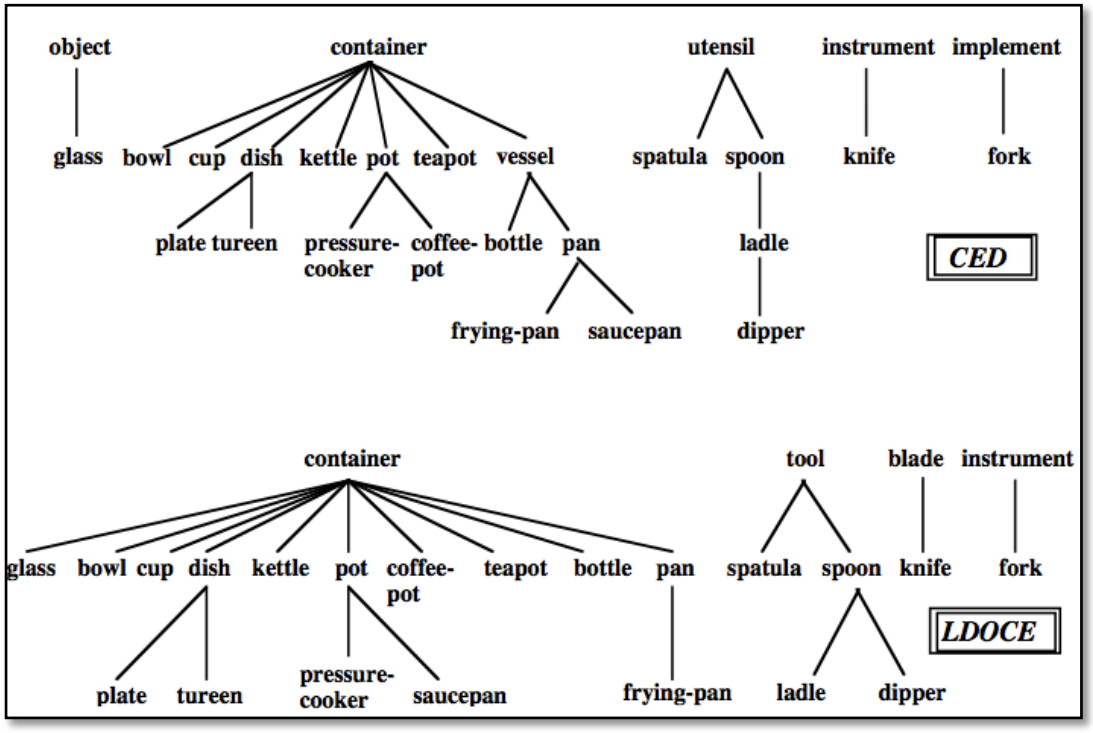


Figure 3.16. Hiérarchies extraites à partir des dictionnaires *CED* et *LDOCE*

(Ide & Véronis, 1994)

3. *Des hyperonymes absents.*

Dans 0-3 % des cas, les hyperonymes récupérés comme *part, piece, set,* etc. ne représentent pas en réalité les hyperonymes pour les noms définis. Dans ce cas, ils correspondent plutôt à d'autres relations sémantiques comme celle de méronymie, etc. Toutefois, le fait de les exclure ne permet pas de déterminer les « vrais » hyperonymes. Ainsi, les auteurs nomment ce phénomène des *hyperonymes absents*.

4. *Polyhiérarchie.*

Un autre problème découle du fait que, lors de l'élaboration de la définition, les lexicographes sont amenés à faire un choix afin de spécifier un seul niveau supérieur, alors que dans le monde réel, les concepts se chevauchent librement. Par exemple, *saucepan* (poêlon) peut être considéré comme *pot* (fait-tout) ainsi que *pan* (casserole). En termes de classes, *pot* (fait-tout) et *pan* (casserole) correspondent à des classes distinctes mais qui se chevauchent, et *saucepan* (poêlon) est un sous-ensemble de leur intersection (cf. figure 3.17.a).

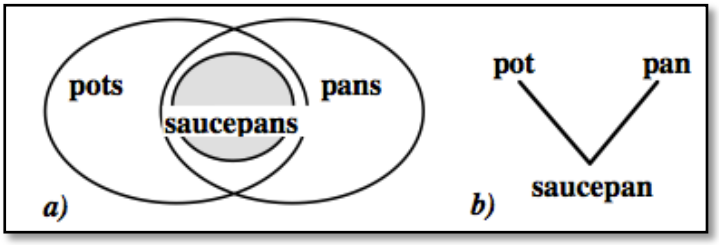


Figure 3.17. Exemple de polyhiérarchie

(Ide & Véronis, 1994)

On n'est donc plus en présence d'une hiérarchie stricte, mais plutôt d'une polyhiérarchie (*overlapping hierarchy*) où un terme peut appartenir à plusieurs termes génériques de niveau immédiatement supérieur (cf. figure 3.17.b). Ainsi, dans différents dictionnaires *saucepan* (poêlon) est défini soit comme *pot* (fait-tout), soit comme *pan* (casserole), mais les deux hyperonymes ne sont jamais utilisés dans la même définition, l'un des hyperonymes étant toujours manquant.

5. *Hyperonymes conjoints.*

Les termes des niveaux supérieurs de la hiérarchie sont très génériques et souvent moins clairement définis dans les dictionnaires, ce qui provoque une confusion dans la hiérarchie.

Ainsi, dans 7-10 % des définitions, les hyperonymes sont séparés par la conjonction *OR* (ou). Par exemple, *utensil* (ustensile) est défini dans le dictionnaire *CED* comme « an implement, tool or container… ». Si les premiers trois termes sont considérés comme des hyperonymes pour *utensil* (ustensile), alors la hiérarchie construite aura la forme présentée dans la figure 3.18.

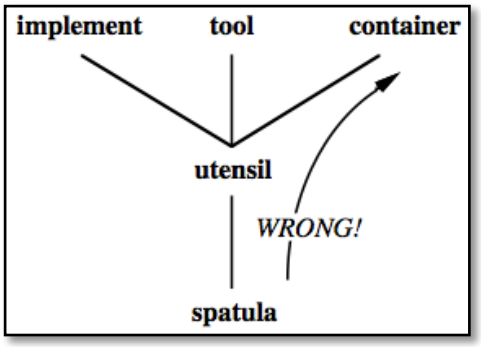


Figure 3.18. Hiérarchie problématique

(Ide & Véronis, 1994)

Ce type de hiérarchie pose des problèmes, car par exemple dans la figure 3.18., le nom *spatula* (spatule) défini comme *utensil* (ustensile) n'est pas un type de *container* (récipient).

6. Circularité.

La circularité est l'un des autres problèmes connus des définitions lexicographiques. Ainsi, 7-11 % des définitions utilisent un hyperonyme qui est lui-même défini circulairement. Les définitions circulaires produisent des hiérarchies contenant des boucles (cf. figure 3.19.a). Par exemple, *tool* (outil) est défini comme *implement* (instrument) et à son tour celui-ci est défini comme *tool* (outil).

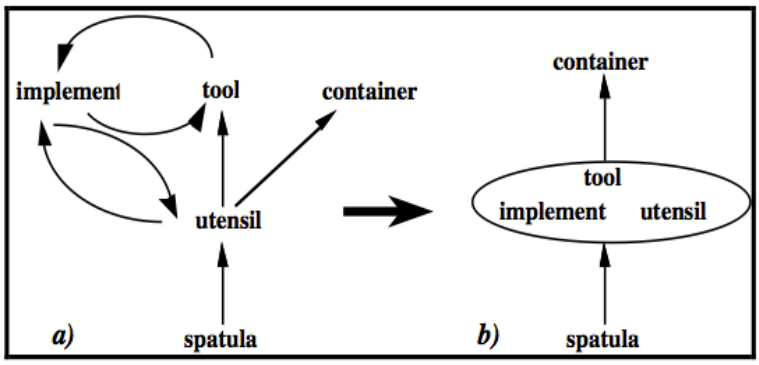


Figure 3.19. Taxonomie avec des boucles

(Ide & Véronis, 1994)

C'est pour cette raison qu'Amsler (1980) suggère de fusionner les concepts définis par la circularité et de les considérer comme synonymes (figure 3.19.b). Cependant, dans la plupart des cas, cette solution conduit à des résultats erronés, car tous les *tools* (outils) ne sont pas de *containers* (récipients).

Toutefois, Ide et Véronis (1995) mentionnent que des bases de connaissances de grande envergure pourront être construites avec l'intervention humaine en utilisant des informations provenant de plusieurs dictionnaires.

Loiseau, Gréa et Magué (2011) s'intéressent à la construction de graphes à partir d'un dictionnaire de langue en les comparant avec ceux construits à partir de dictionnaires de synonymes. Ainsi, ils différencient deux types de graphes qui peuvent être obtenus à partir du *TLFi* :

1. Graphe de définitions, qui connecte une entrée (nominale) à tous les noms de toutes ses définitions contenues dans l'article. Dans ce graphe une arête indique

la présence d'un mot-forme d'un lexème à l'intérieur de la définition d'un autre lexème.

2. Graphe de cooccurrences, qui connecte tous les noms qui apparaissent à l'intérieur d'une même définition. Dans ce graphe une arête indique le fait que deux mots-formes apparaissent simultanément dans au moins une définition.

Loiseau, Gréa et Magué (2011) décrivent aussi plusieurs problèmes spécifiques à l'utilisation d'un dictionnaire de langue ainsi qu'un certain nombre de choix qui doivent être faits dans le cas des multiples définitions, de l'hétérogénéité des relations sémantiques ou de la polysémie. Ils montrent que les graphes de cooccurrences résolvent les problèmes de polysémie et sont plus cohérents. Nous allons aussi nous interroger dans la section §4.1. sur des approches à adopter pour traiter certains de ces cas comme l'homonymie ou surtout quand la question se pose de savoir de quelles définitions il faut tenir compte lors de la construction de hiérarchies sémantiques — d'une seule qui se rapporte au sens « premier » d'un vocable ou de toutes sans faire de distinction. De plus, par rapport aux hiérarchies sémantiques qui explicitent les relations sémantiques entre les termes, les graphes relèvent seulement les principales propriétés lexicales et sémantiques du dictionnaire.

Ainsi, nous voyons que plusieurs problèmes doivent être résolus lors de la construction de hiérarchies sémantiques « parfaites » à partir des dictionnaires. Dans notre travail, pour éviter certains de ces problèmes, surtout celui de circularité, nous nous focalisons sur la construction de hiérarchies sémantiques pour l'article (bloc de texte) de chaque vocable en exploitant les définitions de ses lexèmes et non d'une seule taxonomie hyperonymique du *TLFi*. Ce fait nous permet d'expliciter toute l'information implicitement contenue dans les définitions de chaque article du *TLFi* en construisant des hiérarchies sémantiques ayant au maximum deux niveaux de profondeur[52] (cf. chapitre 5).

3.5. Conclusion

Dans ce chapitre, nous nous sommes intéressée dans un premier temps à l'utilisation des ressources lexicales dans le domaine de l'indexation et de la recherche d'images. Dans un premier temps, nous avons dressé un état de l'art en décrivant les ressources

[52] La profondeur étant mesurée en fonction de nombre de relations de type *is-a* dans la hiérarchie.

lexicales existantes ainsi que celles construites manuellement utilisées pour l'indexation et la recherche d'images. Étant donné que la construction manuelle de ressources lexicales est très coûteuse en temps et en ressources humaines, un grand nombre de travaux, d'un côté, exploitent les ressources lexicales existantes et, d'un autre côté, construisent automatiquement de nouvelles ressources. Puis, dans un deuxième temps, nous avons décrit les principales techniques d'extraction de connaissances et de construction de taxonomies à partir de textes (corpus de domaine, Web) et plus particulièrement à partir de dictionnaires. Nous avons vu que toutes ces techniques nécessitent soit un effort supplémentaire d'annotation manuelle d'un corpus d'entraînement pour déterminer des patrons lexico-syntaxiques, soit, un analyseur syntaxique ou une ressource externe comme WordNet pour structurer les relations extraites. De plus, ces techniques sont concentrées sur l'extraction des seules relations d'hyperonymie (genre prochain) à partir des définitions lexicographiques, tandis que les autres informations (différences spécifiques) des définitions ne sont pas exploitées lors de la construction de hiérarchies sémantiques. Notre objectif est d'extraire des connaissances exclusivement à partir d'un dictionnaire de telle manière qu'elles soient facilement utilisables à la fois pour l'enrichissement d'un thésaurus existant et pour l'indexation et la recherche d'images.

Même si la possibilité de construction de taxonomies à partir de dictionnaires a été prouvée (malgré les problèmes qui se sont posés), aucune utilisation effective de ces connaissances n'est mentionnée dans les systèmes de recherche d'images.

Les chapitres suivants de la thèse vont s'attacher à proposer une méthodologie de construction automatique de hiérarchies sémantiques à partir du *TLFi* et à montrer que l'exploitation automatique d'un grand dictionnaire institutionnel de langue comme le *TLFi* peut améliorer les processus d'indexation et de recherche d'images.

CHAPITRE 4

Extraction automatique d'informations pertinentes des définitions du *TLFi*

Sommaire

4.1. Définition du corpus de travail et son analyse 117
4.1.1. Analyse des définitions 119
4.1.2. Analyse des domaines 122
4.1.3. Analyse des locutions 124
4.2. Propositions de pondérations des noms dans le *TLFi* 124
4.2.1. Analyse des critères de pondération des noms dans une définition 125
4.2.1.1. Analyse de la fréquence du nom dans les définitions du *TLFi* 127
4.2.1.2. Analyse du nombre de définitions 128
4.2.1.3. Analyse de la position du nom dans la définition 128
4.2.1.4. Analyse de l'appartenance des noms aux expressions métalinguistiques 129
4.2.2. Normalisation du corpus de travail 130
4.2.2.1. Normalisation des domaines 130
4.2.2.2. Traitement des expressions métalinguistiques 132
4.2.3. Pondérations retenues pour déterminer l'importance des noms dans les définitions 133
4.2.3.1. Pondération locale 133
4.2.3.2. Pondération globale 134
4.2.3.3. Pondération par position 135
4.3. Évaluation de chaque facteur de pondération par rapport aux CC et CP du projet Definiens 136
4.4. Analyse de l'influence de chaque facteur de pondération sur le calcul de poids des noms 137
4.5. Conclusion 141

Chapitre 4. Extraction automatique d'informations pertinentes des définitions du *TLFi*

Dans ce chapitre, nous nous intéressons à l'extraction automatique d'informations pertinentes dans les définitions du *TLFi*. Ce chapitre commence par présenter une analyse de notre corpus de travail. Nous discutons ensuite des différents critères possibles pour pondérer l'importance des noms dans les définitions du *TLFi*. Puis nous présentons le travail de normalisation qui a été réalisé ainsi que les facteurs de pondération possibles pour la détermination de l'importance des noms dans les définitions du *TLFi*. Nous poursuivons ce chapitre par l'évaluation de chaque facteur de pondération proposé par rapport aux composantes centrales et périphériques du projet Definiens avant de conclure par la présentation de la pondération finale retenue.

4.1. Définition du corpus de travail et son analyse

L'intérêt que nous portons au *TLFi* concerne essentiellement ses définitions hyperonymiques et ses domaines de définitions. Le premier problème qui se pose alors est de définir l'information du *TLFi* qui sera utilisée dans nos recherches. Pour répondre à cette question, nous avons été amenée, dans un premier temps, à choisir notre corpus de travail. En analysant en parallèle le modèle de données du *TLFi* (cf. §1.3.) et l'organisation d'information dans les fichiers XML de SEMEME, nous avons opté pour l'utilisation de ce dernier. La principale raison justifiant notre choix est la lemmatisation préexistante des définitions du *TLFi* dans SEMEME. Toutefois, afin de pouvoir exploiter les données engendrées par SEMEME, nous avons été amenée à créer un modèle de données de celui-ci. Au terme de ces prétraitements nous avons effectué, dans un deuxième temps, une analyse des données de notre corpus que nous présenterons par la suite. La figure 4.1. présente la modélisation que nous avons faite à partir des données issues du projet SEMEME (cf. §1.7.1.).

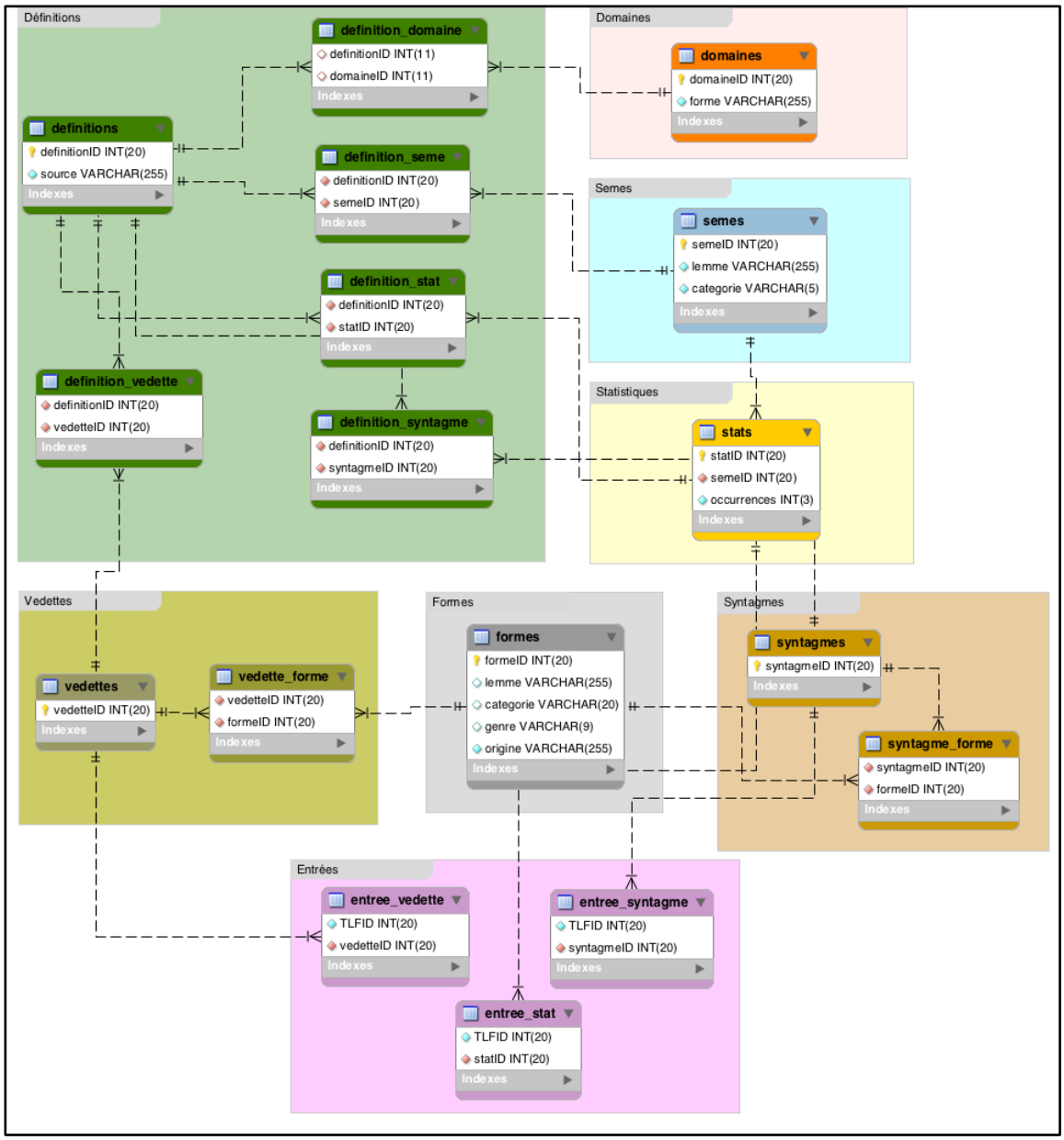


Figure 4.1. Modélisation réalisée à partir des données issues du projet SEMEME

Le *TLFi* est une source de données lexicales très riche. On y retrouve toute l'information synchronique et diachronique d'un lexème. Dans nos recherches, nous proposons l'utilisation de l'information du *TLFi*, dans un premier temps pour l'enrichissement du thésaurus actuel de Xilopix construit manuellement, et dans un deuxième temps nous démontrons que ce thésaurus améliore l'indexation et la recherche d'images.

Nous avons réalisé une analyse des données de notre corpus de travail afin de déterminer quelles informations du *TLFi* pouvaient être utilisées dans nos recherches.

Dans cette étude, nous nous sommes limitée exclusivement aux lexèmes nominaux. Les unités de la catégorie substantif sont en effet les seules capables de dénoter n'importe quel type de référent (les adjectifs et les verbes, eux, ne peuvent pas dénoter un être ou un objet, par exemple) et sont, par conséquent, les plus pertinentes à exploiter pour enrichir le thésaurus existant.

L'hypothèse sur laquelle nous nous appuyons est qu'un lexème donné peut être caractérisé par les noms qui apparaissent dans sa définition au sein du *TLFi*.

4.1.1. Analyse des définitions

Le tableau 4.1. présente l'information sur les données de notre corpus de travail. Compte tenu du fait que, dans l'article d'un vocable, les lexies sont représentées par des lexèmes et des locutions, nous présentons séparément l'information les concernant afin de pouvoir mieux saisir les différences entre ces données.

Ainsi, nous avons déterminé que 78 % des définitions du *TLFi* sont directement rattachées aux lexèmes et 22 % d'entre elles le sont aux locutions. Une remarque s'impose : nous avons constaté que, dans les données de SEMEME, 0,16 % des définitions des lexèmes et 0,31 % des définitions des locutions ne contiennent pas de lemmes. Ceci est dû aux limites (non lemmatisation de définitions) et erreurs qui persistent dans les données issues du projet SEMEME. Par exemple, les mots-formes de la définition « Serre-joint » du lexème SERGENT pour le domaine de la *menuiserie* n'ont pas été lemmatisés.

	Total	Lexèmes	Locutions
Lemmes différents	38 617	35 468	22 779
Définitions	265 475	206 052	59 423
Domaines	7 786[53]	6 851	3 341

Tableau 4.1. Information sur des données du corpus de travail SEMEME

[53] Le grand nombre de domaines s'explique par le fait que certains domaines apparaissent en combinaison avec d'autres comme « Math., Arithm. », « Math., Géom. », « Math.mod », etc. La normalisation des domaines et leur hiérarchisation sont présentées dans la section §4.2.2.1.

Nous avons aussi déterminé qu'en moyenne une définition peut contenir 6 lemmes, avec un maximum de 111 lemmes pour un lexème et de 64 lemmes pour une locution (cf. figure 4.2.).

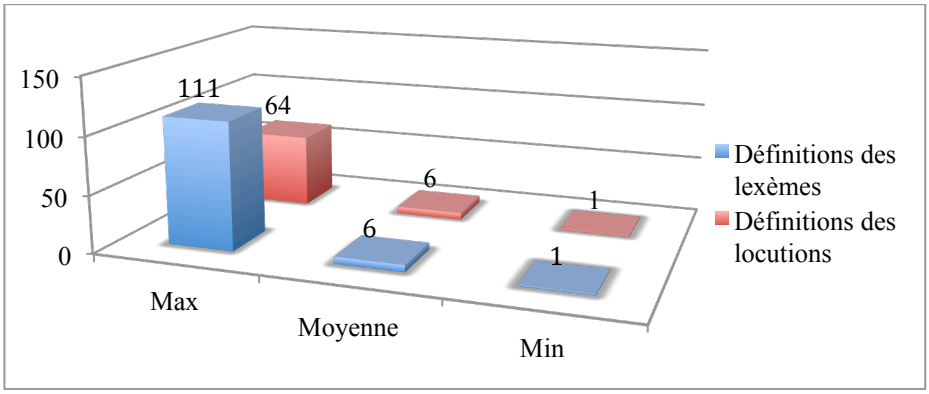


Figure 4.2. Répartition des lemmes dans les définitions du *TLFi*

Les lemmes des définitions du *TLFi* appartiennent aux différentes catégories grammaticales. Toutefois, près de la moitié (50,82 %) des lemmes des définitions des lexèmes appartiennent à la catégorie substantif (cf. figure 4.3.). Cela nous permet de valider notre première hypothèse, à savoir qu'une lexie donnée peut être caractérisée par les noms (ou substantifs) qui apparaissent dans sa définition. Nous remarquons aussi que le taux des lemmes pour chaque catégorie grammaticale dans les définitions des lexèmes et des locutions est presque égal. Par la suite dans nos recherches, nous utilisons donc exclusivement les substantifs[54] des définitions du *TLFi*.

[54] Nous avons également retenu les lemmes des catégories Np qui représentent des noms propres, utilisés ensuite dans les hiérarchies sémantiques construites pour représenter la relation d'association de lieu (TA_location).

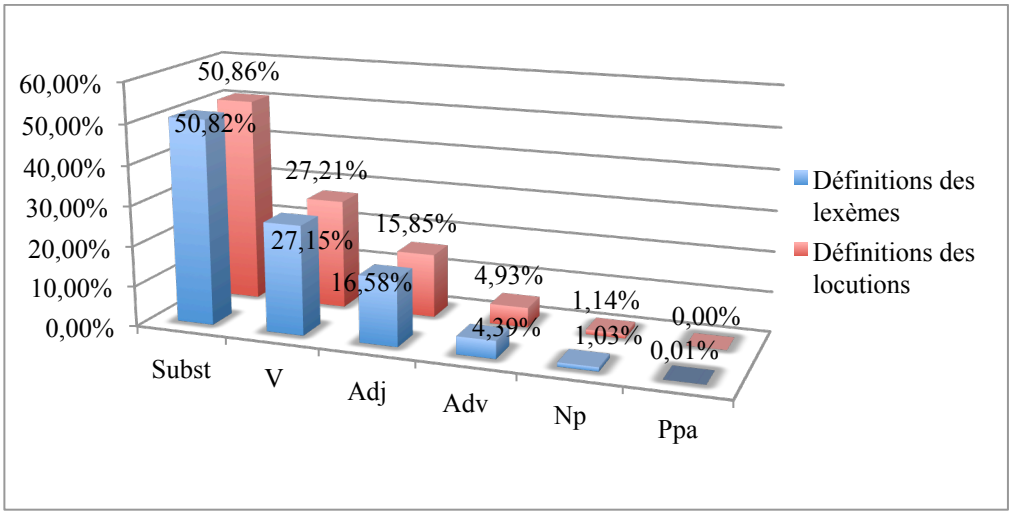


Figure 4.3. Taux des lemmes de différentes catégories grammaticales dans les définitions du *TLFi*

Nous avons aussi déterminé qu'en moyenne 70,04 % des définitions des lexèmes contiennent trois noms (cf. figure 4.4.)[55].

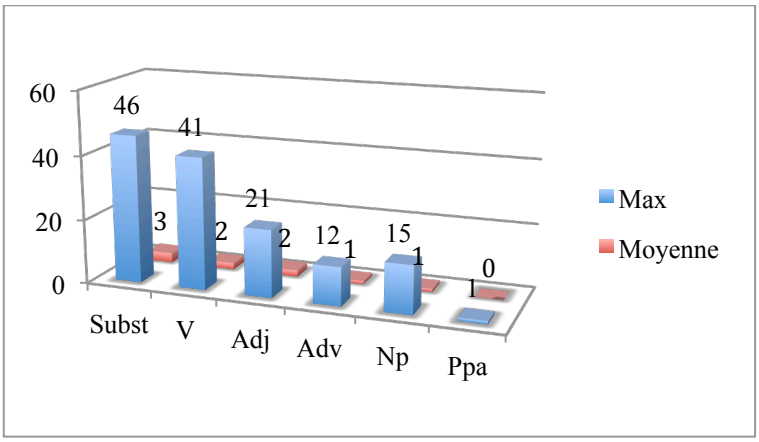


Figure 4.4. Répartition des lemmes des définitions du *TLFi* entre les différentes catégories grammaticales

Nous pouvons donc conclure qu'un lexème donné peut être caractérisé en moyenne par trois noms.

[55] Il existe aussi dans le *TLFi* des définitions qui ne contiennent pas de noms. Ce sont surtout des définitions des lexèmes de catégories grammaticales comme adjectif, verbe, etc.

4.1.2. Analyse des domaines

Dans le *TLFi*, les domaines associés aux définitions indiquent le domaine pour lequel le sens d'une lexie est valide. Les domaines des définitions signifient que la lexie donnée a un sens particulier dans ce domaine. Toutefois, seulement 33,05 % des définitions des lexèmes et 47,30 % des définitions des locutions sont rattachées à un ou plusieurs domaines. Si une définition n'est liée à aucun domaine, la lexie correspondante est non spécialisée (elle relève de la langue générale) ; la définition peut donc être appliqué à tous les domaines. Par exemple, nous pouvons affirmer que le sens du lexème LION[56] comme *mammifère* peut être appliqué dans les autres domaines : astronomie, héraldique, astrologie, etc. Même si le lexème LION dans le domaine d'*astrologie* représente un *signe du zodiaque*, l'image représentative de ce signe continue de faire référence au sens premier de *mammifère*. Nous proposons donc d'attribuer aux définitions sans domaine le domaine nommé *générique*.

Dans le *TLFi* les définitions des lexèmes d'un vocable sont classées selon une hiérarchie de niveaux. Grâce aux marques de plan, aux indicateurs et à l'information entre crochets de la définition, nous pouvons percevoir la hiérarchie des niveaux de définition, qui n'est pas toujours la même[57]. Hélas, dans le modèle de données que nous exploitons, nous ne disposons pas de cette information qui permet de hiérarchiser les définitions. Le seul critère permettant de grouper les définitions reste donc le domaine. Toutefois, cela pose aussi un problème, car nous avons déterminé que, dans les données de SEMEME une définition du *TLFi* peut référer jusqu'à quatre domaines différents, et qu'au sein d'un même article peut exister une hiérarchie de domaines et sous-domaines. Par exemple, les domaines des définitions des lexèmes du vocable SOURIS sont organisés selon la hiérarchie suivante (nous fournissons entre parenthèses un extrait d'une des définitions correspondantes à chaque domaine pour mieux préciser les choses) :

[56] La définition du lexème LION « Mammifère carnivore, de la famille des Félidés, de forte taille, caractérisé par sa face large, sa crinière touffue, son tronc et ses membres trapus, son pelage fauve, et vivant à l'état sauvage surtout en Afrique. » dans le *TLFi* n'est liée à aucun domaine.

[57] Dans les articles de certains vocables, les définitions précédées des indicateurs d'emploi, des domaines, apparaissent aux niveaux plus hauts (I, II) de la hiérarchie, mais pour d'autres entrées, les définitions apparaissent aux niveaux inférieurs.

A. –ZOOLOGIE (Petit mammifère rongeur omnivore)

c) HISTOIRE (Souris grise. Femme soldat de l'armée allemande pendant la guerre de 1939-1945*)*

C. –SC. ET TECHN.

1. *BOTANIQUE (Oreille de souris*. Synon. pop. de myosotis, piloselle *; Queue de souris*. Espèce de Renonculacée*)*
2. *DEFENSE (Pas de souris*. *, «* Escaliers étroits qui font communiquer les différents ouvrages d'une fortification *»)*
3. *ELECTRON., INFORMAT* (Périphérique d'entrée relié à l'ordinateur)
4. *MEDECINE (Souris articulaire*. „Petit fragment osseux ou cartilagineux)
5. *TECHNOL. (Dents de souris*. Entaillures très fines sur une pièce mécanique)
6. *TYPOGR. (morceau de plomb suspendu à une ficelle et qui pend sur le papier pour l'empêcher de glisser dans les pinces du cylindre)*
7. *ZOOLOGIE*
   a) *ANAT., BOUCH., ART CULIN. (*[Chez le mouton] Muscle charnu situé à l'extrémité du gigot)
   b) *ICHTYOL. (Souris (de mer)*. [Dénom. pop. de différents poissons, notamment la baliste)

Compte tenu de la hiérarchie des domaines, les définitions appartenant aux sous-domaines (ex. *7.a) ANAT., BOUCH., ART CULIN.)* sont rattachées aux domaines supérieurs (ex. *7. ZOOLOGIE* et *C. – SC. ET TECHN.*). Dans ce cas, nous convenons que la définition doit être traitée individuellement pour chaque sous-domaine, *ANAT., BOUCH., ART CULIN*.

Comme nous pouvons le remarquer dans la hiérarchie des domaines ci-dessus les dénominations des domaines ne sont pas normalisées et sont souvent abrégées (ex. *ANAT., SC., CULIN.*). Un même domaine peut donc apparaître sous plusieurs formes. Cela représente un inconvénient majeur dans le cas où l'on voudrait par exemple rassembler toutes les hiérarchies sémantiques construites pour les lexèmes des vocables d'un certain domaine. Une normalisation manuelle de domaines s'imposait donc (cf. §4.2.2.1.).

Dans le *TLFi*, plusieurs vocables peuvent être homonymes, appartenir aux mêmes signifiants mais avoir des sens entièrement différents. Par exemple, les vocables CÂPRE[1] (« Bouton à fleurs du câprier ») et CÂPRE[2] (« [Aux Antilles françaises] Personne issue du croisement de nègre et de mulâtre ») sont des vocables homonymes. Afin de traiter les définitions des lexèmes des vocables homonymes, nous les regroupons aussi par domaine si les vocables sont de la même catégorie grammaticale (le genre pouvant être différent).

4.1.3. Analyse des locutions

Nous avons déterminé que 20,72 % des articles des vocables du *TLFi* définissent des locutions (appelées dans le *TLF* « syntagmes figés[58] »). Dans l'exemple du vocable SOURIS présenté à la page précédente, les syntagmes figés sont *Souris grise, Oreille de souris, Queue de souris, Souris articulaire, Dents de souris*, etc. L'un des problèmes liés à l'utilisation des locutions réside dans le fait qu'elles ne sont pas normalisées et peuvent contenir, dans leur structure, des signes de ponctuation particuliers de mise en facteur (ex. *Lion (d'Amérique/du Pérou,* etc.). Cela ne permet pas leur reconnaissance par les outils de TAL. Afin de les rendre exploitables, un travail de normalisation des locutions devrait être effectué manuellement[59]. Compte tenu du fait qu'il existe au total 40 000 locutions à normaliser, et que seulement 20,72 % des articles des vocables traitent les locutions, nous avons décidé de ne pas les utiliser lors de nos recherches. Ainsi, par la suite nous n'exploitons que les définitions des lexèmes. Ceci dit les méthodologies que nous proposons pourraient sans problème s'appliquer aux locutions dès que ces dernières seront normalisées au sein du *TLFi*.

4.2. Propositions de pondérations des noms dans le *TLFi*

L'analyse des données de notre corpus de travail nous a permis de valider l'hypothèse qu'un lexème donné peut être caractérisé par les substantifs inclus dans sa définition. Nous avons déterminé que la moitié des lemmes des définitions sont des substantifs et qu'en moyenne un lexème peut être caractérisé par trois substantifs issus de sa définition.

Dans nos recherches, nous avons décidé d'exploiter tous les substantifs (ou noms) des définitions des lexèmes pour construire des hiérarchies sémantiques à partir du *TLFi*. Afin de pouvoir différencier les noms des définitions du *TLFi* d'un lexème, nous avons besoin d'une formule qui permette de les pondérer selon certains critères en faisant l'hypothèse que le nom de poids maximal devrait représenter l'information la plus pertinente au sein des définitions et ainsi être un bon candidat hyperonyme pour le

[58] Un syntagme que le locuteur utilise comme un tout « préconstruit » dans la langue (Polguère, 2008, p.54).
[59] Un projet d'homogénéisation des locutions du *TLFi* est actuellement en cours à l'ATILF.

lexème donné. On peut d'ailleurs remarquer que tous les travaux sur l'extraction de connaissances à partir de dictionnaires (cf. §3.4.2.) s'appuient sur une hypothèse de même type. Pour construire une telle formule, adaptée aux définitions du *TLFi* nous avons été amenée à analyser un ensemble de critères susceptibles de jouer un rôle lors de la pondération des noms dans ces définitions.

4.2.1. Analyse des critères de pondération des noms dans une définition

Dans cette section, nous allons analyser les critères observables selon lesquels les noms des définitions du *TLFi* pourraient être pondérés. Dans un premier temps, nous avons listé et analysé sept critères :

1. La fréquence du nom dans une définition d'un lexème pour un domaine donné ;
2. La fréquence du nom dans toutes les définitions des lexèmes d'un vocable pour un domaine donné ;
3. La fréquence du nom dans l'ensemble de toutes les définitions des lexèmes des vocables pour un domaine donné ;
4. Le nombre de définitions des lexèmes pour un domaine donné qui contiennent le nom ;
5. La position du nom dans la chaîne de caractères de la définition ;
6. L'appartenance du nom aux expressions métalinguistiques de la classe d'opposition ou de négation constituée par nous et la position des autres noms pas rapport aux expressions de cette classe ;

Nom de la classe	Expressions métalinguistiques
Classe des indices d'opposition ou de négation	Par opposition au (aux)
	Sans
	Pas avec
	Non
	Ni
	Absence de/Manque de

Tableau 4.2. Expressions métalinguistiques de la classe d'opposition ou de négation

7. L’appartenance du nom aux expressions métalinguistiques de la classe générique.

Nom de la classe	Expressions métalinguistiques
Classe générique	Action de
	Manière de
	Fait de
	Partie de
	Ensemble de
	Sorte de
	Genre de
	Sous-genre de
	Vue de
	Forme de
	Famille de
	Espèce de
	Ordre de
	Sous-ordre de
	Variété de
	Nom de
	Représentation de
	Suite de
	Lot de
	Série de
	Groupe de
	Réunion de

Tableau 4.3. Expressions métalinguistiques de la classe générique

Pour réaliser nos analyses, nous avons implémenté chaque critère de pondération pour tous les noms des définitions du *TLFi* et présentons les analyses qui ont été faites dans les sections suivantes.

4.2.1.1. Analyse de la fréquence du nom dans les définitions du *TLFi*

En analysant les fréquences des noms, nous avons constaté que les critères 1-3 sont importants et influencent fortement le poids d'un nom dans une définition. Ainsi, dans 7 113 définitions du *TLFi*, un même nom apparaît plusieurs fois dans une définition (critère 1). Par exemple, dans la définition du lexème DÉTERMINANT[60] on compte quatre occurrences du lemme du nom *mot*, ce qui montre bien son importance. L'occurrence du nom dans la définition dépend aussi de la construction de la définition. Dans les définitions qui contiennent deux énoncés explicatifs séparés par un point virgule (ex. « Surface plane et peu épaisse de quelque chose ; ce qui constitue une telle surface » (PLATEAU[1])) ou dans les énoncés énumératifs séparés par une virgule (ex. « Bouton d'un appareil de radio, bouton de sonnette, interrupteur » (PITON)), le nom qui est plus important (ex. *surface, bouton*) apparaît dans les 2 énoncés. Les occurrences multiples d'un nom dans une définition devraient donc le favoriser par rapport aux autres noms. Nous proposons donc que le poids d'un nom augmente proportionnellement à son nombre d'occurrences dans la définition, ce qui permettrait de bien renforcer ces noms comme candidats hyperonymes possibles pour un lexème donné.

L'analyse de la fréquence du nom dans toutes les définitions des lexèmes d'un vocable pour un domaine donné (critère 2) a permis d'émettre l'hypothèse que plus le nom est fréquent dans les définitions des lexèmes d'un vocable pour un domaine donné, plus le poids de ce nom devrait augmenter par rapport aux autres noms. Par exemple, dans les définitions du vocable HERBE pour le domaine *générique,* le nom *plante* apparaît trois fois, ce qui montre que ce nom est assez important et pertinent pour le vocable donné.

Le poids d'un nom peut aussi varier selon sa fréquence dans l'ensemble de toutes les définitions des lexèmes des vocables pour un domaine donné (critère 3). Plus un nom apparaît dans l'ensemble des définitions d'un domaine, plus il paraît caractéristique

[60] La définition du *TLFi* pour le lexème DÉTERMINANT est « Mot (ou groupe de mots) qui, placé à côté d'un autre mot ou groupe de mots (ou déterminé) a pour fonction de le déterminer, c'est-à-dire d'en préciser le genre, le nombre, éventuellement le sens contextuel et par là de limiter son extension. ».

pour ce domaine. Par exemple, dans les définitions du domaine *droit*, les noms qui possèdent le plus d'occurrences sont *acte, personne, loi,* etc.

4.2.1.2. Analyse du nombre de définitions

Le nombre de définitions des lexèmes pour un domaine donné qui contiennent le nom étudié (critère 4) est sensiblement égal à la fréquence du nom dans l'ensemble de toutes les définitions des lexèmes des vocables pour un domaine donné (critère 3). Toutefois, nous avons remarqué que cette fréquence est un peu plus élevée, car un nom peut apparaître plusieurs fois dans les définitions pour un domaine donné. Par exemple, le nom *justice* apparaît 85 fois dans les définitions pour le domaine *droit* et il existe 83 définitions de lexèmes qui contiennent le nom *justice* pour plus de 2 600 substantifs ayant au moins une définition particulière dans le domaine du *droit.*

4.2.1.3. Analyse de la position du nom dans la définition

La position du nom dans la définition paraît assez importante (critère 5). Nous la déterminons en calculant la position de la première occurrence du nom dans la chaîne de caractères de la source de la définition[61]. Dans la plupart des cas, les noms en positions 1 représentent la tête des genres prochains du lexème défini pour le domaine donné, et ceux en position assez élevée représentent plutôt des caractéristiques spécifiques.

Par ailleurs, il existe des définitions où le nom en position 1 est suivi d'une préposition de type *de, de la, des.* Par exemple, dans le cas du lexème FOURCHETTE[62], se pose la question de savoir s'il faut considérer comme candidat hyperonyme *ustensile* ou le syntagme *ustensile de table*. Il s'agit dans ce cas plutôt du syntagme. Néanmoins, pour notre part, nous considérons que seul le nom *ustensile* représente le candidat hyperonyme du lexème FOURCHETTE, parce que la détermination automatique des syntagmes candidats hyperonymes avec exactitude est

[61] Compte tenu du fait qu'on a déjà la liste de tous les noms de chaque définition, nous avons décidé de calculer la position de la première occurrence du nom dans la chaîne de caractères et non le numéro du nom dans la définition.

[62] Une des définitions pour ce lexème dans le *TLFi* est « Ustensile de table en forme de petite fourche à deux, trois ou quatre dents, dont on se sert pour piquer les aliments ».

en fait très problématique étant donné que les modes de présentation des formes des locutions sont très différentes (ex. Souris (de mer), Mettre (qqc . ou qqn) ablativo tout (ou tous) en un tas, (Branche d') acacia, (Être) sans âge, agent secret, Agent de change, etc.).

4.2.1.4. Analyse de l'appartenance des noms aux expressions métalinguistiques

Chaque type de définition (nominale, verbale, adjectivale, adverbiale) est lié à des expressions métalinguistiques spécifiques caractéristiques de ce type. Dans les définitions nominales, on rencontre souvent les expressions métalinguistiques de type : *action de, manière de, fait de, partie de, manque de,* etc. Pour les définitions verbales sont caractéristiques des locutions verbales et des locutions comprenant les verbes fonctionnels *faire, laisser,* etc. Quant aux définitions adjectivales, elles ont plus souvent la structure *Qui + être + Adj., Qui + verbe, propre à, relatif à, se dit de,* etc.

Au sein du *TLFi* il existe 7 362 définitions qui commencent par le nom *action*. Dans ces définitions, *action* représente plutôt l'expression métalinguistique de type *action + de* suivie de noms ou de verbes.

Le nom *partie* apparaît, pour sa part, comme le premier nom dans 1 567 définitions. Ces définitions ont la structure de type *partie + de + nom* où il convient de considérer *partie de* comme une expression métalinguistique. Cette expression indique une relation méronymique : les noms qui suivent l'expression *partie de* désignent le tout duquel le lexème donné fait partie. Par exemple, dans la définition « Partie de l'atmosphère comprise entre deux niveaux et dans laquelle les nuages de certain genre se présentent normalement » (ÉTAGE*)*, *partie de* indique que le référent du lexème ÉTAGE fait partie de l'*atmosphère*.

Si la classe des événements ou la classe des transports est définissable, la classe des actions, des manières, est plus difficile à définir. Les candidats hyperonymes de type *action, manière, caractère,* etc. sont trop génériques pour représenter des concepts. C'est pourquoi nous avons créé une liste de toutes les expressions métalinguistiques rencontrées dans les définitions du *TLFi*. Les noms inclus dans ces expressions ne seront pas pris en compte comme candidats hyperonymes pour les lexèmes. Dans ce cas, il s'agit le plus souvent de définitions par approximation.

Ainsi, suite à nos études sur les définitions du *TLFi*, il apparaît que c'est plutôt la position des noms dans les définitions par rapport à ces expressions métalinguistiques qui est plus importante à prendre en compte. Un candidat hyperonyme d'un lexème se trouve en effet souvent après une expression métalinguistique.

4.2.2. Normalisation du corpus de travail

Dans la section §4.1.2. nous avons vu que les domaines du *TLFi* n'étaient pas normalisés, ce que nous ne permettait pas par exemple de calculer le nombre exact de définitions pour un domaine donné. Par ailleurs, nous avons vu aussi qu'un traitement des noms appartenant aux expressions métalinguistiques pourrait nous aider à déterminer plus facilement les candidats hyperonymes d'un lexème dans ses définitions.

La section suivante sera consacrée à la présentation des normalisations des données que nous avons faites.

4.2.2.1. Normalisation des domaines

Comme nous l'avons déjà mentionné, le problème majeur des domaines du *TLFi* est qu'ils ne sont pas normalisés. Un même domaine peut être indiqué de différentes manières (ex. *MAR., MARINE*), ce qui induit un biais lors du calcul de la fréquence d'un nom dans un ensemble de définitions pour un domaine donné.

Nous avons donc effectué une normalisation manuelle des domaines du *TLFi*. Cela nous a permis de réduire considérablement le nombre de domaines, de 7 786 à 758. En parallèle à cette normalisation manuelle, nous avons procédé à une hiérarchisation des domaines, obtenant ainsi un thésaurus des domaines du *TLFi*.

Pour réaliser cette normalisation des domaines, nous avons utilisé la documentation sur le thésaurus des domaines techniques du *TLF* qui contient tous les domaines et les sous-domaines utilisés lors de la rédaction des définitions. Le processus de normalisation consiste à attribuer à un domaine sa forme normalisée de la manière suivante :

« bactériol. », « bactériologie. » → BACTÉRIOLOGIE

« MAR. », « MARINE. » → MARINE

En même temps, nous avons effectué un travail de hiérarchisation des domaines en créant une base de données dans laquelle on attribue à un domaine son domaine père. La hiérarchie des domaines a été construite en reprenant celle qui existait déjà pour le *TLFi*.

La normalisation des domaines a été effectuée en suivant les règles suivantes :

1. Si, dans les annotations, un domaine est séparé d'un deuxième domaine par une virgule ou une conjonction, et qu'il s'agit de domaines différents sans lien hiérarchique entre eux, plusieurs domaines sont attribués à la définition.

« MAR., AVIAT. » → MARINE, AVIATION

« ANAT., ZOOL. »→ ANATOMIE, ZOOLOGIE

2. Si, dans les annotations, un domaine est séparé d'un deuxième domaine par une virgule ou une conjonction et que l'un des domaines représente le domaine père du second, nous attribuons la définition seulement au domaine du plus bas niveau de la hiérarchie des domaines.

« PHYS., OPT. » → OPTIQUE

« MÉD., BIOL. » → MÉDECINE

3. Si le domaine spécifique existe dans la hiérarchie constituée, nous l'attribuons à la définition, sinon nous lui attribuons le domaine de la hiérarchie qui l'inclut (père virtuel).

« SYNTAXE » → LINGUISTIQUE

« NAVIG. MAR. » → NAVIGATION

Lors de cette hiérarchisation des domaines, il ne s'agissait pas de construire une nouvelle hiérarchie ou de faire une réorganisation de celle existante. Néanmoins étant donné que ce thésaurus des domaines servira pour l'enrichissement du thésaurus Xilopix, qui à son tour sera utilisé pour l'indexation des images, certaines modifications ont été faites dans la hiérarchie. Nous avons ajouté[63] dans la hiérarchie certains domaines spécifiques (ex. *antiquité grecque, iconographie,* etc.) jugés susceptibles d'être intéressants pour l'indexation d'une image. Des domaines plus globaux comme

[63] Nous précisons qu'il s'agit d'ajouter des domaines présents dans le *TLFi* mais absents dans la hiérarchie.

science de la nature, sciences humaines ont aussi été ajoutés pour regrouper des sous domaines en domaines plus génériques. Pour réaliser cette hiérarchisation, nous avons également consulté le thésaurus MOTBIS[64].

4.2.2.2. Traitement des expressions métalinguistiques

L'analyse des expressions métalinguistiques a montré que la plupart des expressions métalinguistiques sont situées en tête des définitions et sont souvent suivies par le candidat hyperonyme du lexème. Cet état de fait pouvait avoir une incidence néfaste pour la détermination des candidats hyperonymes qui devra s'appuyer, en partie du moins, sur leur position dans la définition. La solution que nous avons retenue pour régler ce problème est d'éliminer les expressions métalinguistiques en créant une liste de mots-formes à filtrer. Toutefois, pour pouvoir conserver[65] cette information lors de nos recherches, nous avons choisi de les ajouter comme propriété du nom qui les suit de la manière suivante :

Expression métalinguistique + Nom => Nom (Expression métalinguistique)

En revanche, nous avons décidé de ne pas traiter de cette manière les expressions métalinguistiques de type *fait*, *action*, *manière* (*façon)*. Nous considérons en effet que ces expressions permettent de réaliser une classification des noms en FAIT, ACTION et MANIÈRE. Par opposition aux expressions comme *famille de, espèce de,* etc*.,* toujours suivies dans les définitions par un nom, ces expressions métalinguistiques (EM) sont en effet utilisées dans les constructions de genre *EM +V., EM + Adv*. Ce traitement des expressions métalinguistiques permet de remonter au premier rang les noms qui les suivent. Par exemple, pour le lexème LION nous déterminons comme candidat hyperonyme le nom *mammifère (famille des)* au lieu du nom *famille*.

[64] http://www.cndp.fr/motbis/

[65] Nous jugeons important de garder les expressions métalinguistiques afin de pouvoir spécifier le nom qui les suit. Par exemple, *continent* ne peut pas être considéré comme *monde* quand en réalité il représente *une partie de ce monde*.

4.2.3. Pondérations retenues pour déterminer l'importance des noms dans les définitions

En analysant plusieurs critères susceptibles de jouer un rôle lors de la pondération des noms, notre approche, basée sur l'analyse de la structure, de la taille et du métalangage spécifique des définitions du dictionnaire, nous a permis de définir trois facteurs de pondérations des noms dans une définition : la pondération locale, la pondération globale et la pondération par position.

Comme nous l'avons vu au chapitre 3 (cf. §3.4.), la plupart des travaux qui s'intéressent à la construction de hiérarchies sémantiques à partir de dictionnaires utilisent en partie des patrons lexico-sémantiques définis manuellement pour extraire la relation d'hyperonymie à partir de définitions lexicographiques. Pour notre part, nous tenons à exploiter le maximum d'information sémantique contenue dans les définitions lexicographiques, sans nous limiter à la seule extraction des hyperonymes. Cela nous permettra, d'une part, d'enrichir le thésaurus Xilopix avec de nouvelles relations et, d'autre part, d'enrichir les descriptions textuelles d'images avec les autres termes des définitions afin de déterminer ensuite les meilleurs termes d'indexation. Compte tenu de notre domaine d'application spécifique (indexation et recherche d'images), nous optons plutôt pour une formule de pondération permettant de mesurer l'importance des noms dans les définitions des lexèmes que la définition d'un processus d'extraction des seuls hyperonymes à l'aide des patrons lexico-syntaxiques.

Les pondérations que nous proposons sont inspirées de schéma de pondération TD-IDF présenté dans la section §2.1.1. mais adaptées à nos propres objectifs. Nous allons maintenant présenter les facteurs retenus pour cette formule de pondération avant d'évaluer l'importance de chaque facteur et de conclure par la pondération finale retenue.

4.2.3.1. Pondération locale

La pondération locale d'un nom représente le nombre d'occurrences du nom dans une définition, normalisé par la somme des nombres d'occurrences de tous les noms de la définition. La normalisation du nombre d'occurrences d'un nom permet la comparaison de deux définitions de longueurs différentes. Ainsi, nous ne pénalisons pas

les définitions de petite taille dans lesquelles un nom peut se rencontrer moins de fois que dans les définitions de grande taille. La pondération locale permet d'évaluer l'importance d'un nom dans une définition.

Soit une définition *d* et un nom *t*, alors le facteur fréquentiel *tf* normalisé est :

$$tf = \frac{N(t)}{\sum_i N(t_i)} \tag{4.1.}$$

où

$N(t)$: fréquence d'un nom *t* dans la définition *d,*

$\sum_i N(t_i)$: somme des fréquences de tous les noms t_i dans la définition *d.*

4.2.3.2. Pondération globale

La pondération globale permet d'apprécier l'importance d'un nom dans la collection des définitions pour un domaine donné. Elle représente le nombre de définitions qui contiennent le nom, normalisé par le nombre total de définitions dans la collection pour un domaine donné.

Soit une définition *d*, une collection de définitions *c* et un nom *t*, alors le facteur *df* (*definition frequency*) normalisé est :

$$df = \frac{N(d_t, c)}{N(d, c)} \tag{4.2.}$$

où

$N(d_t, c)$: nombre des définitions dans la collection pour un domaine donné qui contiennent le nom,

$N(d, c)$: nombre des définitions dans la collection pour un domaine donné.

Contrairement à la pondération *IDF* (cf. §2.1.1.) qui donne plus de poids aux noms rares, notre pondération favorise les noms fréquents. Nous considérons en effet que les noms fréquents dans les définitions pour un domaine particulier représentent des classificateurs possibles (genre prochain) et que les noms moins fréquents correspondent plutôt à des spécifications de ces définitions. Par exemple, les noms les plus fréquents dans le domaine *botanique* sont *plante, fleur, arbre* parce qu'ils sont les plus utilisés en tant qu'hyperonymes ou classificateurs pour définir les lexèmes de ce

domaine tandis que les noms *ornementation, couleur, rameau* représentent des spécifications au sein de ce domaine.

4.2.3.3. Pondération par position

Dans la plupart des définitions du *TLFi*, nous avons vu que les noms en première position représentent de bons candidats hyperonymes des lexèmes définis. Nous proposons donc un facteur de pondération lié à la position du nom dans la définition pour rendre compte de la dépendance existante au sein des définitions entre l'importance d'un nom pour la définition d'un lexème et sa position dans la définition lexicographique correspondante. Ce facteur de pondération permet de favoriser les noms situés au début de la chaîne.

Soit *ch* une chaîne de caractères d'une définition *d*, un nom *t*, alors le facteur *itpos* (*inverse terme position*) est :

$$itpos = \log \frac{N_{pos}}{N_{pos}(t, ch)} \tag{4.3.}$$

où

N_{pos} : nombre total de positions dans la chaîne de caractères d'une définition,

$N_{pos}(t, ch)$: numéro de la position du nom *t* dans la chaîne de caractères *ch*.

Étant donné que la position des noms est calculée dans les définitions sources du *TLFi* qui ne sont pas lemmatisées, certains noms (ex. *cheval*) pourraient ne pas être présents bien que, dans la définition source, apparaisse une de leurs flexions (ex. *chevaux*). Il est donc nécessaire de calculer la position minimale de toutes les flexions d'un nom, pour ce faire nous utilisons le lexique morphosyntaxique MORPHALOU mis à disposition par le laboratoire (cf. http://www.cnrtl.fr/lexiques/morphalou/).

Nous avons aussi opté pour un traitement spécifique des noms qui suivent les expressions métalinguistiques de la classe d'opposition ou de négation. Ces noms sont affectés automatiquement d'une position négative, ce qui nous permet ensuite de les différencier des autres noms car ils ne représentent pas de caractéristiques spécifiques dans une définition.

4.3. Évaluation de chaque facteur de pondération par rapport aux CC et CP du projet Definiens

Afin d'évaluer les pondérations proposées des noms dans les définitions du *TLFi*, nous sommes partie du constat qu'en général les définitions des entrées nominales du *TLFi* sont hyperonymiques, définies via une approche aristotélicienne, où le genre prochain représente un bon candidat hyperonyme pour un lexème donné du *TLFi*. Compte tenu du fait qu'une telle évaluation manuelle serait très coûteuse en temps (le *TLF* contient plus de 50 000 vedettes nominales), nous avons choisi de valider chaque pondération automatiquement par rapport à une ressource de référence définie par ailleurs et validée manuellement : les définitions structurées du *TLFi* produites dans le cadre du projet Definiens (cf. §1.7.2.), où les composantes centrales (CC) et périphériques (CP) ont été déterminées et validées manuellement.

Pour montrer l'influence de chaque facteur de pondération sur le calcul du poids des noms dans un premier temps, nous calculons les pondérations proposées en essayant de les combiner afin d'obtenir une formule de pondération finale. Dans un deuxième temps, nous évaluons les résultats en utilisant les définitions structurées du projet Definiens.

Étant donné que le projet Definiens n'est pas encore terminé, nous n'avons pu valider nos pondérations que pour 30 % des entrées nominales du *TLFi*, mais cet ensemble nous est apparu suffisamment représentatif (plus de 55 000 définitions) pour en valider les conclusions.

Corpus de travail (*TLFi*)	Total
Entrées nominales	50 471
Définitions	127 089

Tableau 4.4. Information sur le corpus de travail

Corpus de référence (Definiens)	Total
Entrées nominales	14 749 (30 %)
Définitions	55 865 (44 %)

Tableau 4.5. Information sur le corpus de référence

Ainsi, nous faisons l'hypothèse que les noms de poids maximal, représentent les informations les plus pertinentes au sein des définitions et doivent se trouver dans les composantes centrales (CC) des définitions annotées du projet Definiens. Nous nous appuyons aussi sur une seconde hypothèse, à savoir que les candidats hyperonymes des lexèmes doivent faire partie des informations pertinentes des définitions.

Pour valider notre première hypothèse, nous calculons la précision qui représente le rapport entre le nombre de noms de poids maximal annotés comme CC et le nombre total de noms de poids maximal. Cette précision est calculée pour chaque pondération. Nous présentons les diverses analyses effectuées dans la section suivante.

4.4. Analyse de l'influence de chaque facteur de pondération sur le calcul de poids des noms

Les analyses effectuées ont comme but d'une part, de montrer l'importance de chaque facteur de pondération et d'autre part, de déterminer un schéma de pondération finale des noms dans les définitions lexicographiques.

Dans un premier temps, nous combinons la pondération locale (TF) et globale (TD) en proposant une formule de pondération (4.4.) et comparons les résultats avec la pondération TD-IDF.

$$TF - DF = TF(t, d) * DF(t, c) \quad (4.4.)$$

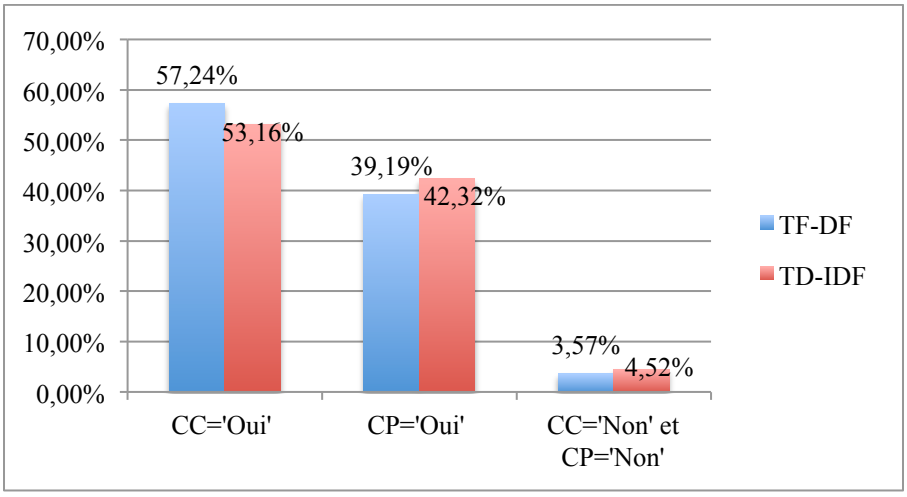


Figure 4.5. Précision pour des noms de poids maximal obtenue avec la formule de pondération TF-DF proposée et TD-IDF

Les résultats présentés dans la figure 4.5. montrent que la précision avec laquelle la pondération TF-DF détermine les noms de poids maximal situés dans les composantes centrales du projet Definiens, est plus élevée de 4,08 % que celles obtenues avec la pondération TD-IDF. Ainsi, nous pouvons affirmer que dans un dictionnaire les noms fréquents dans les définitions pour un domaine donné sont plus discriminants que ceux qui apparaissent moins fréquemment.

Toutefois, nous voyons qu'avec la pondération TF-DF 39,19 % des noms de poids maximal ont été déterminés dans les composantes périphériques. En même temps, 3,57 % des noms de poids maximal n'ont été identifiés ni dans les composantes centrales (CC) ni dans les composantes périphériques (CP). Ceci est dû au fait que, dans le projet Definiens, certaines définitions des lexèmes n'ont pas été annotées. Nous tenons à préciser ici qu'à cause des différences de modèles de données de Definiens et de SEMEME l'évaluation n'a pas été effectuée pour chaque définition distincte, mais pour l'ensemble des définitions des lexèmes d'un vocable du *TLFi*. Étant donné qu'un même nom peut être identifié dans une définition dans les CC et dans une autre dans les CP, lors du calcul de la précision nous l'avons considéré plutôt comme CC.

Pour améliorer les résultats de la précision calculée avec la pondération TF-DF nous ajoutons à celle-ci la pondération par position ITPOS. Nous comparons les résultats avec la formule de pondération TD-IDF à laquelle nous ajoutons aussi la pondération par position :

$$TF - DF - ITPOS = TF(t,d) * DF(t,c) * ITPOS(t,ch) \quad (4.5.)$$

$$TD - IDF - ITPOS = TD(t,d) * IDF(t,D) * ITPOS(t,ch) \quad (4.6.)$$

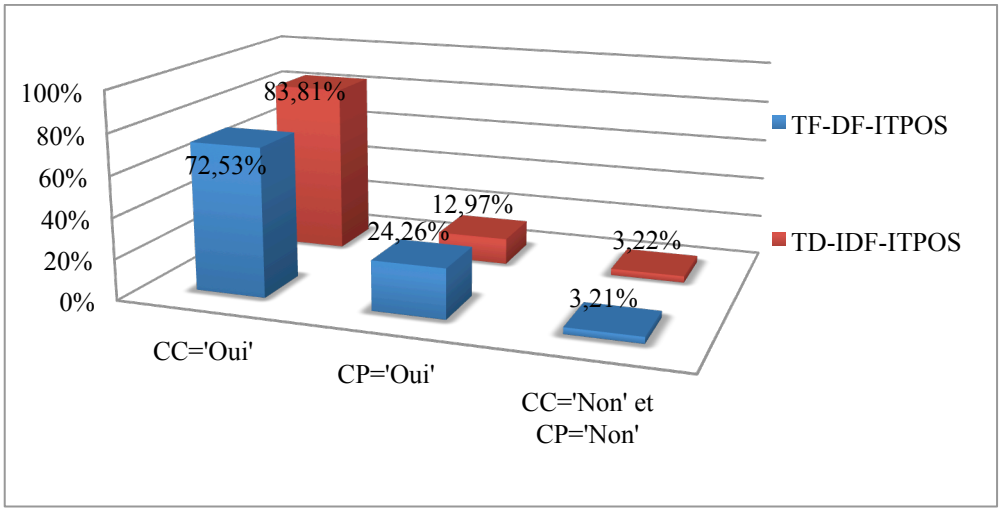


Figure 4.6. Précision pour des noms de poids maximal obtenue avec la formule de pondération TF-DF-ITPOS et TD-IDF-ITPOS

En analysant les résultats présentés dans la figure 4.6. avec ceux de la figure 4.5. nous remarquons que la prise en compte de la pondération ITPOS permet d'améliorer les résultats, la meilleure précision étant toutefois obtenue avec la pondération TD-IDF-ITPOS. Ces résultats s'expliquent par le fait que dans la pondération TF-DF-ITPOS c'est la pondération globale qui fait baisser la précision, car en analysant les résultats nous constatons que, dans un domaine donné, les noms de poids maximal sont parfois les noms les plus spécifiques pour un domaine donné. Ils ne représentent pas toujours de bons candidats hyperonymes des lexèmes du *TLFi*. C'est pour cette raison que ces noms sont identifiés dans les CP (ex. le nom *instrument* pour le lexème PISTON dans le domaine *musique*). Ainsi, nous pouvons affirmer que dans un dictionnaire les noms fréquents dans un domaine donné ne sont pas toujours classifiants. C'est dans le cas de la pondération TD-IDF-ITPOS que la meilleure précision est obtenue, car la pondération TD-IDF donne moins de poids aux termes les plus fréquents et par conséquent c'est la pondération par position qui domine. Reste à déterminer maintenant si ce n'est pas la pondération par position qui joue en fait le rôle plus important.

Pour prouver l'hypothèse que la position du nom dans la définition est le critère sur lequel on peut s'appuyer lors de l'extraction d'informations pertinentes des définitions nous calculons la précision pour la pondération par position seule (ITPOS).

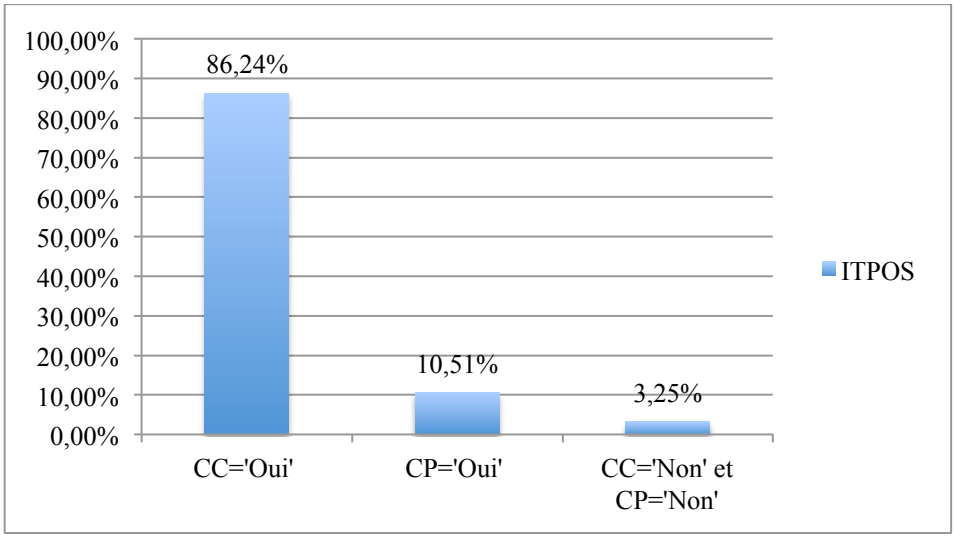


Figure 4.7. Précision pour des noms de poids maximal obtenue avec la formule de pondération ITPOS

Ainsi, nous voyons qu'avec la seule pondération par position la précision augmente avec 2,43 % par rapport à la précision obtenue avec la pondération TD-IDF-ITPOS. Ce fait permet de valider l'hypothèse que, dans des définitions lexicographiques des vocables nominaux, la position des noms joue un rôle plus important que leurs fréquences dans l'ensemble de définitions pour un domaine donné. Dans ces cas seulement 10,51 % des noms de poids maximal ont été déterminés dans une CP. L'une des causes étant la structure de certaines définitions. En effet, les vocables de ces lexèmes peuvent appartenir à deux catégories, nominale et adjectivale, qui ne sont pas distinguées pour certaines vedettes du *TLFi*[66] (ex. CLINIQUE[1], adj. et subst. fém., ACIDE[1], adj. et subst. et ACIDE[2], adj. et subst. masc., ACTUEL, ELLE, adj. et subst., ASIATIQUE, adj. et subst., etc.). Les définitions des adjectifs sont plus complexes que celles des substantifs et ne suivent pas le schéma *hyperonyme* + *spécificités* ou ne l'intègrent que partiellement (en réutilisant les définitions nominales). Les principales structures possibles sont : *Qui +Verbe +X*, *Celui +qui +Verbe, Adj. +Nom*. C'est pourquoi dans ce type de structures, les noms de poids maximal sont identifiés dans les

[66] Pour s'en rendre compte, il suffit d'effectuer une interrogation simple sur ces entrées.

CP. Par exemple, dans la définition structurée de Definiens *</CC> Qui est provoqué <CP> par les boissons alcooliques </CP>* du lexème ALCOOLIQUE*,* le nom *boissons* est annoté comme CP.

Une autre cause d'identification des noms de poids maximal comme composantes périphériques provient du fait que, dans le projet Definiens, certaines définitions des lexèmes n'ont pas été structurées, et que, par ailleurs, quelques erreurs ont été commises lors du balisage des définitions au sein du *TLFi* (ex. *balises non fermées,* etc.).

Ainsi, les évaluations effectuées renforcent l'hypothèse que la position des noms est le seul critère sur lequel on peut s'appuyer lors de l'extraction d'informations pertinentes des définitions. Dans la suite de nos recherches, nous retenons seulement la pondération par position (ITPOS) comme notre schéma de pondération finale.

4.5. Conclusion

Pour pouvoir utiliser le *TLFi* lors de l'enrichissement du thésaurus Xilopix ainsi que pour améliorer l'indexation et la recherche d'images, nous avons eu besoin d'une caractéristique quantitative (mesurable) et comparable pour les noms de ses définitions. Cette caractéristique doit permettre de comparer les noms entre eux et de reconnaître les noms susceptibles d'être des hyperonymes possibles d'un lexème.

Ainsi, dans un premier temps nous avons fait une analyse de notre corpus de travail afin de déterminer quelles informations du *TLFi* seront utilisées lors de nos recherches. Ensuite, nous avons analysé des différents critères de pondération des noms dans les définitions du *TLFi*. Ce fait nous a permis de définir dans un deuxième temps, trois facteurs de pondération des noms pour mesurer l'importance des noms dans les définitions lexicographiques. Pour pouvoir analyser l'influence de chaque facteur de pondération proposé sur le calcul de poids des noms et ainsi déterminer un schéma de pondération finale, nous avons réalisé une évaluation des informations pertinentes extraites des définitions du *TLFi* avec chaque pondération par rapport aux composantes centrales (CC) et périphériques (CP) définies dans le projet Definiens. L'analyse des résultats de l'évaluation a montré que seule la pondération par position détermine avec une précision très élevée les informations pertinentes des définitions. Toutefois, nous n'avons pas pu valider à ce stade si les candidats hyperonymes des lexèmes font partie de ces informations extraites.

Ainsi, dans le chapitre suivant nous allons procéder à la construction automatique de hiérarchies sémantiques du *TLFi* et réaliser une évaluation plus fine des informations extraites à l'aide de schéma de pondération proposé au travers d'un processus d'évaluation manuellement des relations d'hyperonymie elles-mêmes. L'évaluation étant confiée à des documentalistes de Xilopix.

CHAPITRE 5

Construction automatique de hiérarchies sémantiques à partir du *TLFi*

Sommaire

5.1. Hiérarchisation des noms 146
5.1.1. Règles d'inclusion 146
5.1.2. Règles d'association 148
5.1.3. Règles de hiérarchisation 148
5.2. Méthodologie de construction automatique de hiérarchies sémantiques 149
5.2.1. Vue globale de l'approche 149
5.2.2. Description détaillée 149
5.3. Évaluation manuelle des relations hyperonymiques 159
5.3.1. Présentation du corpus d'évaluation 160
5.3.2. Analyse des résultats d'évaluation 161
5.3.2.1. Analyse des résultats positifs de l'évaluation manuelle 162
5.3.2.2. Analyse des résultats négatifs de l'évaluation manuelle 163
5.4. Comparaison des hiérarchies sémantiques avec le thésaurus Xilopix 166
5.4.1. Description du corpus d'évaluation 166
5.4.2. Analyse des résultats d'évaluation 167
5.5. Proposition de méthodologie d'enrichissement du thésaurus Xilopix 171
5.6. Conclusion 172

Chapitre 5. Construction automatique de hiérarchies sémantiques à partir du *TLFi*

Dans le chapitre 3 de cette thèse, nous avons présenté les principales ressources linguistiques utilisées pour la recherche d'informations et la recherche d'images. Nous avons vu que les thésaurus sont utilisés avec succès dans ce domaine à la fois comme outils d'indexation et de recherche. Toutefois, leur construction manuelle est assez coûteuse en temps et en argent.

Afin de pouvoir expliciter et exploiter les connaissances d'un dictionnaire de langue, le *TLFi*, pour enrichir le thésaurus construit manuellement au sein de la société Xilopix et améliorer l'indexation et la recherche d'images, nous avons proposé dans le chapitre 4 une heuristique de pondération des noms des définitions du *TLFi*. Cela nous permet d'attribuer un poids à chaque nom de la définition, en faisant l'hypothèse que le nom de poids maximal représente un candidat hyperonyme du lexème donné. Cependant, les noms pondérés ne sont pas suffisants pour être utilisés lors de l'enrichissement du thésaurus Xilopix, il conviendrait de pouvoir les insérer dans une hiérarchie sémantique pour faciliter un tel enrichissement.

Dans ce chapitre, nous nous intéressons donc à la façon d'organiser les noms des définitions de tous les lexèmes d'un même vocable, afin de construire des hiérarchies sémantiques. Ainsi, les hiérarchies sémantiques construites automatiquement à partir du *TLFi* seront utilisées, d'un côté, pour enrichir le thésaurus existant et, d'un autre côté, pour l'indexation et la recherche d'images. Nous commençons par présenter les principales règles de hiérarchisation des noms que nous avons retenues. Nous présentons ensuite l'algorithme de construction automatique de hiérarchies sémantiques à partir des définitions des lexèmes du *TLFi* d'un vocable en donnant d'abord une vue globale de notre approche, puis en détaillant les étapes principales de l'algorithme. Par la suite, nous analysons les résultats de l'évaluation manuelle de relations hyperonymiques ainsi construites et ceux de la comparaison des hiérarchies sémantiques construites automatiquement avec le thésaurus Xilopix. Nous concluons ce chapitre en proposant une méthodologie d'enrichissement du thésaurus existant en utilisant les hiérarchies sémantiques générées à partir du *TLFi*.

5.1. Hiérarchisation des noms

Dans cette section, nous présentons les trois types de règles qui sont utilisées pour la hiérarchisation des noms pondérés de la définition d'un lexème donné.

5.1.1. Règles d'inclusion

Les règles d'inclusion indiquent qu'un ensemble est un sous-ensemble d'un second. Dans le *TLFi* la définition de chaque lexème d'un vocable est associée à un domaine. Ainsi, un lexème X' du vocable X peut être inclus dans un sous-ensemble du domaine D :

$$\boldsymbol{R} : X' \subset D \tag{5.1.}$$

$$Ananas \subset Botanique$$

$$Ananas \subset Générique$$

Nous pouvons aussi établir des relations d'inclusion entre le lexème X' et les noms de sa définition $X' = \{X_1,\ X_2\ ,\ \ldots,\ X_n\}$. Ainsi, nous considérerons qu'un nom X_n dont le poids est maximal dans la définition du lexème X' est un bon candidat hyperonyme pour le lexème et en conséquence qu'il peut inclure celui-ci :

$$\boldsymbol{R}\ :\ \boldsymbol{Si}\ P(X_n, X') = max\ \boldsymbol{alors}\ X' \subset X_n \tag{5.2.}$$

où

P est le poids du nom X_n dans la définition du lexème X'.

Dans le cas de l'étude du lexème ANANAS de définition « Plante monocotylédone de la famille des Broméliacées, croissant dans les contrées chaudes de l'Asie, de l'Afrique, de l'Amérique, à feuilles radicales et pointues, bordées d'épines, ressemblant à celles de l'aloès », nous obtenons :

$$P(plante, ANANAS) = max\ \Rightarrow\ Ananas \subset Plante$$

À partir des règles (5.1.) et (5.2.), nous déduisons une nouvelle règle :

$$\boldsymbol{R} : \boldsymbol{Si}\ X' \subset D\ et\ X' \subset X_n\ \boldsymbol{alors}\ X_n \subset D \tag{5.3.}$$

$$Ananas \subset Botanique\ et\ Ananas \subset Plante\ \Rightarrow\ Plante \subset Botanique$$

La règle générale peut être écrite sous la forme suivante :

$$\boldsymbol{R : Si}\ X' \subset X_n\ et\ X_n \subset D\ \boldsymbol{alors}\ X' \subset X_n \subset D \tag{5.4.}$$

$$Ananas \subset Plante \subset Botanique$$

Afin de permettre la croissance de l'arbre en profondeur, nous recherchons dans le *TLFi* pour chaque nom de la définition du lexème X' le vocable correspondant dont les lexèmes sont inclus dans le même domaine que celui du X' et qui contiennent le nom X_n de poids maximal. Soit Z un tel vocable et Z' son lexème qui contient dans sa définition le nom X_n, nous appliquons alors une nouvelle règle :

$$\begin{aligned} \boldsymbol{R : Si}\ & P(X_n, Z') = max\ et\ X', Z' \subset D \\ & avec\ D(X') = D(Z')\ ou\ D(Z') = générique \\ & \boldsymbol{alors}\ Z' \subset X_n \end{aligned} \tag{5.5.}$$

où

$Z' = \{Z_1, Z_2, \dots, Z_n\}$ est un lexème d'un nouveau vocable Z,

$P(X_n, Z')$ est le poids du nom X_n dans la définition du lexème Z',

$D(X')$, $D(Z')$ sont les domaines des lexèmes X' et Z'.

Soit lexème X' = ANANAS dont D(ANANAS) = botanique et le vocable Z = BROMÉLIACÉES avec son lexème Z' de définition « Famille de plantes tropicales de la classe des monocotylédones comprenant notamment l'ananas » où D(BROMÉLIACÉES) = botanique. En appliquant la règle 5.5. on obtient que si P(plante, BROMÉLIACÉES) = max et D(ANANAS) = D(BROMÉLIACÉES) alors Broméliacées ⊂ Plante.

La structure choisie dans le *TLFi* pour les définitions des domaines liés aux *sciences naturelles* typiquement *botanique* et *zoologie* fait apparaître systématiquement le nom de la famille ainsi que des exemples prototypiques de cette famille. Ainsi, une simple vérification de l'apparition du mot-forme du lexème X' dans la définition du lexème Z' nous permet d'ajouter une nouvelle règle :

$$\begin{aligned} & \boldsymbol{R : Si}\ X' \subset X_n,\ Z' \subset X_n\ et\ X' \in Z' \\ & et\ X', Z' \subset D(Botanique, Zoologie)\ \boldsymbol{alors}\ X' \subset Z' \end{aligned} \tag{5.6.}$$

$$Ananas \subset Plante,\ Broméliacées \subset Plante\ et\ Ananas \in Broméliacées$$
$$et\ Plante,\ Broméliacées \in\ D(Botanique) => Ananas \subset Broméliacées$$

Nous proposons enfin une règle de généralisation du type :

$$\boldsymbol{R : Si}\ \ Z' \subset X_n\ \ et\ X' \subset\ Z\ \boldsymbol{alors}\ X' \subset\ Z' \subset X_n \tag{5.7.}$$

Soit X′ = ANANAS, Z′ = BROMÉLIACÉES et X_n = plante. Si Broméliacées ⊂ Plante et Ananas ⊂ Broméliacées alors Ananas ⊂ Broméliacées ⊂ Plante. Par la transitivité, en appliquant la règle (5.3.) où D = botanique, nous obtenons : Ananas ⊂ Broméliacées ⊂ Plante ⊂ Botanique.

5.1.2. Règles d'association

En complément des règles d'inclusion présentées ci-dessus, nous proposons des règles d'association qui permettent d'établir des liens d'association entre un lexème et les noms de sa définition de sorte que l'évocation d'un nom fasse surgir le lexème (ex. *Ananas* → *Plante, Afrique*). Une règle d'association est une relation d'implication de la forme :

$$\boldsymbol{R :}\ X\ \rightarrow \{X_1, X_2, \dots, X_n\} \tag{5.8.}$$

où

X' est un lexème du vocable X et $\{X_1, X_2, \dots, X_n\}$ est un ensemble de noms.

Ainsi, selon cette règle, nous pouvons établir que le lexème ANANAS implique tous les noms de sa définition :

$$ANANAS \rightarrow \{Asie, Afrique, feuille, épine, aloès, \text{etc.}\}$$

5.1.3. Règles de hiérarchisation

Les règles de hiérarchisation doivent permettre la construction de hiérarchies sémantiques à arborescence simple. La hiérarchisation des noms d'un lexème se construit à partir des règles d'inclusion spécifiées ci-dessus. Les noms sont hiérarchisés en vertu du principe selon lequel le nom qui inclut un autre nom figure en haut de la hiérarchie et représente son nœud père. Le nom inclus dans un autre nom représente

pour sa part un nœud fils. Les règles d'association sont appliquées à la fin de hiérarchisation.

5.2. Méthodologie de construction automatique de hiérarchies sémantiques

Les règles présentées dans la section précédente se trouvent à la base de notre approche de construction automatique de hiérarchies sémantiques[67] que nous allons décrire par la suite.

5.2.1. Vue globale de l'approche

Le principe de l'algorithme est de construire pour chaque vocable du *TLFi* des hiérarchies sémantiques en exploitant les définitions de ses lexèmes. La figure 5.1. présente une vue globale de notre approche de construction automatique de hiérarchies sémantiques. Ainsi, pour chaque vocable du *TLFi* des hiérarchies sémantiques différentes sont construites.

5.2.2. Description détaillée

L'algorithme de construction automatique de hiérarchies sémantiques pour un vocable du *TLFi* est fondé sur deux processus. Le premier processus a pour objectif de transformer chaque lexème du vocable, les lemmes et les domaines de ses définitions en nœuds de hiérarchies sémantiques. Le second processus permet de hiérarchiser les nœuds obtenus lors du processus précédent en créant des arbres hiérarchiques.

Nous allons décrire ci-dessous les deux processus qui sont à la base de l'algorithme de construction automatique de hiérarchies sémantiques correspondantes à un vocable X du *TLFi*.

[67] Des hiérarchies qui peuvent se représenter sous la forme des arbres hiérarchiques à arborescence simple (cf. §3.2.3.).

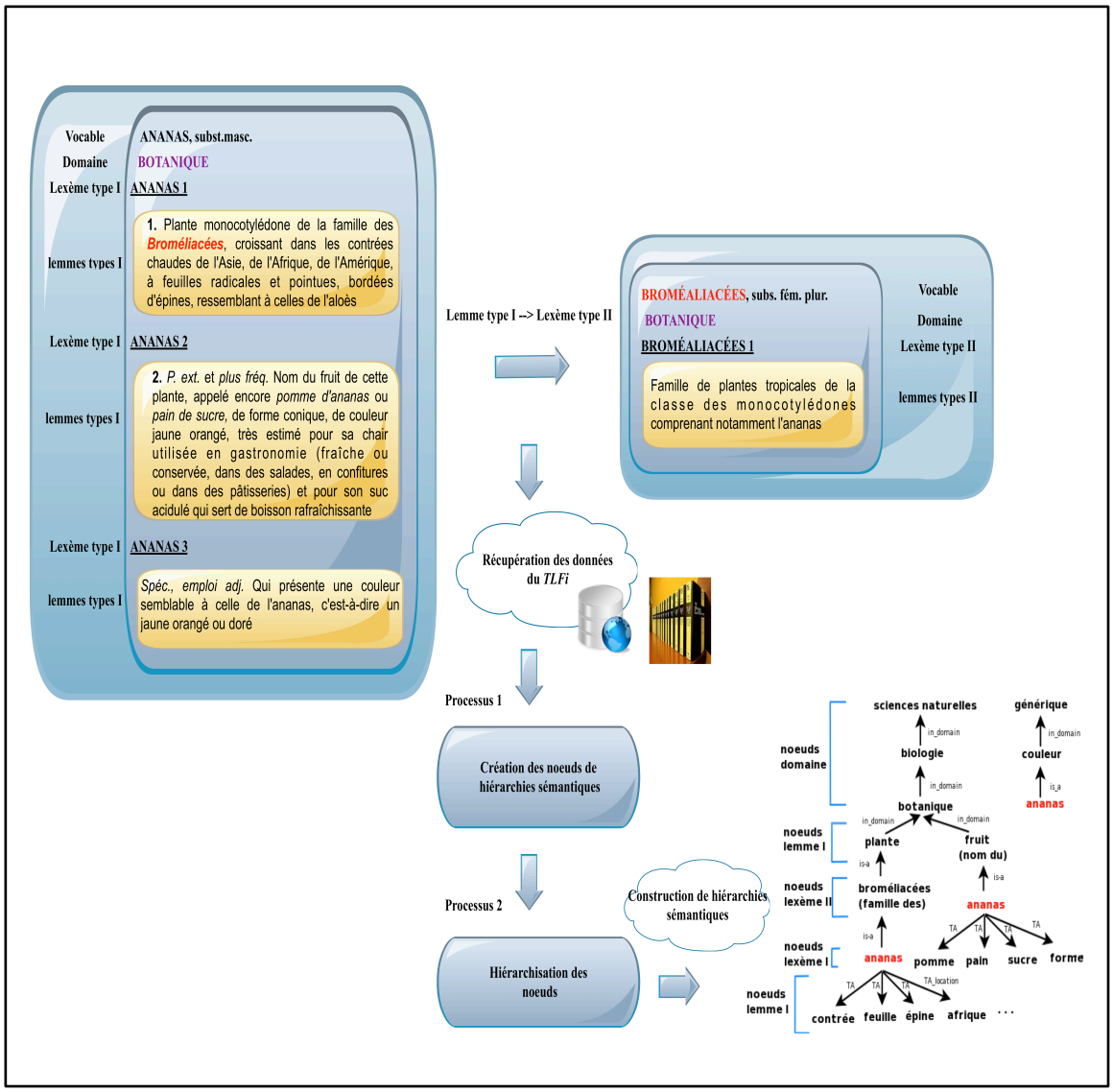


Figure 5.1. Vue globale de l'algorithme de construction de hiérarchies sémantiques

Processus 1 : création des nœuds de hiérarchies sémantiques

L'objectif de ce processus est de transformer toute l'information contenue dans la microstructure d'un vocable X du *TLFi* en nœuds de hiérarchies sémantiques. Ce processus se fait en deux étapes :

1.1. Extraction des données issues du *TLFi* pour le vocable X.

Pour chaque lexème du vocable X, nous extrayons, à partir du *TLFi*, après le filtrage des mots vides et ceux liés au métalangage lexicographique, les noms des définitions avec leurs pondérations et leur domaine de définition.

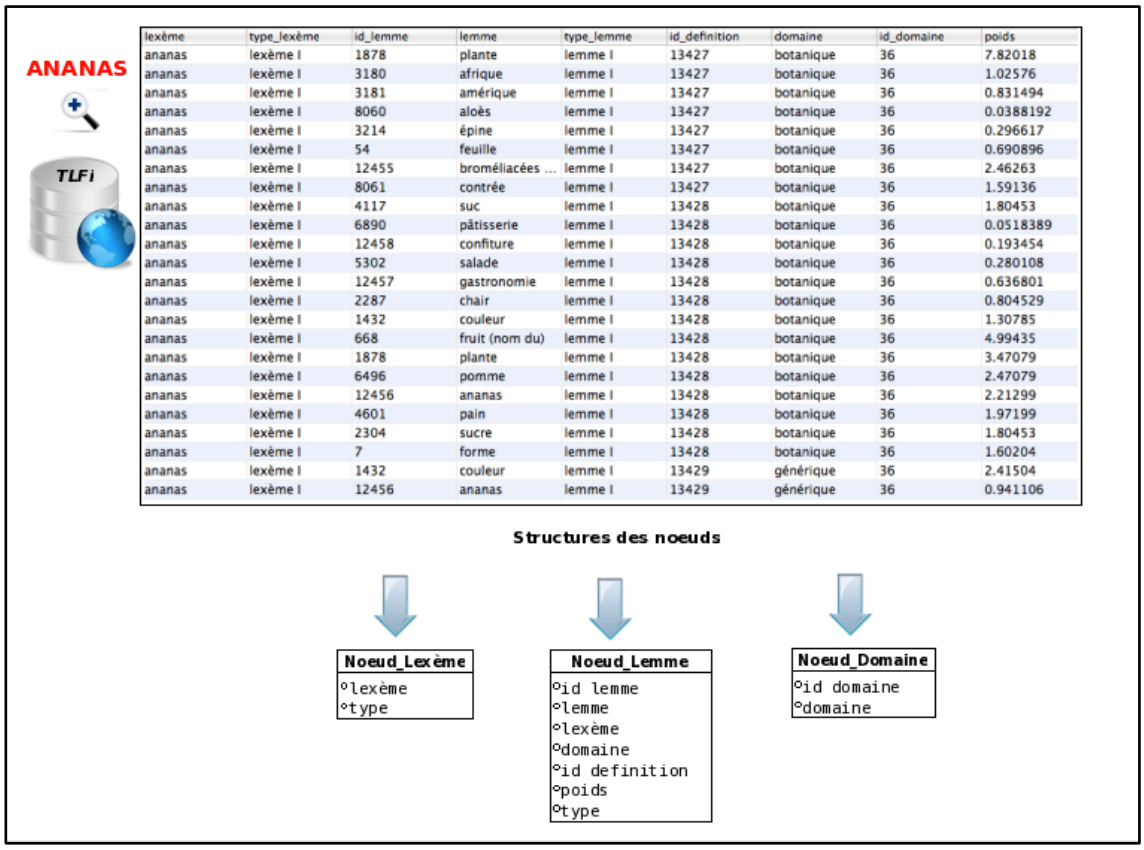
1.2. Transformation de la liste de données d'un vocable X en nœuds de hiérarchies sémantiques.

À partir des données extraites pour un vocable X, nous procédons à la création des nœuds de hiérarchies sémantiques lui correspondant. Un nœud représente une structure de données. Nous considérons comme nœuds de telles hiérarchies sémantiques chaque lexème du vocable *X* du *TLFi* ainsi que les lemmes des noms de la définition et le domaine auquel se rattache cette définition. Par conséquent, nous distinguons trois types de nœuds : nœud lexème, nœud lemme et nœud domaine (cf. figure 5.1.). Suivant le type de nœud, la structure de données est différente :

- Pour le nœud domaine, la structure de données est limitée au nom de domaine et son identifiant.
- La structure du nœud lexème est formée uniquement du nom du lexème et de son type.
- Quant à la structure du nœud lemme, elle contient son identifiant, le nom du lexème, le nom du lemme, son poids, l'identifiant de la définition, son domaine et son type.

Ainsi, chaque acception d'un vocable X du *TLFi* représente un nœud lexème et l'ensemble des lemmes des noms de sa définition correspond à des nœuds lemmes. Nous considérons que les nœuds lexèmes et les nœuds lemmes d'un vocable X pour lequel les arbres hiérarchiques sont construits sont de type I. Par exemple, dans la figure 5.1. le vocable ANANAS a trois lexèmes ANANAS 1, ANANAS 2, ANANAS 3 qui représentent des nœuds de type I ; les lemmes de leurs définitions comme *plante, famille, broméliacées, Asie, Afrique, feuille, épine, aloès,* etc. sont considérés comme des nœuds lemmes de type I. Compte tenu du fait que les lemmes de type I sont à leur tour définis dans le *TLFi*, nous les considérons eux-mêmes ensuite comme des nœuds lexèmes de type II ; les lemmes de leurs définitions sont quant à eux considérés comme des nœuds lemmes de type II. Par exemple, le lemme de type I *broméliacées* est considéré à son tour comme lexème de type II (cf. figure 5.1.). Bien que ce processus puisse être poursuivi aux niveaux supérieurs (III, IV, etc.) nous limitons le processus aux seuls deux premiers niveaux considérant que cela devrait être suffisant pour extraire l'information principale à partir de la définition d'un lexème. La figure 5.2. présente un

exemple du processus 1 pour le vocable ANANAS, qui consiste en la récupération de l'information sur chacun de ses lexèmes à partir de la base de données du *TLFi* et la création de trois types de nœuds de structures différentes. Ainsi, nous obtenons 3[68] nœuds lexèmes de type I, 27 nœuds lemmes de type I et 2 nœuds domaine (*botanique* et *générique*).



lexème	type_lexème	id_lemme	lemme	type_lemme	id_definition	domaine	id_domaine	poids
ananas	lexème I	1878	plante	lemme I	13427	botanique	36	7.82018
ananas	lexème I	3180	afrique	lemme I	13427	botanique	36	1.02576
ananas	lexème I	3181	amérique	lemme I	13427	botanique	36	0.831494
ananas	lexème I	8060	aloès	lemme I	13427	botanique	36	0.0388192
ananas	lexème I	3214	épine	lemme I	13427	botanique	36	0.296617
ananas	lexème I	54	feuille	lemme I	13427	botanique	36	0.690896
ananas	lexème I	12455	broméliacées ...	lemme I	13427	botanique	36	2.46263
ananas	lexème I	8061	contrée	lemme I	13427	botanique	36	1.59136
ananas	lexème I	4117	suc	lemme I	13428	botanique	36	1.80453
ananas	lexème I	6890	pâtisserie	lemme I	13428	botanique	36	0.0518389
ananas	lexème I	12458	confiture	lemme I	13428	botanique	36	0.193454
ananas	lexème I	5302	salade	lemme I	13428	botanique	36	0.280108
ananas	lexème I	12457	gastronomie	lemme I	13428	botanique	36	0.636801
ananas	lexème I	2287	chair	lemme I	13428	botanique	36	0.804529
ananas	lexème I	1432	couleur	lemme I	13428	botanique	36	1.30785
ananas	lexème I	668	fruit (nom du)	lemme I	13428	botanique	36	4.99435
ananas	lexème I	1878	plante	lemme I	13428	botanique	36	3.47079
ananas	lexème I	6496	pomme	lemme I	13428	botanique	36	2.47079
ananas	lexème I	12456	ananas	lemme I	13428	botanique	36	2.21299
ananas	lexème I	4601	pain	lemme I	13428	botanique	36	1.97199
ananas	lexème I	2304	sucre	lemme I	13428	botanique	36	1.80453
ananas	lexème I	7	forme	lemme I	13428	botanique	36	1.60204
ananas	lexème I	1432	couleur	lemme I	13429	générique	36	2.41504
ananas	lexème I	12456	ananas	lemme I	13429	générique	36	0.941106

Figure 5.2. Exemple du processus 1 de l'algorithme

Processus 2 : hiérarchisation des nœuds

L'objectif de ce processus est de hiérarchiser les nœuds obtenus pour un même vocable en construisant des arbres hiérarchiques. La hiérarchisation se réalise en comparant les structures de données de chaque nœud avec les autres.

Ce processus est réalisé en cinq étapes :

[68] En réalité il existe que 3 nœuds lexèmes, mais dans la base de données nous en avons 27 car à chaque lexème est associé un lemme correspondant.

2.1. Détermination des nœuds pères pour les nœuds lexèmes.

À partir des nœuds créés, nous déterminons les nœuds pères de chaque nœud lexème de type I. Pour ce faire, nous regroupons les nœuds lemmes de type I par leur identifiant de définition et pour chaque groupe ainsi créé nous déterminons le nœud lemme dont la pondération est maximale. Ce nœud devient le nœud père pour un nœud lexème de type I. Ainsi, plusieurs nœuds pères différents (autant que le nombre des lexèmes d'un vocable) sont créés ayant comme nœud fils le nœud lexème de type I correspondant. Puis le pas suivant est effectué :

SI pour deux groupes différents émergent deux nœuds lemmes de type I de poids maximal ayant même domaine et même nom

ALORS seul un nœud devient le nœud père du nœud lexème de type I

FIN SI

Dans les arbres hiérarchiques construits, nous considérons que les nœuds pères déterminés représentent des candidats hyperonymes pour le lexème donné et qu'ils sont liés aux nœuds fils par la relation de type *is-a*. Dans le cas où les nœuds pères ont un poids négatif, la relation est de type *not-is-a* (cf. §4.2.3.3.). Par exemple, dans la définition « Manque de sommeil, insomnie » du lexème AGRYPNIE, le nom *sommeil* a un poids négatif car il est situé après une expression métalinguistique de la classe de négation (ex. manque).

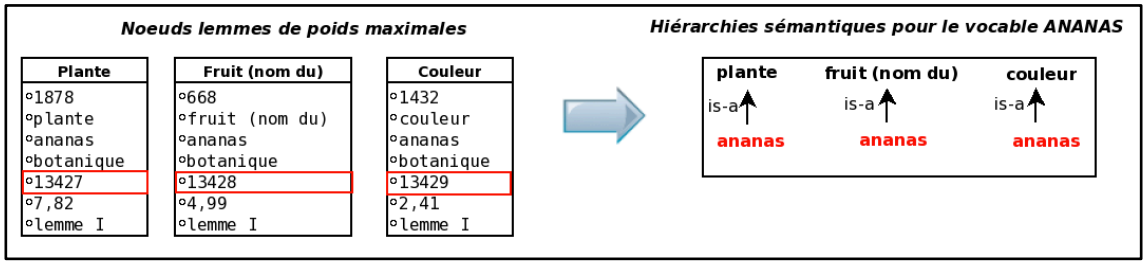


Figure 5.3. Exemple de détermination des nœuds pères pour les nœuds lexèmes du vocable ANANAS

La figure 5.3. présente un exemple de détermination des nœuds pères pour les nœuds lexèmes du vocable ANANAS. Ainsi, en regroupant les nœuds lemmes par leur identifiant de la définition, l'algorithme détermine pour chaque groupe (chaque lexème

du même vocable) le nœud lemme de poids maximal (*plante, fruit (nom du), couleur*) qui est considéré comme le nœud père correspondant à chaque nœud lexème du vocable ANANAS.

2.2. Détermination des nœuds fils pour des nœuds pères créés.

Afin de déterminer des nœuds fils pour des nœuds pères créés, nous procédons à la création des autres nœuds. En effet, de nouveaux nœuds sont créés à partir des nœuds lemmes de type I existants. Cette procédure consiste à appliquer pour chaque nœud lemme I le processus 1, en d'autres termes chaque nœud lemme de type I est considéré comme un vocable du *TLFi* pour lequel on récupère la liste de données à partir du *TLFi*, qui sont ensuite transformées en nœuds correspondants. Les nouveaux nœuds obtenus (nommés respectivement nœuds lexèmes de type II et nœuds lemmes de type II) servent à la détermination des autres nœuds fils pour des nœuds pères créés. Pour cela, nous comparons les structures de nouveaux nœuds obtenus avec celles des nœuds pères créés en déterminant parmi les nœuds lemmes de type II ceux qui ont le même nom du lemme que les nœuds pères créés.

SI les nœuds lemmes de type II ont le même nom du lemme que les nœuds pères créés

ALORS

SI leur poids est maximal et s'ils ont le même domaine que les nœuds pères ou le domaine *générique*

ALORS ces nœuds lemmes de type II sont remplacés par les nœuds lexèmes de type II correspondants aux noms des lexèmes des nœuds lemmes de type II et ils deviennent des nœuds fils du nœud père créé

FIN SI

FIN SI

Dans les arbres hiérarchiques ainsi construits, les nouveaux nœuds fils déterminés sont liés avec les nœuds pères par la relation de type *is-a* ou *not-is-a*.

La figure 5.4. présente un exemple de détermination des nœuds fils pour des nœuds pères des nœuds lexèmes du vocable ANANAS. Ce procédé consiste dans un premier

temps, à créer de nouveaux nœuds à partir des nœuds lemmes de type I déjà existants (*plante, contrée, broméliacées, épine,* etc.) en utilisant le processus 1. Par exemple, à partir du vocable BROMÉLIACÉES, de nouveaux nœuds ont été créés : 3 nœuds lemmes de type II *(plante, classe, ananas*), 1 nœud lexème de type II (*broméliacées (famille des)*). Le nœud domaine *botanique* n'a pas été créé de nouveau parce qu'il existe déjà. Dans un deuxième temps, les nouveaux nœuds lemmes de type II sont comparés avec les nœuds pères (*plante, fruit (nom du), couleur*) des nœuds lexèmes de type I en déterminant seulement un nouveau nœud fils pour le nœud père *plante*. Ensuite, le nœud fils *plante (famille de)* déterminé, qui est un nœud lemme de type II, est remplacé par le nœud lexème de type II *broméliacées (famille des)* correspondant. Finalement, dans la hiérarchie sémantique le nœud lexème de type II *broméliacées (famille des)* devient le nœud fils du nœud père *plante*.

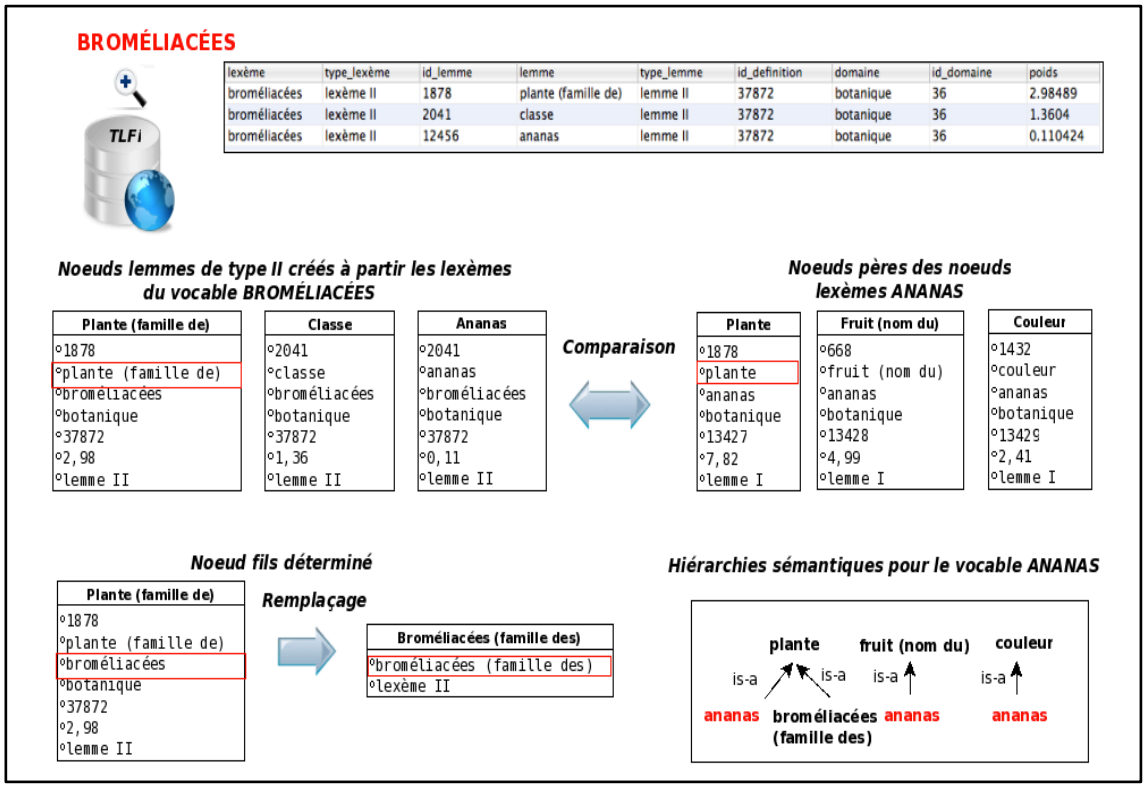


Figure 5.4. Exemple de détermination des nœuds fils pour des nœuds pères des nœuds lexèmes du vocable ANANAS

Cette comparaison des nœuds entre eux est nécessaire pour déterminer si éventuellement un nom de la définition d'un lexème donné a le même candidat

hyperonyme que celui-ci. Dans l'exemple présenté (cf. figure 5.4.), nous déterminons que *plante* est aussi le candidat hyperonyme du lexème BROMÉLIACÉES.

2.3. Transformation des nœuds fils en nœuds pères.

Afin de permettre la croissance de hiérarchies sémantiques en profondeur, nous déterminons pour les nœuds lexèmes de type I du vocable X de nouveaux nœuds pères. Ainsi, nous comparons les nœuds fils obtenus avec les nœuds lemmes de type II en vérifiant :

SI il existe des nœuds lemmes de type II dont le nom du lemme et du lexème sont respectivement identiques à des noms des nœuds lexèmes de type I et à ceux de type II, et si ces nœuds sont dans le domaine *botanique* ou *zoologie*

ALORS les nœuds lexèmes de type II deviennent les nœuds pères des nœuds lexèmes de type I

FIN SI

Dans les hiérarchies construites, les nouveaux nœuds pères sont liés avec les nœuds lexèmes de type I par la relation de type *is-a*. Nous considérons que ces nœuds représentent des candidats *hyperonymes indirects* du lexème parce qu'ils ont été déterminés en consultant des définitions des lexèmes des autres vocables. Le candidat *hyperonyme direct* du lexème donné représente le nœud père de son candidat hyperonyme indirect. Les hiérarchies sémantiques de ce type qui ont deux niveaux de relations hyperonymiques ne sont construites que pour les lexèmes des domaines *botanique* et *zoologie*. Ces deux niveaux de relations ne sont construits que pour ces domaines car la structure de leurs définitions fait toujours apparaître le nom de la famille à laquelle appartient le lexème défini. Nous reviendrons sur ce point dans la section §5.4.2.

La figure 5.5. présente un exemple de transformation des nœuds fils déterminés lors de l'étape précédente en nœuds pères des nœuds lexèmes I. Cette étape de l'algorithme a pour but de comparer les nœuds fils de l'arborescence, le nœud lexème II *broméliacées (famille des)* avec le nœud lexème de type I *ananas* afin de déterminer si le nœud lexème II *broméliacées (famille des)* inclut le nœud lexème de type I *ananas*, c'est-à-dire si le lexème BROMÉLIACÉES contient dans sa définition du *TLFi* pour le domaine *botanique* le nom *ananas*. Pour cela, nous analysons les nœuds lemmes de

type II en déterminant ceux qui ont comme lemme *ananas* et comme lexème *broméliacées* et appartiennent au domaine *botanique*. Cela permet de sélectionner le nœud lemme de type II *ananas* car le nom *ananas* se rencontre dans la définition du lexème BROMÉLIACÉES pour le domaine *botanique*. Nous en concluons donc que le sens du lexème ANANAS est inclus dans le sens du lexème BROMÉLIACÉES, en d'autres termes que le nœud lexème de type II *broméliacées (famille des)* est le nœud père pour le nœud lexème de type I *ananas*. Dans la hiérarchie construite, le nom *broméliacées (famille des)* représente le candidat hyperonyme indirect du lexème ANANAS.

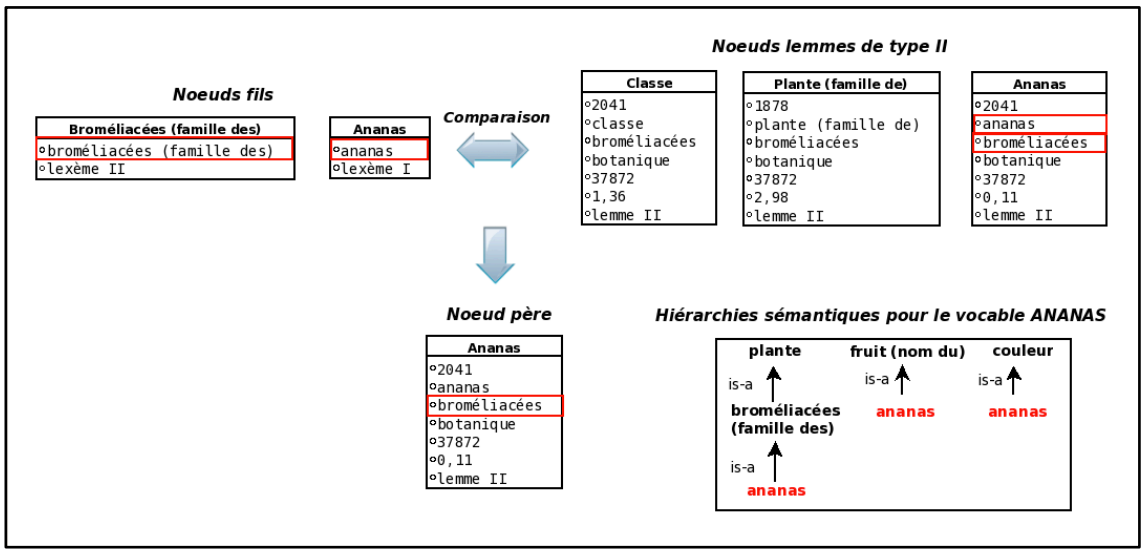


Figure 5.5. Exemple de transformation des nœuds fils en nœuds pères des nœuds lexèmes du vocable ANANAS

2.4. Détermination de la hiérarchie pour les nœuds domaines.

Pour chaque nœud domaine qui représente le domaine associé à une définition, nous déterminons son nœud père en interrogeant le thésaurus des domaines du *TLFi* construit auparavant. Pour agrandir nos hiérarchies sémantiques en profondeur, nous comparons les nœuds domaines avec les nœuds situés en haut de la hiérarchie déjà créée. Si ces nœuds ont le même domaine, alors les nœuds domaines deviennent des nœuds pères pour ceux-là.

Les nœuds domaines sont liés entre eux par la relation de type *in_domain*. Ce type de relation est aussi utilisé pour relier le thésaurus des domaines avec les hiérarchies

sémantiques construites pour les lexèmes d'un vocable donné.

La figure 5.6. présente un exemple de thésaurus des domaines construit à partir du domaine *botanique*. Tenant compte du fait que les nœuds situés en haut de la hiérarchie déjà créée *(plante, fruit (nom du))* ont le même domaine *botanique* et que le nœud *couleur* relève du domaine *générique,* la hiérarchie des domaines issue du *TLFi* est ajoutée à la hiérarchie préalablement construite pour le vocable ANANAS.

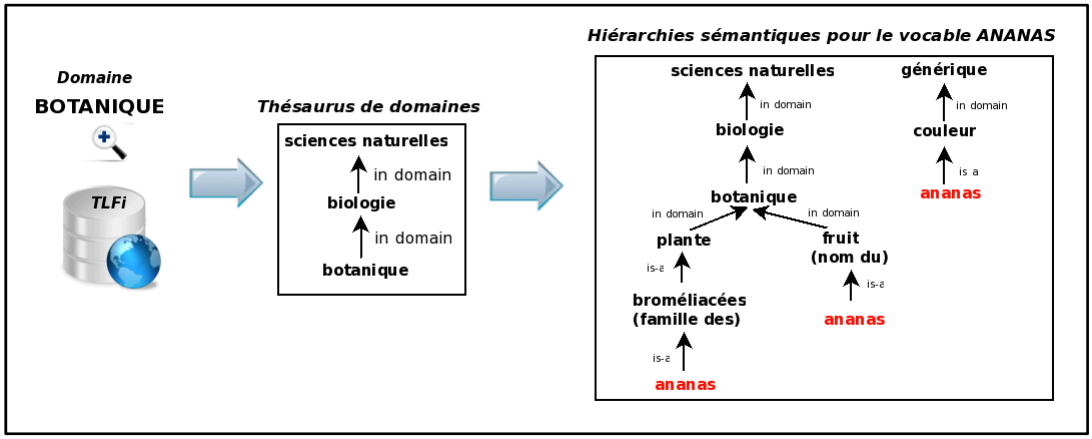


Figure 5.6. Exemple de fusion du thésaurus des domaines aux hiérarchies sémantiques construites automatiquement

2.5. Attribution des nœuds associés aux nœuds lexèmes.

Afin d'élargir les hiérarchies sémantiques construites en profondeur, nous introduisons des nœuds associés. Ce sont des nœuds qui, lors de la recherche, seront utilisés pour diriger l'utilisateur vers des nœuds situés plus bas dans la hiérarchie. Les nœuds associés sont des nœuds lemmes de type I qui n'ont encore pas été utilisés pour la création de hiérarchies sémantiques. L'attribution de ces nœuds associés correspond au processus suivant :

SI les domaines et les identifiants des définitions des nœuds lemmes de type I non utilisés pour la création de hiérarchies sémantiques sont identiques à ceux des nœuds pères des nœuds lexèmes de type I

ALORS ces nœuds sont associés aux nœuds lexèmes de type I correspondants

FIN SI

Il existe trois types de relations entre les nœuds associés et les nœuds lexèmes :

- relation *TA* (terme associé) ;

- relation *TA_location*, dans le cas quand le nœud associé indique un lieu géographique[69] ;
- relation *is_not_TA*, dans le cas quand le nœud associé a un poids négatif (cf. §4.2.3.3.). Cette relation indique que le nœud n'est pas caractéristique pour un lexème donné.

Nous tenons à préciser que dans les hiérarchies sémantiques construites à partir du *TLFi* nous considérons la relation d'association toutes relations autres que la relation *is-a* (*not-is-a*). Alors que dans l'état de l'art concernant le thésaurus (cf. §3.2.2.), les relations d'association ne sont pas des relations hiérarchiques, ils sont plutôt de renvois vers d'autres termes.

La figure 5.7. présente un exemple des hiérarchies sémantiques construites pour le vocable ANANAS, dans lequel, aux hiérarchies, ont été ajoutés des nœuds associés correspondant à chaque nœud lexème de type I.

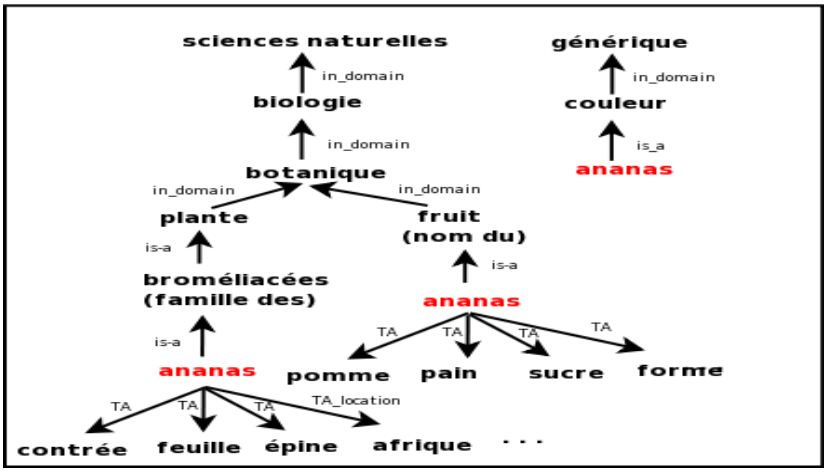


Figure 5.7. Exemple des hiérarchies sémantiques construites pour le vocable ANANAS

5.3. Évaluation manuelle des relations hyperonymiques

L'évaluation des relations hyperonymiques des hiérarchies sémantiques construites a été réalisée manuellement par des documentalistes de Xilopix. Plus précisément, il s'agissait d'évaluer, dans les hiérarchies sémantiques construites automatiquement pour des vocables donnés, la relation d'hyperonymie entre les nœuds lexèmes et ses nœuds

[69] À partir des définitions du *TLFi*, nous avons créé une liste composée de noms propres (109 de lieux), ce qui nous permet d'obtenir des relations de lieu entre un nœud lexème et ses nœuds associés.

pères. Cette évaluation n'est pas subjective. L'évaluateur devait évaluer la relation d'hyperonymie en se basant sur les définitions lexicographiques des lexèmes des vocables dans le *TLFi*[70]. Il devait évaluer si, dans la définition pour un domaine donné d'un lexème, son nœud père de la hiérarchie construite représente ou non son hyperonyme.

5.3.1. Présentation du corpus d'évaluation

Nous avons choisi comme corpus d'évaluation cent[71] vocables appartenant à dix catégories différentes[72] comme *vaisselle*, *ustensiles de cuisine, métiers, matériel de sport*, etc. Pour chaque vocable ont été générées automatiquement ses hiérarchies sémantiques qui représentent des arbres hiérarchiques. Les hiérarchies sémantiques construites contiennent un seul niveau de relations de type *is-a* et *TA*. Leur profondeur dépend plus de la profondeur du thésaurus des domaines du *TLFi* qui est lié aux arbres construits par la relation *in_domain.* Par exemple, dans la figure 5.8. l'arbre hiérarchique du lexème RAQUETTE du domaine *sports* a seulement une seule relation de type *in_domain* tandis que l'arbre du lexème RAQUETTE du domaine *chasse* contienne deux relations de ce type, ce qui fait que cet arbre est plus profond par rapport aux arbres des autres lexèmes du même vocable.

[70] Ceci pour ne pas biaiser l'évaluation par un différentiel trop grand entre les connaissances des documentalistes et celles fournies dans le *TLFi*, comme tenu de la richesse extrême de ce dictionnaire. Ainsi, la connaissance de ROSETTE comme « *Encre* rouge obtenue à partir d'une essence de bois du Brésil » (cf. définition (a) du domaine TECHNOLOGIE) était accessible par les documentalistes (cf. www.cnrtl.fr/lexicographie/rosette).
[71] La limitation du nombre de vocables choisis fut faite pour rester dans un coût raisonnable d'évaluation.
[72] Nous avons choisi volontairement des catégories différentes de celles sur lesquelles nous avions travaillé pour définir notre algorithme.

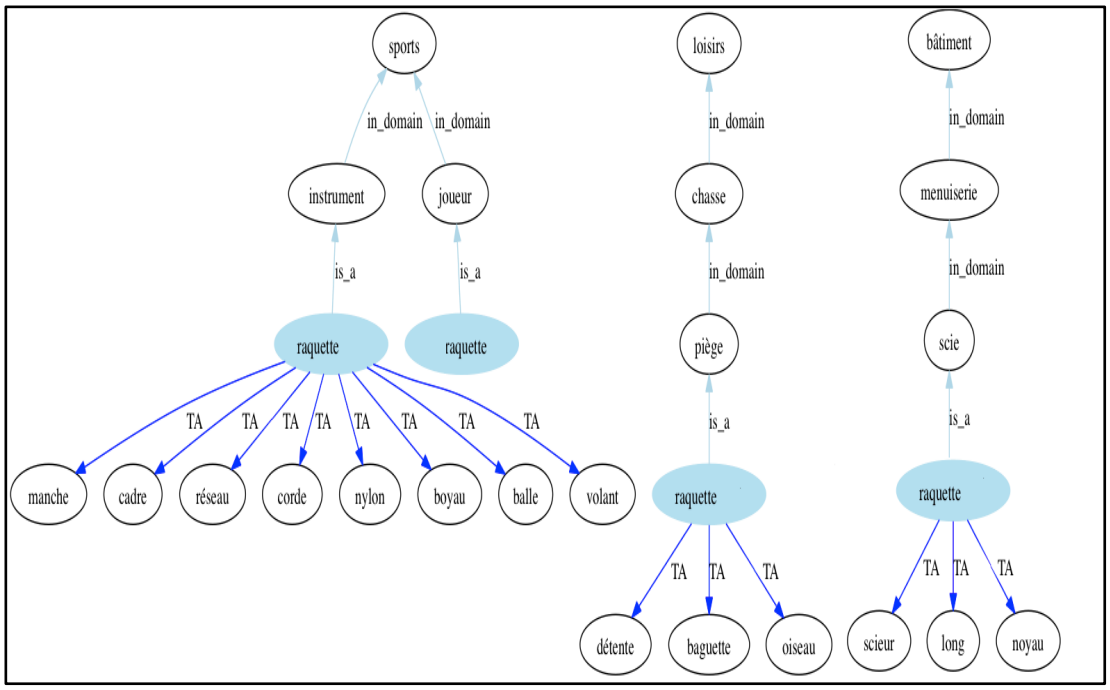


Figure 5.8. Exemple des hiérarchies sémantiques construites pour le vocable RAQUETTE

L'évaluation a été effectuée par trois documentalistes de Xilopix qui travaillent sur la conception manuelle du thésaurus de l'entreprise. Un formulaire a été mis à leur disposition. Ainsi, les évaluateurs ont pu évaluer la relation d'hyperonymie d'un lexème et son nœud père à travers deux réponses :

- « Oui » quand le nœud père représente l'hyperonyme du lexème pour un domaine donné ;
- « Non » quand le nœud père n'est pas l'hyperonyme du lexème.

5.3.2. Analyse des résultats d'évaluation

L'évaluation a été effectuée sur l'ensemble de 620 relations d'hyperonymie ainsi déterminées automatiquement. La figure 5.9. présente les résultats obtenus. Ainsi, 82,07 % des relations ont été évaluées comme étant hyperonymiques (ex. *manteau-vêtement*, *compassion-sentiment*, etc.). Toutefois, 17,92 % des relations ont été évaluées comme non hyperonymiques (ex. *batteur-branche*, *palme-forme*, etc.).

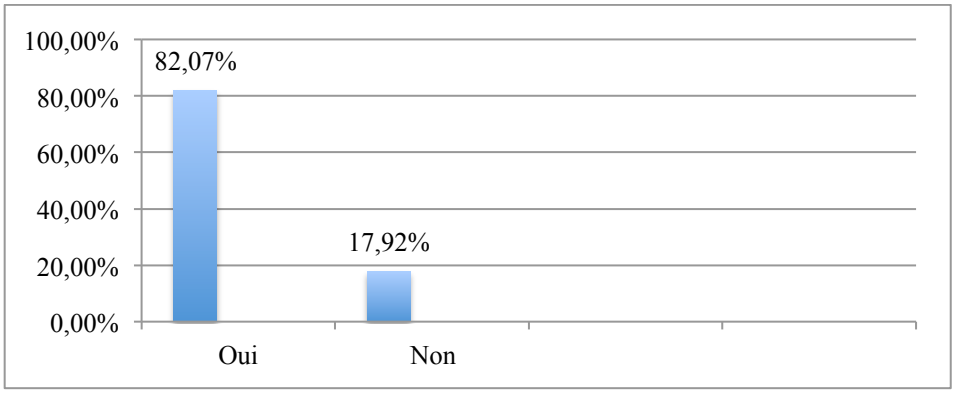


Figure 5.9. Résultats de l'évaluation manuelle des relations hyperonymiques

Nous allons maintenant présenter les analyses détaillées de l'évaluation ainsi réalisée.

5.3.2.1. Analyse des résultats positifs de l'évaluation manuelle

Le tableau 5.1. présente des exemples de relations hyperonymiques évaluées positivement.

Lexème	Hyperonyme	Domaine	Définition du *TLFi*
ROSETTE	Encre	Technologie	*Encre* rouge obtenue à partir d'une essence de bois du Brésil.
CANOË	Embarcation	Générique	*Embarcation* légère manœuvrée à la pagaie …
SKI	Patin	Sports	Chacun des deux *patins* (de bois, de métal ou de fibre synthétique) longs …
POÊLE	Appareil	Générique	*Appareil* de chauffage.
BALLON	Pneu	Cyclisme	Gros *pneu* de bicyclette.
RAQUETTE	Piège	Chasse	*Piège* à détente constitué par une baguette flexible tendue …
AMOUR	Attirance	Générique	*Attirance*, affective ou physique, qu'en raison d'une certaine affinité …

DOULEUR	Souffrance	Générique	*Souffrance* plus ou moins vive, produite par une blessure …
CHÂTEAU	Forteresse	Histoire	*Forteresse* souvent construite sur une hauteur et/ou difficilement accessible …
OBÉLISQUE	Monument	Générique	*Monument* ou objet ayant la forme d'un obélisque.

Tableau 5.1. Exemples de relations identifiées comme hyperonymiques

Ainsi, nous pouvons observer que dans les relations évaluées, l'hyperonyme déterminé pour un lexème est toujours situé en tête de la définition, même dans les définitions synonymiques ou dans le cas des définitions qui contiennent des syntagmes hyperonymiques (ex. *appareil de chauffage*).

5.3.2.2. Analyse des résultats négatifs de l'évaluation manuelle

Même si la grande majorité (82,07 %) des relations hyperonymiques déterminées automatiquement a été évaluée positivement (cf. tableau 5.1.), 17,92 % des relations d'hyperonymie identifiées ont été évaluées négativement (cf. tableau 5.2.).

Lexème	Hyperonyme	Domaine	Définition du *TLFi*
CASSEROLE	Épreuve	Générique	Subir une *épreuve* désagréable.
CHAUFFEUR	Feu	Générique	Celui qui s'occupe du *feu* d'une forge, d'un fourneau, du fonctionnement d'une chaudière.
CITADELLE	Force	Générique	Personne rappelant, par sa *force* de résistance.
FORCE	Nature	Générique	Ce qui meut, anime la *nature*, l'univers.
POUVOIR	Fonction	Générique	Droits d'exercer certaines *fonctions* et prérogatives attachées à ces droits.
SALADIER	Salade	Générique	Personne qui raconte des *salades.*
SOUCOUPE	Consomma-tion	Générique	Montant d'une *consommation* dans un café.

POMME	Excroissan-ce	Générique	*Excroissance* qui se trouve sur la sauge passifère et qui est le fruit de la piqûre d'un insecte.
COCOTTE	Familier	Générique	Appellatif *familier* adressé à une jument.
AVOCAT	Intérêt	Générique	Personne défendant les *intérêts* de quelqu'un ou de quelque chose.

Tableau 5.2. Exemples des relations identifiées comme non hyperonymiques

Nous allons maintenant détailler les différents facteurs qui ont conduit à proposer des relations hyperonymiques erronées.

1. Structure des définitions du *TLFi*.

Notre modèle de pondération fonctionne plutôt correctement à partir des définitions hyperonymiques des vocables nominaux. Toutefois, les définitions de certains lexèmes ne sont pas toujours hyperonymiques, elles peuvent en effet avoir une autre structure. C'est par exemple le cas des définitions pour les lexèmes qui désignent un métier comme CHAUFFEUR, AVOCAT ou qui désignent des choses abstraites comme FORCE. Les définitions de ces lexèmes commencent toujours par un pronom démonstratif *celui, celle, ce*, etc. ou pronom relatif qui indiquent une personne ou une chose : « Celui qui s'occupe du feu d'une forge » (CHAUFFEUR), « Ce qui meut, anime la nature, l'univers » (FORCE). Dans ce type de définition, l'hyperonyme du lexème défini est le pronom démonstratif ou relatif. Dans SEMEME, ces pronoms n'ont pas été annotés. Ce fait nous ne permet pas de les identifier et de traiter ce type de structure des définitions.

2. Syntagmes hyperonymiques.

On relève aussi des définitions qui débutent avec un verbe, par exemple l'une des définitions des lexèmes du vocable CASSEROLE est « Subir une épreuve désagréable ». La construction du type *Verbe + Nom + Adjectif* rencontrée dans la définition représente un genre prochain de celle-ci complexe, constitué de plusieurs éléments. La pondération que nous avons proposée ne permet pas de déterminer ce type de construction complexe. Ainsi, la relation hyperonymique déterminée automatiquement *casserole-épreuve* a été évaluée négativement par l'évaluateur.

Toutefois, ce fait est discutable, car cette relation pourrait être bien considérée comme une relation hyperonymique.

3. Annotation et lemmatisation erronées des définitions du SEMEME.

Lors de l'annotation morphosyntaxique des définitions du *TLFi* dans le cadre du projet SEMEME, certains lemmes ont été annotés incorrectement. C'est surtout le cas des lemmes qui peuvent avoir plusieurs catégories grammaticales (ex. *avoir, être, animal, mammifère, voisin, félin,* etc.). Ainsi, certains lemmes des définitions sont annotés avec d'autres catégories que « nom », ce qui fait qu'ils ne sont pas pris en compte lors du calcul de poids qui ne traite que les noms. Par exemple, dans la définition du lexème POUVOIR, le lemme *droit* est annoté comme adjectif. Ce sont les mêmes cas pour le lexème SOUCOUPE où le lemme *montant* est annoté comme verbe et le lexème COCOTTE où le lemme *appellatif* est annoté comme adjectif et le lemme *familier* comme nom.

4. Des lemmes manquants dans SEMEME.

Nous avons remarqué que dans SEMEME, certains lemmes des définitions manquent. Par exemple, dans une des définitions des lexèmes du vocable CITADELLE, le lemme *personne* manque dans la liste des lemmes de la définition :

<definition>
<source> **Personne** rappelant, par sa force de résistance.*</source>*
<semes>
<seme lemme="rappeler" categorie="v"/>
<seme lemme="force" categorie="subst"/>
<seme lemme="résistance" categorie="subst"/>
/semes>
</definition>

De ce fait, notre procédure automatique a déterminé comme hyperonyme du lexème CITADELLE le nom *force* et non *personne*. Le même cas de figure se présente pour les lexèmes SALADIER et AVOCAT.

5. Les définitions des locutions d'un vocable.

Dans notre recherche, nous utilisons seulement les définitions des lexèmes d'un vocable donné et non les définitions des locutions. Toutefois, dans SEMEME, certaines définitions des locutions d'un vocable donné ont été annotées comme étant des définitions des lexèmes. Ce balisage erroné a contribué à la récupération des définitions

de locutions d'un vocable donné. Par exemple, la définition « Excroissance qui se trouve sur la sauge passifère et qui est le fruit de la piqûre d'un insecte » de la locution POMME DE SAUGE a été considérée comme l'une des définitions des lexèmes du vocable POMME. De ce fait, nous déterminons des hyperonymes qui sont valables plutôt pour les locutions du vocable et non pour les lexèmes. Ainsi, le nom *excroissance* ne peut pas être considéré comme l'hyperonyme du lexème POMME mais uniquement de la locution POMME DE SAUGE.

5.4. Comparaison des hiérarchies sémantiques avec le thésaurus Xilopix

Pour notre deuxième évaluation, nous avons décidé de comparer les hiérarchies sémantiques construites avec celle du thésaurus Xilopix. Ainsi, nous avons choisi les vocables de la catégorie *flore* et *faune*, deux catégories pour lesquelles le thésaurus Xilopix est le plus développé, tant en largeur qu'en profondeur, en faisant l'hypothèse que les arbres hiérarchiques construits doivent se rapprocher de la structure du thésaurus construit manuellement. Dans un premier temps, l'évaluation des relations hyperonymiques a été faite manuellement. Dans un deuxième temps, les relations hyperonymiques obtenues ont été évaluées automatiquement. Pour ce faire, à partir des arbres hiérarchiques construits, ont été extraits les nœuds pères des lexèmes des vocables donnés qui à leur tour ont été comparés avec ceux du thésaurus Xilopix.

5.4.1. Description du corpus d'évaluation

À cet effet, nous avons choisi deux cents vocables appartenant aux catégories *faune* et *flore* pour les raisons suivantes : d'une part, les définitions des lexèmes des vocables de ces catégories par rapport aux autres ont une autre structure qui permet de construire des hiérarchies sémantiques plus profondes et, d'autre part, le thésaurus construit manuellement pour ces vocables au sein de Xilopix est le plus développé. Par conséquent pour chaque vocable, des hiérarchies sémantiques ont été construites automatiquement. Nous avons évalué manuellement les relations d'hyperonymie entre les lexèmes du vocable donné et leurs nœuds pères de hiérarchies construites pour le domaine *générique*, *zoologie (ornithologie, entomologie)* et *botanique*. Ensuite, les

relations d'hyperonymie obtenues ont été comparées automatiquement avec celles extraites à partir du thésaurus construit manuellement.

5.4.2. Analyse des résultats d'évaluation

Dans les sections suivantes, nous présentons les analyses de cette seconde évaluation.

1. *Résultats de l'évaluation manuelle des hiérarchies sémantiques du domaine flore et faune.*

Pour la catégorie *flore*, nous avons évalué en total 123 relations hyperonymiques et, pour la catégorie *faune*, 100 relations hyperonymiques. Comme dans le cas de la première évaluation manuelle, les évaluateurs ont été amenés à évaluer les relations seulement par deux types de réponses :

- « Oui » quand le nœud père représente l'hyperonyme du lexème ;
- « Non » quand le nœud père n'est pas l'hyperonyme du lexème.

La figure 5.10. présente les résultats de cette évaluation. Ainsi, 99,18 % des relations hyperonymiques de la catégorie *flore* ont été évaluées positivement et seulement 0,81 % négativement.

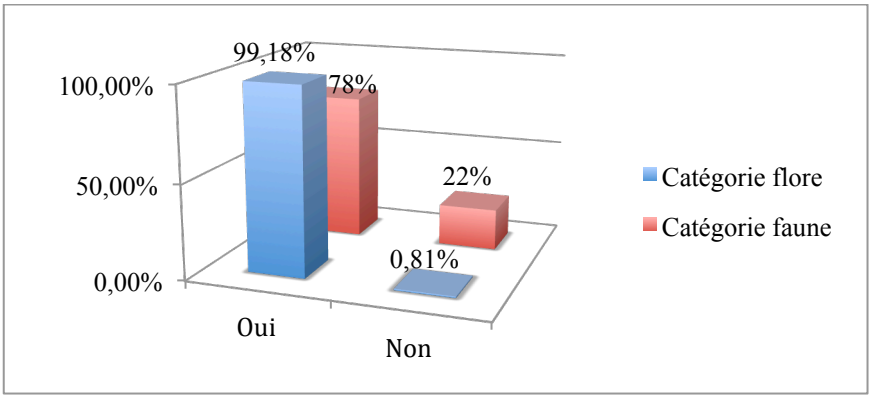


Figure 5.10. Résultats de l'évaluation manuelle des relations hyperonymiques du domaine *flore* et *faune*

La cause principale de la précision très élevée de l'évaluation des relations hyperonymiques des vocables du domaine *flore* s'explique par le fait que, dans ce domaine, les lexèmes sont définis par les noms situés toujours en tête de la définition comme *plante*, *fleur*, *arbre*, etc. De plus ce sont des noms les plus caractéristiques pour

ce domaine. Le tableau 5.3. ci-dessous présente des exemples des relations identifiées comme hyperonymiques dans le domaine *flore*.

Lexème	Hyperonyme	Domaine	Définition du *TLFi*
ANCOLIE	Plante	Botanique	*Plante* herbacée et vivace de la famille des renonculacées à fleurs de couleurs variées …
CAMOMILLE	Plante	Botanique	*Plante* connue pour ses vertus fébrifuges et digestives, de la famille des Composées.
ROMARIN	Plante	Botanique	*Plante* arbustive méditerranéenne de la famille des Labiacées …
MARGUERITE	Fleur	Botanique	*Fleur* de cette plante.
MARGUERITE	Plante	Botanique	*Plante* à fleurs de la famille des Composées, à pétales généralement blancs et à coeur jaune.
EUCALYPTUS	Arbre	Botanique	Grand *arbre* d'origine exotique aux feuilles bleuâtres longues et minces…
CITRONNIER	Arbrisseau	Botanique	*Arbrisseau* de la famille des Rutacées, haut de trois à cinq mètres …
NÉFLIER	Arbuste	Botanique	*Arbuste* donnant des nèfles, dont le bois très dur est employé à divers usages en menuiserie.
AIRELLE	Arbrisseau	Botanique	*Arbrisseau* de la famille des vacciniées qui porte une petite baie d'un noir violacé …
VIOLETE	Fleur	Botanique	*Fleur* de cette plante.

Tableau 5.3. Exemples des relations identifiées comme hyperonymiques dans le domaine *flore*

Toutefois, nous avons obtenu 0,81 % de relations hyperonymiques évaluées négativement, la cause principale étant la structure des définitions correspondantes. Par exemple, pour le lexème ACACIA « En dehors de la variété du mimosa, cet arbre exotique est surtout connu comme producteur de la gomme arabique, d'où le composé

acacia-gommier » le nom *mimosa* a été déterminé comme hyperonyme au lieu du nom *arbre* qui n'est pas situé en tête de la définition.

Pour la catégorie *faune*, seulement 78 % des relations hyperonymiques ont été évaluées positivement (cf. figure 5.10.). Ainsi, les hyperonymes des lexèmes du domaine *faune* comme ceux du domaine *flore* sont situés en tête des définitions (cf. tableau 5.4.).

Lexème	Hyperonyme	Domaine	Définition
PUMA	Mammifère	Zoologie	*Mammifère* carnassier d'Amérique, de la famille des Félidés …
BUBALE	Mammifère	Zoologie	*Mammifère* ruminant d'Afrique de la famille des antilopes …
AUTRUCHE	Échassier (genre d')	Ornithologie	Genre d'*échassiers* brévipennes vivant à l'état sauvage dans les steppes africaines …
CHAMOIS	Mammifère	Zoologie	*Mammifère* quadrupède ruminant du genre antilope, à cornes creuses et lisses …
GIRAFE	Mammifère	Zoologie	*Mammifère* ruminant, ongulé d'Afrique, que caractérisent sa haute taille …
ROUGE-QUEUE	Oiseau	Ornithologie	*Oiseau* (ordre des Passereaux, genre Fauvette, famille des Turdidés) à gorge noire …
GYPAÈTE	Vautour	Zoologie	*Vautour* de la famille des Falconidés …
OUTARDE	Échassier	Zoologie	*Échassier* au corps lourd et à fortes pattes, à long cou et à bec court …
CARCAJOU	Blaireau	Zoologie	*Blaireau* du Labrador.
LÉZARD	Reptile	Zoologie	*Reptile* saurien (de la famille des Lacertidés) …

Tableau 5.4. Exemples des relations identifiées comme hyperonymiques dans le domaine *faune*

Toutefois, 22 % des relations hyperonymiques ont été évaluées négativement principalement à cause de l'annotation erronée des définitions de SEMEME ou de l'absence des lemmes. Ainsi, les lemmes de certaines définitions, par exemple *mammifère*, *félin*, *animal*, *reptile* ont été annotés comme des adjectifs au lieu des noms ce qui n'a pas permis de les déterminer comme des hyperonymes des lexèmes OTARIE, OURS, etc. (cf. tableau 5.5.). En même temps les lemmes *martre* et *reptile* des définitions des lexèmes PÉKAN et respectivement ORVET sont absents dans la liste des lemmes des définitions du SEMEME.

Lexème	Hyperonyme	Domaine	Définition du *TLFi*
OTARIE	Marin	Zoologie	Mammifère *marin* pinnipède, au corps fusiforme, à la tête petite, allongée …
OURS	Corps	Générique	Mammifère au *corps* volumineux et massif, à fourrure épaisse …
LYNX	Patte	Générique	Félin, haut sur *pattes*, de la taille d'un gros chat, au pelage roux,…
CROCODILE	Vertébré	Zoologie	Animal *vertébré*, reptile de grande taille, à corps allongé couvert d'écailles …
ORVET	Saurien	Générique	Reptile *saurien* sans pattes, ovovivipare, insectivore et inoffensif …
PÉKAN	Canada	Générique	Martre du *Canada*.

Tableau 5.5. Exemples des relations identifiées comme non hyperonymiques dans le domaine *faune*

2. Comparaison des structures hyperonymiques avec ceux du thésaurus Xilopix.

Les relations hyperonymiques obtenues ont été comparées automatiquement avec celles extraites à partir du thésaurus construit manuellement pour les vocables donnés. Les résultats ont coïncidé seulement pour 9 % de relations où l'hyperonyme du lexème représentait la famille ou le genre auquel appartient le lexème du vocable (cf. figure 5.11., tableau 5.6.).

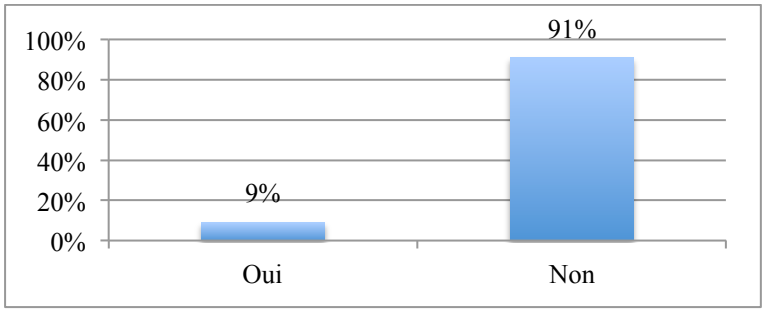


Figure 5.11. Coïncidence des structures hyperonymiques construites manuellement et celles construites automatiquement

Relations hyperonymiques	
Nénuphar – Nymphéacée	Éléphant – Pachyderme
Amaryllis – Amaryllidées	Perroquet – Psittacidés
Bananier – Musacées	Pigeon – Columbidés
Oranger – Citrus	Chauve-souris – Chéiroptères

Tableau 5.6. Exemple des relations hyperonymiques existantes dans les deux hiérarchies

Le thésaurus construit manuellement pour la *flore* et *faune* est plus profond, il contient plusieurs niveaux qui indiquent l'ordre, la famille, le genre, l'espèce, etc[73]. Alors que, pour notre part, seul 9 % de vocables ont conduit à la construction de hiérarchies sémantiques à deux niveaux de relations hyperonymiques où le premier niveau représente la classe (ex. *plante, mammifère*, etc.) du lexème et le deuxième niveau représente la famille (ex. *broméliacées, musacées,* etc.) à laquelle appartient le lexème du vocable donné.

5.5. Proposition de méthodologie d'enrichissement du thésaurus Xilopix

Les hiérarchies sémantiques construites automatiquement pour chaque vocable du *TLFi* peuvent être utilisées pour l'enrichissement semi-automatique du thésaurus construit manuellement par l'équipe de documentalistes de Xilopix. La principale limite

[73] Dans le thésaurus, pour le domaine *flore* et *faun*e plus de 20 niveaux de représentation sont possibles.

du thésaurus existant est qu'il n'est pas développé de la même manière pour tous les domaines. Par exemple, la structure des micro-thésaurus des domaines *flore* et *faune* contient beaucoup plus de niveaux que celle des autres domaines. Par ailleurs, on a pu remarquer que, lors de construction manuelle du thésaurus, les documentalistes ne prennent pas toujours en compte tous les sens possibles d'un terme. Ainsi, dans le thésaurus existant le terme *lion* apparaît seulement dans les micro-thésaurus de trois domaines *faune*, *astrologie* et *architecture*, tandis que dans les hiérarchies sémantiques construites à partir du *TLFi*, il appartient à sept domaines différents comme le domaine *numismatique*, *héraldique*, *astronomie*, *iconographie*, etc. La méthodologie d'enrichissement du thésaurus existant par les hiérarchies sémantiques construites automatiquement que nous proposons est la suivante :

1. Pour chaque terme du thésaurus existant, on construit ses hiérarchies sémantiques à partir du *TLFi*. Ensuite, on détermine dans les arbres obtenus les nœuds pères du terme donné ainsi que les domaines et on vérifie s'ils existent dans le thésaurus actuel. Dans le cas où ils n'existent pas, une liste avec de nouvelles relations et de nouveaux domaines est proposée aux documentalistes qui peuvent l'accepter ou la rejeter.
2. Les hiérarchies sémantiques générées à partir du *TLFi* pour chaque vocable peuvent être utilisées comme une ressource linguistique directe, ce qui permettra aux documentalistes d'avoir immédiatement une vision globale sur les possibles sens d'un terme et ses domaines d'utilisation.

Cette méthodologie d'enrichissement proposée sera implémentée lors de la mise en production de notre prototype par l'entreprise.

5.6. Conclusion

Dans ce chapitre, nous avons proposé une méthodologie de construction automatique de hiérarchies sémantiques pour chaque vocable du *TLFi*. Nous avons réalisé deux évaluations manuelles des relations hyperonymiques construites sur des corpus différents et avons ensuite comparé les hiérarchies sémantiques construites pour les vocables des catégories *flore* et *faune* avec le thésaurus Xilopix. Au terme de cette évaluation, la meilleure précision (99,18 %) a été obtenue lors de l'évaluation manuelle des relations sémantiques des vocables de la catégorie *flore*. La détermination de

relations non hyperonimiques est principalement due à des imperfections (annotations morphosyntaxiques, lemmes manquants) de la ressource initiale utilisée SEMEME et à la structure des définitions.

En conclusion, les évaluations de relations hyperonymiques nous ont permis de valider notre algorithme de construction automatique de hiérarchies sémantiques ainsi que notre formule de pondération des noms, en montrant de plus que la position dans les définitions du *TLFi* est l'un des facteurs importants pour la détermination des hyperonymes. Nous avons par ailleurs proposé une méthodologie d'utilisation des hiérarchies sémantiques construites pour l'enrichissement du thésaurus actuel de l'entreprise.

Dans le chapitre suivant, nous allons montrer, à titre d'exemple, une application concrète des hiérarchies sémantiques construites à partir du *TLFi* pour l'indexation et la recherche d'images.

CHAPITRE 6

Exploitation des hiérarchies sémantiques du *TLFi* pour l'indexation et la recherche d'images

Sommaire

6.1. Intégration des relations *is-a* dans un algorithme simple d'indexation textuelle automatique d'images 177
6.1.1. Évaluation des performances du prototype d'indexation 180
6.1.2. Discussion sur les termes d'indexation déterminés 182
6.2. Prise en compte de relations d'association dans un algorithme simple d'indexation textuelle automatique d'images 188
6.2.1. Analyse des résultats d'évaluation de ce second algorithme 189
6.2.2. Bilan sur l'indexation d'images 192
6.3. Exploitation de l'indexation dans un algorithme simple de recherche d'images 193
6.3.1. Évaluation des résultats de recherche 195
6.3.2. Analyse des résultats d'évaluation 198
6.3.2.1. Domaines erronés 198
6.3.2.2. Relations d'association 200
6.3.3. Avantages de notre approche de recherche d'images 201
6.3.3.1. Structuration des résultats de recherche selon les domaines 201
6.3.3.2. Recherche associative d'images 203
6.3.3.3. Requêtes complexes 204
6.3.4. Bilan sur la recherche d'images 205
6.4. Conclusion 205

Chapitre 6. Exploitation des hiérarchies sémantiques du *TLFi* pour l'indexation et la recherche d'images

La finalité de notre travail porte sur l'application de ressources lexicales, et plus particulièrement lexicographiques, pour l'indexation et la recherche d'images. Nous avons choisi le *TLFi* comme ressource lexicale principale. Étant donné que l'information contenue dans le *TLFi* n'est pas explicite, nous avons présenté, dans les chapitres précédents, une heuristique de pondération des noms des définitions des lexèmes du *TLFi* (cf. §4.2.3.3.), qui nous a permis d'expliciter l'information du *TLFi* en construisant de hiérarchies sémantiques pour les lexèmes d'un vocable donné du *TLFi* (cf. §5.2.).

Ce chapitre a comme objectif de montrer que les hiérarchies sémantiques obtenues à partir du *TLFi* sont applicables dans le domaine de l'indexation et de la recherche d'images. Pour ce faire, nous avons mis en place un prototype d'indexation et de recherche d'images. Nous commençons donc par présenter dans un premier temps, un algorithme d'indexation textuelle automatique des images qui intègre des relations *is-a* déterminées, en montrant les résultats de l'évaluation réalisée ainsi que les analyses effectuées. Dans un deuxième temps, nous présentons un algorithme d'indexation d'images qui prend en compte les relations d'association. Ensuite, nous présentons un algorithme de recherche d'images qui s'appuie sur les indexations réalisées et analysons des résultats d'évaluation. Enfin, avant de conclure, nous décrivons les principaux avantages apportés par l'utilisation des hiérarchies sémantiques du *TLFi* dans un processus d'indexation et de recherche d'images.

6.1. Intégration des relations *is-a* dans un algorithme simple d'indexation textuelle automatique d'images

Dans le cadre de Xilopix jusqu'à 2011, l'indexation des images de la photothèque se réalisait manuellement par les auteurs des images. Ce qui était une tâche coûteuse en temps. Ainsi pour indexer une image il fallait tout d'abord l'annoter avec une liste des

mots-clés et ensuite de trouver des correspondances avec les termes du thésaurus de l'entreprise. Afin d'automatiser le processus d'indexation à l'aide de hiérarchies sémantiques du *TLFi* nous présentons un algorithme simple d'indexation automatique d'images à partir de leurs descriptions textuelles associées. L'objectif est de trouver, parmi les mots-clés de la description de l'image, les termes d'indexation qui correspondent aux concepts principaux qui y sont représentés. Nous faisons l'hypothèse que dans les hiérarchies sémantiques construites, les concepts principaux de l'image doivent être reliés par des relations sémantiques. Pour ce faire, l'algorithme d'indexation, après un filtrage de mots-clés, exploite pour chaque image les relations des hiérarchies sémantiques obtenues à partir du *TLFi* pour les mots-clés de la description textuelle associée. Le résultat de cet algorithme d'indexation est une liste de termes d'indexation (descripteurs de l'image) ordonnés par spécificité décroissante associée à leur domaine d'utilisation. La figure 6.1. présente une vue globale de l'algorithme d'indexation automatique d'images à travers un exemple d'image à indexer représentant le fruit *ananas*.

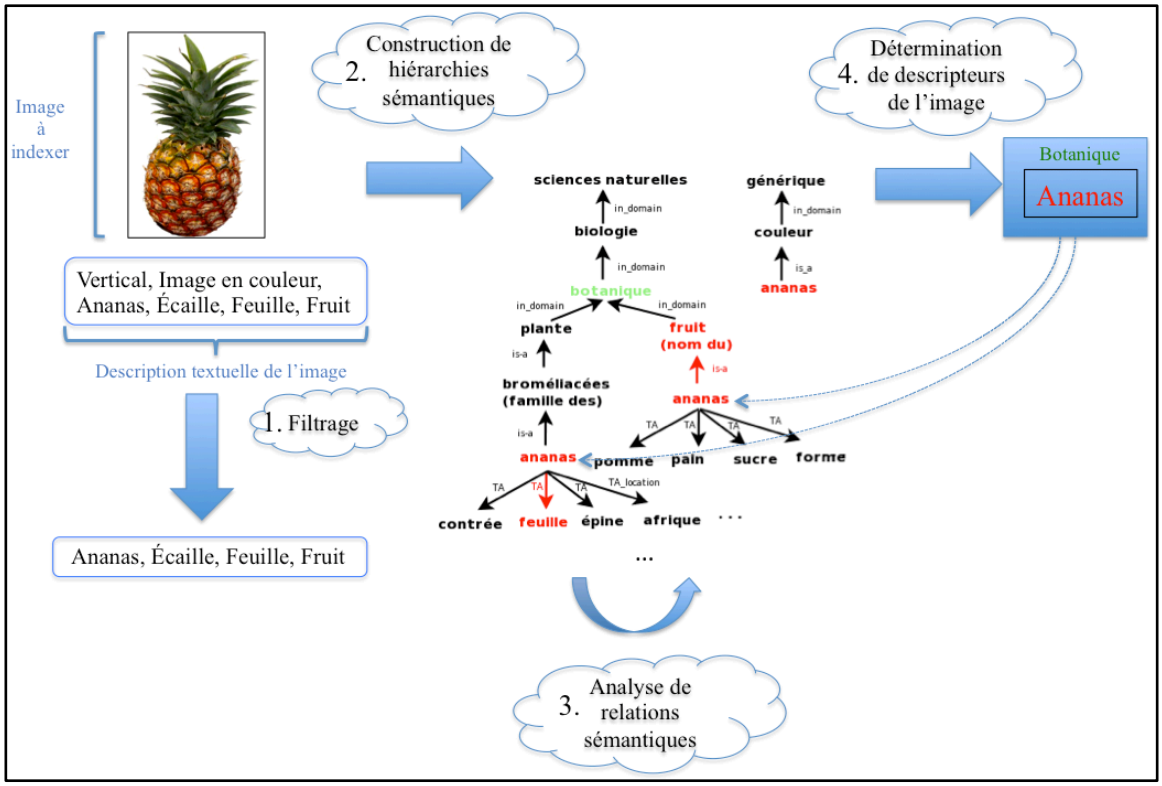


Figure 6.1. Vue globale de l'algorithme d'indexation automatique d'images

Afin d'analyser l'influence sur l'indexation des images de chaque type de relations existantes dans les hiérarchies sémantiques construites à partir du *TLFi*, nous présentons ci-dessous tout d'abord un algorithme d'indexation qui prend en compte dans un premier temps les relations de type *is-a* et, seulement dans un second temps, les relations de type *TA* (termes associés) :

1. Lemmatisation et annotation morphosyntaxique des mots-clés de l'image en utilisant le lexique morphosyntaxique Morphalou[74], puis filtrage par élimination des mots-clés appartenant aux catégories autres que *nom* et ceux faisant partie du métalangage de description d'une prise de vue (ex. : *Horizontal*, *Cadrage en pied*, *Prise de vue en extérieur*, *Plan rapproché*, *Vue latérale*, etc.).

2. Pour chaque mot-clé de l'image à indexer, analyse de ses arbres hiérarchiques construits préalablement en utilisant notre algorithme de construction automatique de hiérarchies sémantiques (cf. chapitre 5).

3. À partir des arbres hiérarchiques de chaque mot-clé, extraction des relations de filiation :

SI détermination des relations de filiation où les nœuds fils et les nœuds pères représentent les mots-clés de l'image à indexer (cf. figure 6.1., *fruit-ananas*, *feuille-ananas*)

ALORS détermination des filiations où les nœuds sont liés par la relation *is-a* (ex. *ananas-fruit*)

SI telles relations existent

ALORS indexation de l'image au nœud fils de la relation déterminée (ex. *ananas*)

SINON détermination des filiations où les nœuds sont liés par la relation *TA* (ex. *feuille-ananas)*

SI telles relations existent

ALORS indexation de l'image au nœud père de la relation déterminée (ex. *ananas*)

FIN SI

[74] Morphalou (http://www.cnrtl.fr/lexiques/morphalou/) est un lexique ouvert des formes fléchies du français qui contiennent 539,413 formes fléchies, appartenant à 68,075 lemmes.

FIN SI

SINON indexation de l'image par les trois premiers [75] mots-clés de la description textuelle de l'image (ex. *ananas, écaille, feuille*)

FIN SI

Nous présentons par la suite l'évaluation de l'algorithme proposé d'indexation d'images.

6.1.1. Évaluation des performances du prototype d'indexation

La plupart des travaux n'évaluent l'indexation qu'à travers les résultats de la recherche, l'évaluation de l'indexation elle-même n'étant pas effectuée. Cette problématique a été abordée dans les travaux de Soergel (1994), Kefi, Berrut et Gaussier (2005) (cf. §2.3.1.). Dans notre approche, nous avons décidé d'évaluer séparément les deux processus d'indexation et de recherche pour, d'une part, mieux mettre en évidence l'apport des hiérarchies sémantiques construites automatiquement à partir du *TLFi* dans chacun de ces processus et, d'autre part, de cibler les problèmes demeurant ouverts dans une telle approche.

Afin d'évaluer l'algorithme d'indexation proposé ci-dessus, un prototype a été mis en place pour valider la qualité des termes d'indexation déterminés par cet algorithme pour chaque image. Pour cette évaluation, nous avons utilisé un corpus constitué de 503 images avec leurs descriptions textuelles extraites de Getty Images[76], appartenant à différents domaines (ex. *flore, faune, architecture, jeux, sport,* etc.). Nous avons choisi d'évaluer les images de Getty car les descriptions textuelles associées aux images nous sont apparues pertinentes, étant attribuées semble-t-il manuellement. Dans un premier temps, l'évaluation a été effectuée par un documentaliste qui s'occupe de l'indexation manuelle des images au sein de l'entreprise Xilopix qui devait préciser si les termes d'indexation déterminés par notre algorithme pour chaque image correspondent ou non aux concepts généraux représentés dans l'image. Dans le cas où le terme représente le concept de l'image, l'évaluateur devait préciser aussi le niveau de représentation selon

[75] L'ordre de termes est déterminé en calculant leur position dans la chaîne de caractères (la description textuelle de l'image), en prenant seulement les trois termes dont la position est croissante.
[76] Getty Images : banque d'images http://www.gettyimages.fr/

le modèle de Shatford (1986) : générique (TG), spécifique (TS) ou à propos (TAP) (cf. §2.2.1.). Dans un deuxième temps, nous avons comparé les termes d'indexation obtenus par notre processus d'indexation automatique avec ceux générés manuellement par un documentaliste à partir la description textuelle associée à l'image. Ainsi, nous utilisons l'indexation manuelle comme référence afin de déterminer si tous les termes d'indexation ont été bien extraits.

En analysant plusieurs mesures utilisées dans l'état de l'art (cf. §2.3.1.) pour l'évaluation des performances de l'indexation d'un système, nous n'avons retenu que deux mesures pour évaluer la qualité de l'indexation de notre prototype : la complétude et la justesse.

- La complétude de l'indexation indique la présence des termes d'indexation corrects. Elle est corrélée au rappel de la recherche. La complétude est le nombre d'images correctement indexées par le terme *t* rapporté au nombre d'images qui devraient être indexées par le terme *t*.
- La justesse de l'indexation est corrélée à la précision de la recherche. C'est le nombre d'images correctement indexées par le terme *t* rapporté au nombre total d'images indexées par le terme *t*. La justesse permet de déterminer le pourcentage de termes corrects affectés à l'image.

Les résultats de l'évaluation de l'indexation sont présentés dans le tableau 6.1.

	Exactitude des termes d'indexation
Complétude	0,59
Justesse	0,91

Tableau 6.1. Résultats du processus d'indexation intégrant des relations *is-a*

Nous obtenons une justesse d'affectation des termes d'indexation très élevée, de 91 %. Toutefois, la complétude des termes d'indexation est moins satisfaisante. Parmi tous les termes d'indexation qui devraient être affectés aux images, seulement 59 % ont été affectés. Cela s'explique, d'une part, par le manque de relations sémantiques directes entre certains mots-clés de la description textuelle de l'image (ex. *jaguar-animal*), ce qui n'a pas permis la détermination des termes d'indexation et, d'une autre

part, par le fait que notre algorithme privilégie dans un premier temps les relations hyperonymiques, même si certaines relations d'association ont été déterminées.

Dans la section suivante, nous présentons une analyse détaillée des termes d'indexation déterminés.

6.1.2. Discussion sur les termes d'indexation déterminés

L'évaluation de la qualité de termes d'indexation déterminés par notre prototype d'indexation montre que 91,51 % des termes d'indexation sont déterminés correctement. Le tableau 6.2. présente des exemples d'images avec les termes d'indexation par lesquels elles ont été indexées et leur niveau de représentation évalué par l'évaluateur. Nous voyons que les termes d'indexation obtenus représentent bien les concepts principaux des images. De plus, les hiérarchies sémantiques construites à partir du *TLFi* nous permettent en même temps de déterminer les domaines d'utilisation des termes d'indexation et par conséquent de les désambiguïser. Par exemple, l'image numéro 1 (cf. tableau 6.2.) a été indexée au domaine *botanique* et l'image numéro 2 au domaine *générique*[77]. Toutefois, à ce niveau nous ne sommes pas intéressée d'évaluer le domaine déterminé. Nous le ferons dans les sections suivantes (cf. §6.3.1.) à travers le processus de recherche.

Nr.	Image	Mots-clés de l'image	Termes d'indexation et le niveau de représentation
1.		Simplicité, Alimentation et boisson, Vertical, Prise de vue en studio, Aliment en portion, Fruit, Graine, Inachevé, Avocat, Image en couleur, Coupe transversale, Sans personnage, Photographie, Hygiène alimentaire, Fond blanc, Un seul objet	Fruit TG Avocat TS

[77] Ceci grâce aux domaines du *TLFi* du terme d'indexation *avocat*.

2.		Autorité, Confiance en soi, Concentration, Contemplation, Bureau, Vertical, Cadrage à la taille, Prise de vue en intérieur, Document, Personne de race blanche, Noir américain, Groupe multi-ethnique, Micro, Être debout, Gestes, Parler, Expliquer, Écouter, Avocat, Palais de justice, Droit, Adulte d'âge mûr, Trentenaire, Persuasion, Verre, Image en couleur, Série, Deux personnes, Hommes d'âge moyen, Femmes d'âge mûr, Photographie, Mains dans les poches, Costume complet, Seulement des adultes	Adulte TG Avocat TS Homme TS Trentenaire TG
3.		Vêtements décontractés, Ville, Vertical, Cadrage en pied, Prise de vue en extérieur, 2-3 ans, Cheveux blonds, Affolé, Personne de race blanche, Banlieue pavillonnaire, Assis, Rue, Jour, Enfance, Accident bénin, Divertissement, Image en couleur, Trottinette, Deux personnes, Petits garçons, Petites filles, Bâtiment vu de l'extérieur, Temps libre, Photographie, Visage expressif, Seulement des enfants, Monter un animal ou sur un moyen de transport	Enfant TG Petit TG Garçon TS Fille TS Enfance TAP
4.		Vêtements décontractés, Brillant, Jeans, Transport, Carré, Prise de vue en studio, 4-5 ans, Partie inférieure, Jambe humaine, Pied humain, Activité de loisirs, Bleu, En métal, Jour, Enfance, Une seule personne,	Enfant TG Garçon TS Trottinette TS Enfance TAP

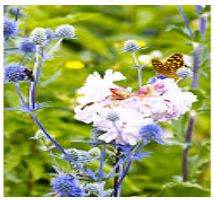
		Image en couleur, Argenté, Trottinette, Un seul petit garçon, Temps libre, Photographie, Chaussure en toile, Seulement des enfants, Fond coloré.	
5.		Vertical, Insecte, Abeille, Papillon, Fleur, Chardon bleu des dunes, Image en couleur, Photographie, Œillet, Symbiose, Saponaire officinale, Juillet, Apiécée	Papillon TG Abeille TG Œillet TS
6.		Architecture, Horizontal, Prise de vue en extérieur, Dôme, Mosquée, Nuage, Crépuscule, Brunei, Bandar Seri Begawan, Islam, Image en couleur, Minaret, Mosquée Omar Ali Saifuddin, Sans personnage, Bâtiment vu de l'extérieur, Photographie, Capitales internationales	Minaret TG
7.		Ensemble, Horizontal, Prise de vue en extérieur, Natation, Partie du corps d'un animal, Tête d'un animal, Manger, Attraper, Honduras, Poisson, Grand dauphin, Jour, Mer des Caraïbes, Image en couleur, Roatan, Avoir faim, Deux animaux, Sans personnage, Photographie, Monde marin, Remonter à la surface, Bouche ouverte	Poisson TS

Tableau 6.2. Exemples d'images avec les termes d'indexation par lesquels elles ont été indexées

La figure 6.2. montre toutefois que la plupart des termes d'indexation déterminés (65,04 %) représentent le niveau générique en décrivant les objets généraux et les

actions. Ceci est tout à fait normal, car généralement une image représente plutôt des objets génériques (ex. *portrait d'une femme, musée, ville, cathédrale,* etc.) qui, ensemble, forment l'idée générale de l'image (ex. *la bataille de Stalingrad, le portrait de la Joconde, la cathédrale Notre-Dame de Paris,* etc.), ou font référence à certains symboles, abstractions ou émotions.

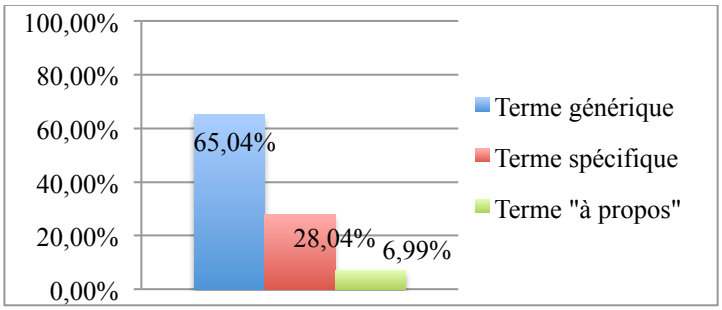


Figure 6.2. Évaluation des niveaux de représentation des termes d'indexation

Ces résultats peuvent aussi s'expliquer par :

1. Le niveau culturel de l'annotateur. La plupart des images ne contiennent pas dans leur description textuelle de mots-clés spécifiques. C'est par exemple le cas des images du domaine *flore*, où le nom précis de la fleur ou de l'arbre n'est souvent pas spécifié par méconnaissance de la part de l'annotateur. C'est par exemple le cas de l'image 5 du tableau 6.2., pour laquelle seul le terme générique *fleur* a été déterminé.
2. Des problèmes de vocabulaire. Par exemple, les images de *lion*, *jaguar* contiennent dans leur description textuelle le mot-clé *animal*, tandis que dans le *TLFi* les lexèmes LION et JAGUAR sont définis comme des *mammifères*. Cela ne permet pas de trouver directement la relation d'hyperonymie entre les termes descriptifs de l'image, *lion* et *animal*, et donc par application de notre algorithme ces images ne seront indexées que par les trois premiers termes de la description.
3. Noms propres. Les termes d'indexation spécifiques sont souvent des noms propres. L'absence d'une nomenclature avec les noms propres ne nous permet pas de les déterminer. Par exemple, l'image 6 (cf. tableau 6.2.) contient dans sa description le nom de la mosquée (*Omar Ali Saifuddin*), qui n'a pas pu être déterminé.

Cependant, nous avons observé que 8,48 % des termes ont été mal déterminés. Les causes principales qui conduisent à une indexation par des termes incorrects sont :

1. Les annotations qui contiennent des termes relevant plutôt du métalangage de description de l'image (temps, lieu) que de son contenu. Par exemple, pour l'image 1 du tableau 6.3., le terme d'indexation déterminé est *sport*, car dans sa description textuelle l'image ne contient aucun terme relevant de son contenu propre.
2. L'ambiguïté des mots-clés descriptifs de l'image. Un nom peut avoir des sens différents comme *rose*, qui peut référer à la *couleur* ou à la *fleur*. Par exemple, l'annotation de l'image 2 (cf. tableau 6.3.) contient le nom *rose* qui fait dans ce cas référence à la *couleur*. Toutefois, du fait que le nom *couleur* a été éliminé parce qu'il fait partie de la description d'une prise de vue et non de l'image, le système a déterminé la relation de type *is-a* entre les noms *rose-fleur* et non *magnolia-fleur*[78]. Par conséquent, le nom *rose* a été déterminé comme terme d'indexation, ce qui n'est pas correct.
3. Le manque de relations sémantiques entre les mots-clés de l'image. Ainsi, vu le manque de relations sémantiques entre les noms *désert* et *cactus*, l'image 3 (cf. tableau 6.3.) a été indexée par terme *golf*.
4. La non prise en compte des locutions. Notre prototype d'indexation n'étant pas capable de reconnaître les locutions nominales, il détermine plusieurs termes d'indexation au lieu d'un seul. Par exemple, pour l'image numéro 4 (cf. tableau 6.3.) le terme d'indexation déterminé est *lion* et non *lion de mer*. Ce problème découle du fait que les locutions du *TLFi* ne sont pas normalisées, ce qui ne permet pas actuellement leur reconnaissance dans les descriptions des images.

Nr.	Image	Mots-clés de l'image	Terme d'indexation et son évaluation
1.		Plein, Sport, Horizontal, Football américain, États-Unis, Stade, Californie, San Diego, NFL, Denver Broncos	Sport NON

[78] Dans le *TLFi*, le lexème MAGNOLIA est défini comme un « Arbre ou arbrisseau de la famille des Magnoliacées » et non comme la *fleur* de cet arbre ou arbrisseau.

2.		Horizontal, Prise de vue en extérieur, Plan rapproché, Flore, Gris, Rose, Blanc, Ouvert, Fleur, Corolle, Jour, Fond, Image en couleur, Sans personnage, Photographie, Une seule fleur, Magnolia	Rose NON
3.		S'élever, Soleil, Personne humaine, Sport, Horizontal, Image en noir et blanc, Golf, États-Unis, Cactus, Désert, Arizona, Parcours de golf, Une seule personne, Terme sportif, Scottsdale, Lever de soleil, Silhouette	Golf NON
4.		Nature, Vie sauvage, Horizontal, Cadrage aux genoux, Cadrage en pied, Prise de vue en extérieur, Namibie, Comportement animal, Faune sauvage, Mammifère, Lion de mer, Rocher, Jour, Mer, Plage, Swakopmund, Grand groupe d'animaux, Thème des animaux, Sans personnage, Photographie, Monde marin, Premier plan net, Bouche ouverte	Lion NON

Tableau 6.3. Exemples des images indexées par les termes d'indexation incorrects

Afin de voir si la prise en compte de relations d'association peut améliorer les résultats, nous avons aussi testé un second algorithme d'indexation qui considère comme termes d'indexation tous les termes retrouvés par le prototype et nous avons procédé ensuite à une seconde évaluation manuelle.

6.2. Prise en compte de relations d'association dans un algorithme simple d'indexation textuelle automatique d'images

La précision assez élevée obtenue lors de la première évaluation est due au fait que l'algorithme d'indexation privilégie les relations d'hyperonymie de type *is-a*, les autres relations n'étant pas prises en compte si une relation *is-a* est trouvée. Pour augmenter la complétude (rappel) de l'indexation, nous avons modifié l'algorithme afin que toutes les relations sémantiques déterminées soient considérées. La deuxième version de l'algorithme est présentée ci-dessous, il ne diffère de l'algorithme présenté dans la section §6.1. que par la seule étape 3 :

1. Lemmatisation et annotation morphosyntaxique des mots-clés de l'image en utilisant le lexique morphosyntaxique Morphalou, puis filtrage par élimination des mots-clés appartenant aux catégories autres que *nom* et ceux faisant partie du métalangage de description d'une prise de vue (ex. : *Horizontal*, *Cadrage en pied*, *Prise de vue en extérieur*, *Plan rapproché*, *Vue latérale*, etc.).

2. Pour chaque mot-clé de l'image à indexer, analyse de ses arbres hiérarchiques construits préalablement en utilisant notre algorithme de construction automatique de hiérarchies sémantiques (cf. chapitre 5).

3. À partir des arbres hiérarchiques de chaque mot-clé, extraction des relations de filiation :

SI détermination des relations de filiation où les nœuds fils et les nœuds pères représentent les mots-clés de l'image à indexer,

ALORS

POUR CHAQUE relation *is-a*

RÉPÉTER indexation de l'image au nœud fils de la relation déterminée

FIN POUR

POUR CHAQUE relation *TA*

RÉPÉTER indexation de l'image au nœud père de la relation déterminée

FIN POUR

SINON l'image est indexée par les trois premiers mots-clés de la description textuelle de l'image

FIN SI

Nous présentons ci-dessous l'analyse des résultats d'évaluation.

6.2.1. Analyse des résultats d'évaluation de ce second algorithme

Lors de l'évaluation de cette seconde version de notre algorithme, nous avons utilisé le même corpus et la même procédure d'évaluation que lors de l'algorithme d'indexation initial intégrant les relations *is-a*.

L'analyse des résultats d'évaluation nous a montré que la prise en compte de relations d'association construites automatiquement à partir du *TLFi*, a contribué à l'augmentation du nombre des termes d'indexation de 1 392 à 3 153, où 73,04 % des termes d'indexation représentent le niveau générique de l'image. C'est pourquoi les images du tableau 6.5. obtiennent davantage de termes d'indexation que ceux du tableau 6.2.

Nous avons aussi mesuré la qualité de notre prototype d'indexation qui prend en compte les relations d'association, en calculant la moyenne globale de la complétude et de la justesse pour tous les termes d'indexation déterminés. Les valeurs obtenues lors de cette nouvelle évaluation sont présentées dans le tableau 6.4.

	Exactitude des termes
Complétude	0,68
Justesse	0,79

Tableau 6.4. Résultats de l'évaluation du processus d'indexation intégrant les relations d'association

Les résultats obtenus (cf. tableau 6.4.) montrent une augmentation de la complétude de seulement 9 %, mais une baisse de la justesse de 12 %. Cela montre, semble-t-il, que la prise en compte des relations d'association contribue à l'augmentation de la complétude, mais induit alors une diminution de la précision. La principale cause de la

baisse du taux de précision (pourcentage de termes d'indexation déterminés correctement) réside dans le fait que la prise en compte des relations d'association lors de l'indexation a contribué à l'émergence de termes qui ne représentent pas les concepts de l'image. Par exemple, plusieurs termes d'indexation ont été déterminés pour l'image 2 (cf. tableau 6.5.) par rapport à l'indexation intégrant les relations *is-a*, parmi lesquels certains ne représentent pas les concepts de l'image comme *avoir, ombre*. Les images 1 et 5 (cf. tableau 6.5.) sont dans le même cas. En contrepartie, la meilleure prise en compte des relations d'association a permis la détermination d'index relevant bien de l'image, comme *dauphin* pour l'image 1, *bague* et *solitaire* pour l'image 3, *acacia* et *désert* pour l'image 4 et *poisson* pour l'image 5 (cf. tableau 6.5.).

Nr.	Image	Mots-clés de l'image	Termes d'indexation II	Termes d'indexation I
1.		Ensemble, Horizontal, Prise de vue en extérieur, Natation, Partie du corps d'un animal, Tête d'un animal, Manger, Attraper, Honduras, Poisson, Grand dauphin, Jour, Mer des Caraïbes, Image en couleur, Roatan, Avoir faim, Deux animaux, Sans personnage, Photographie, Monde marin, Remonter à la surface, Bouche ouverte	Poisson TS Bouche TG Monde NON Surface NON Natation TAP Dauphin TS Mer TG Animal TG	Poisson TS
2.		Personne humaine, Temps qui passe, Exactitude, Horizontal, Nombre, Vue en plongée, Vue de dos, Horloge, Être debout, États-Unis, Ombre, New York City, Une seule personne, Cadran d'horloge, Enfant, Aiguille de montre, Personnes masculines, Un	Enfant TG Garçon TS Montre TS Horloge TS Être TG Cadran TG Nombre TG Avoir NON	Enfant TG Garçon TS

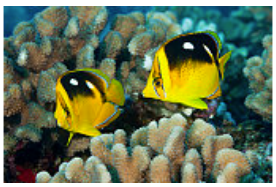
		seul petit garçon, Photographie, Premier plan net, Seulement des enfants, Avoir une bonne idée	Ombre NON Idée NON	
3.		Objet, Luxe, Liens affectifs, S'impliquer à fond, Bijou, Finance, Bague de fiançailles, Gemme, Pierre précieuse, Coffret à bijoux, Sans personnage, Solitaire, La vie chère	Gemme TS Bague TG Solitaire TS	Gemme TS
4.		Chaleur, Être seul, Mort, Danger, La Fin, Nature, Horizontal, Prise de vue en extérieur, Vue en contre-plongée, Namibie, Flore, Rouge, Arbre, Végétation morte, Ciel, Branche, Nuage, Aride, Désert, Jour, Paysage aride, Dune de sable, Sécheresse, Désert du Namib, Paysages, Craquelé, Sec, Absence, Lac asséché, Sans personnage, Acacia de la girafe, Dead Vlei, Parc national de Sossusvlei	Sec TG Mort NON Acacia TS Dune TG Désert TS Sable TG Végétation TG Sécheresse TAP	Sec TG
5.		Sécurité, États-Unis, Faune sauvage, Corail, Récif corallien, Jour, Îles Hawaï, Îles du Pacifique, Protection, Image en couleur, Beauté de la nature, Groupe moyen d'animaux, Sans personnage, Photographie, Monde marin, Poisson papillon à 4 taches, Au fond de l'océan	Corail TG Poisson TS Récif TG Île NON Sécurité NON Animal TG	Corail TG

Tableau 6.5. Exemples des termes d'indexation déterminés par l'intégration de relations *is-a* (termes I) et la prise en compte de relations d'association (termes II)

6.2.2. Bilan sur l'indexation d'images

Nous avons proposé un algorithme d'indexation automatique d'images qui explore les descriptions textuelles associées aux images afin de déterminer les termes d'indexation correspondants présentant les concepts principaux de l'image. Pour ce faire, le système d'indexation s'appuie sur les relations sémantiques entre les mots-clés extraites à partir des hiérarchies sémantiques du *TLFi*. Nous avons fait l'hypothèse que les mots-clés de la description textuelle qui représentent les concepts principaux de l'image sont reliés entre eux par des relations sémantiques. Ainsi, à partir des descriptions textuelles des images composées d'une liste de mots-clés, le prototype d'indexation détermine automatiquement les termes d'indexation retenus. Les évaluations manuelles des termes d'indexation montrent que les termes d'indexation sont déterminés avec une précision très élevée, mais que la complétude reste moyenne car pendant l'indexation certains termes ont été omis. L'analyse détaillée de résultats de l'intégration séparée de chaque type de relation dans le processus d'indexation a montré que les relations d'hyperonymie permettent de déterminer des termes d'indexation avec une précision très élevée, alors que les relations d'association contribuent à l'augmentation du rappel au prix d'une certaine baisse de la précision. Nous avons aussi évalué le niveau de représentation des termes d'indexation et avons pu noter que la plupart des termes représentent le niveau générique de l'image.

L'inconvénient principal de cette approche tient au fait que, pour la détermination des termes d'indexation, la description textuelle de l'image doit être suffisamment riche en mots-clés pour faire émerger des relations sémantiques entre ces mots-clés, ce qui, hélas, n'est pas souvent le cas des images issues du Web. En revanche, son intérêt est que l'utilisation de connaissances lexicographiques, ici celles extraites automatiquement à partir du *TLFi*, nous permet de déterminer non seulement des termes d'indexation, mais aussi leurs domaines d'utilisation. Nous proposons d'exploiter cet avantage ultérieurement lors de la recherche d'images.

6.3. Exploitation de l'indexation dans un algorithme simple de recherche d'images

Dans cette section, nous décrivons un algorithme simple de recherche d'images qui utilise les hiérarchies sémantiques construites à partir du *TLFi*. Le but de l'algorithme est de retrouver toutes les images pertinentes à la requête d'un utilisateur et de les structurer selon le domaine auquel les mots-clés de la requête peuvent faire référence. La figure 6.3. présente une vue globale de l'algorithme de recherche proposé. Les principales étapes de l'algorithme sont décrites ci-dessous :

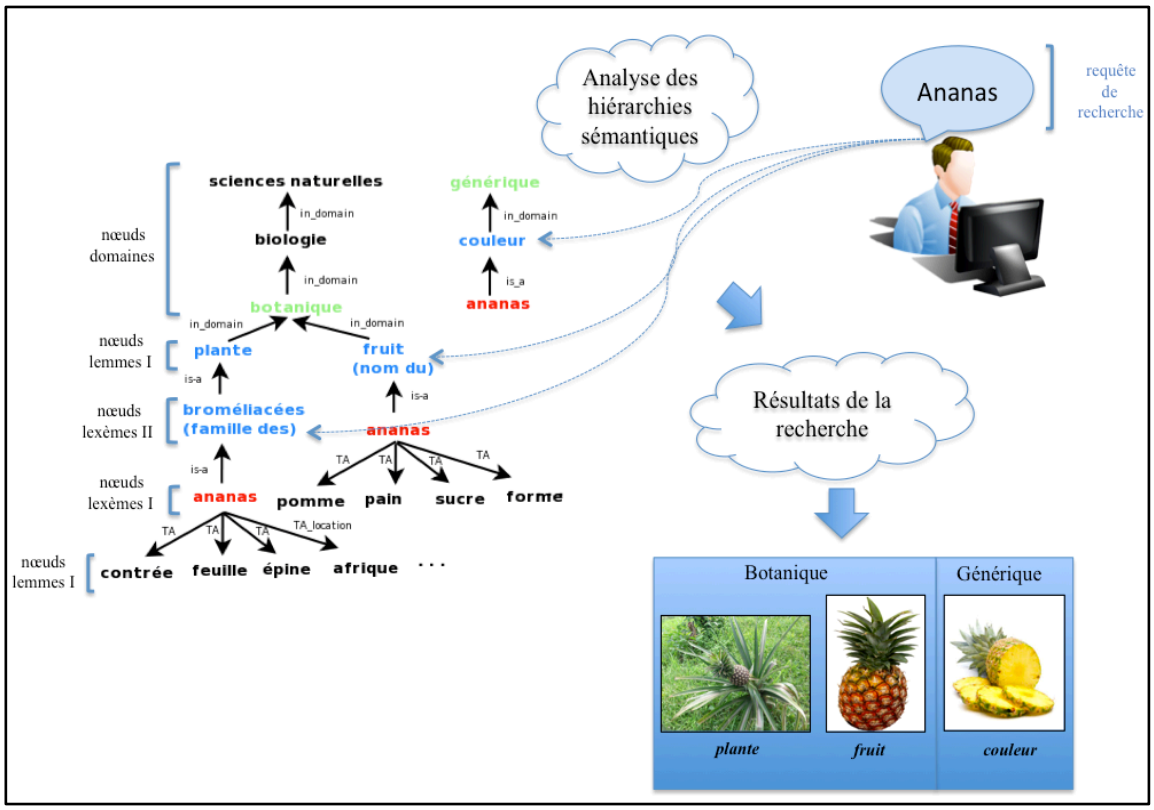


Figure 6.3. Vue globale de l'algorithme de recherche d'images

1. Lemmatisation et annotation morphosyntaxique des mots-clés de la requête d'utilisateur en utilisant le lexique morphosyntaxique Morphalou, puis filtrage par élimination des mots-clés appartenant aux catégories autres que *nom*.

2. Pour chaque mot-clé de la requête, analyse de ses arbres hiérarchiques construits préalablement en utilisant notre algorithme de construction automatique de hiérarchies sémantiques.

3. À partir des arbres hiérarchiques de chaque mot-clé, extraction des relations de filiation :

SI détermination des relations de filiation où les nœuds fils et les nœuds pères représentent les mots-clés de la requête

ALORS

POUR CHAQUE relation *is-a*

RÉPÉTER sélection des images indexées au nœud fils de la relation déterminée

FIN POUR

POUR CHAQUE relation *TA*

RÉPÉTER sélection des images indexées au nœud père de la relation déterminée

FIN POUR

SINON détermination dans les hiérarchies sémantiques de l'ensemble des nœuds avec leur type pour chaque mot-clé de la requête

POUR CHAQUE nœud de type lexème I (cf. figure 6.3. ex. *ananas*)

RÉPÉTER sélection des images indexées à ce nœud

FIN POUR

POUR CHAQUE nœud de type lemme I (cf. figure 6.3. ex. *plante, fruit, couleur*) dont le nœud père est de type domaine

RÉPÉTER sélection des images indexées aux nœuds fils de ce nœud

FIN POUR

POUR CHAQUE nœud de type lemme I (cf. figure 6.3. ex. *contrée, feuille, pomme,* etc.) dont le nœud père est de type lexème I

RÉPÉTER sélection des images indexées aux nœuds pères de ce nœud

FIN POUR

FIN SI

4. Les images trouvées sont groupées selon les domaines du nœud auquel elles ont été indexées de la manière suivante :

SI il n'y a qu'un seul domaine

ALORS le nom du domaine n'est pas affiché

SINON

SI il y a plusieurs domaines différents dont le domaine *générique*,

ALORS tout d'abord seront présentées les images du domaine *générique* sans présenter le nom du domaine, et ensuite seront affichées les autres domaines dans l'ordre dans lequel ils ont été trouvés

FIN SI

FIN SI

5. Les résultats de la recherche sont présentés à l'utilisateur sous forme d'une mosaïque d'images pour chaque domaine auquel les images trouvées peuvent appartenir.

6.3.1. Évaluation des résultats de recherche

Afin d'évaluer cet algorithme de recherche d'images, nous avons implémenté un prototype simple, en calculant ensuite le rappel et la précision des résultats de recherche en fonction du domaine. Le but était d'évaluer la qualité des résultats de recherche obtenus, en déterminant si les images trouvées par le prototype de recherche sont ou non pertinentes pour la requête dans le domaine sélectionné. L'évaluation a été effectuée par un seul évaluateur.

	Formule de calcul
Rappel	$\frac{nb\ \ images\ pertinentes\ trouvées}{nb\ images\ qui\ devraient\ \text{être}\ trouvées}$[79]
Précision	$\frac{nb\ \ images\ pertinentes\ trouvées}{nb\ total\ d'images\ trouvées}$

Tableau 6.6. Formule de calcul de la précision et du rappel de la recherche

[79] Le nombre d'images qui devraient être trouvées pour un domaine donné a été calculé manuellement par l'évaluateur.

Pour mesurer la qualité de la recherche, nous avons calculé les valeurs de la précision et du rappel selon les formules reportées dans le tableau 6.6. Ces deux valeurs ont été calculés pour tous les résultats retournés pour un domaine et pas seulement pour les *n* premiers résultats.

Ainsi, nous avons calculé le rappel et la précision pour trente-quatre concepts (ex. *lion, aigle, peur, trottinette,* etc.) appartenant à des domaines différents (ex. *architecture, flore, faune, sport,* etc.). Pour ce faire, nous avons utilisé un corpus constitué d'images indexées par notre prototype d'indexation réalisé auparavant qui prend en compte les relations d'association (cf. §6.2.).

Les valeurs obtenues après l'évaluation de ces trente-quatre concepts sur notre échantillon sont présentées dans le tableau 6.7.

Nr.	Concept	Domaine	Rappel	Précision
1.	Abeille	générique	1	1
2.	Avocat	botanique	1	1
2.1.	Avocat	générique	1	1
3.	Bébé	générique	0,4	1
4.	Bédouin	géographie	1	1
5.	Cactus	générique	0,36	1
6.1.	Canard	zoologie	0	0
6.2.	Canard	générique	0,63	1
6.3	Canard	technologie	0	0
7.	Cerise	générique	0,33	1
8.	Cygne	générique	0,66	1
9.	Dauphin	technologie	0	0
10.1.	Désert	générique	0,18	1
10.2.	Désert	géographie	0,18	1
11.1.	Église	architecture	0	0
11.2.	Église	beaux-arts	0	0
11.3.	Église	générique	0,66	0,8

12.	Élégance	générique	0,5	1
13.	Éléphant	générique	0,14	1
14.	Étang	générique	0,7	1
15.1.	Fleur	horticulture	0,06	1
15.2.	Fleur	générique	0,70	1
15.3.	Fleur	biologie	0,06	1
16.1.	Football	sports	0	0
16.2.	Football	jeux	0,38	0,71
17.1.	Forteresse	histoire	0,33	1
17.2.	Forteresse	générique	0,5	1
17.3.	Forteresse	défense	0	0
18.1.	Hélicoptère	aviation	0,03	1
18.2.	Hélicoptère	aéronautique	0,20	1
19.1.	Homme	agriculture	0	0
19.2.	Homme	histoire	0	0
19.3.	Homme	générique	0,70	0,58
20.1.	Île	générique	0,05	0,25
20.2.	Île	géologie	0,01	1
21.	Jaguar	générique	0,41	1
22.	Kiwi	botanique	1	1
23.1.	Lion	générique	1	0,4
23.2.	Lion	iconographie	0	0
23.3.	Lion	mythologie	0	0
23.4.	Lion	astrologie	1	1
23.5.	Lion	héraldique	0,25	0,16
23.6.	Lion	numismatique	1	1
24.	Lotus	générique	0,5	1
25.	Minaret	générique	0,57	1
26.	Nénuphar	générique	1	1

27.	Palmier	botanique	0,33	1
28.	Papillon	générique	0,57	1
29.	Pétanque	générique	0,66	1
30.	Tomate	art culinaire	0	0
31.	Trottinette	générique	0,11	1
32.1.	Tulipe	botanique	0,35	1
32.2.	Tulipe	générique	0,55	1
33.	Vampire	folklore	1	1
34.1.	Vitesse	générique	0,36	1
34.2.	Vitesse	sports	0	0

Tableau 6.7. Évaluation des résultats de recherche

Nous présentons ci-dessous l'analyse des résultats de cette évaluation.

6.3.2. Analyse des résultats d'évaluation

En analysant ces résultats d'évaluation, nous avons constaté que la précision de la recherche est plus élevée que le rappel, c'est-à-dire que le prototype de recherche trouve bien des images pertinentes, mais pas la totalité des celles-ci. Cela découle du fait que les images n'ont pas été indexées par tous les termes par lesquels elles devraient être indexées lors de l'indexation. Pour certains concepts, la précision et le rappel égaux à 1 signifient que le prototype a trouvé la totalité des images pertinentes à la requête ne faisant aucune erreur. Toutefois, pour certaines requêtes, la précision et le rappel sont égaux à 0 parce que les images proposées ne correspondaient pas au concept de la requête de l'utilisateur ou au domaine de celui pour lequel elles ont été trouvées.

Une analyse plus détaillée des résultats est présentée par la suite.

6.3.2.1. Domaines erronés

En analysant les résultats de la recherche, nous avons rencontré des cas où les images proposées pour un domaine ne correspondaient pas tout à fait à celui-ci. La figure 6.4. ci-dessous rend compte des images trouvées pour le concept *lion* dans le domaine

iconographie. Bien que toutes les images proposées fassent bien référence au concept *lion* elles ne correspondent pas toutes au domaine *iconographie*, qui, nous pouvons le noter, n'est pas un domaine facile à définir. C'est aussi le cas pour le concept *dauphin*, qui a été indexé au domaine *technologie* (cf. figure 6.5.). Même si les images représentent bien le concept de la requête, elles ne correspondent pas ici au domaine auquel elles ont été indexées. Cela est dû au processus de l'indexation qui a trouvé dans les hiérarchies sémantiques une relation d'association entre les mots-clés de la description textuelle d'images. Ici, la relation d'association *dauphin-tête,* dans la définition du *TLFi* du lexème DAUPHIN : « Tuyau d'écoulement d'une fontaine, représentant la tête d'un dauphin », a conduit à l'indexation de l'image par le terme *dauphin* du domaine *technologie*.

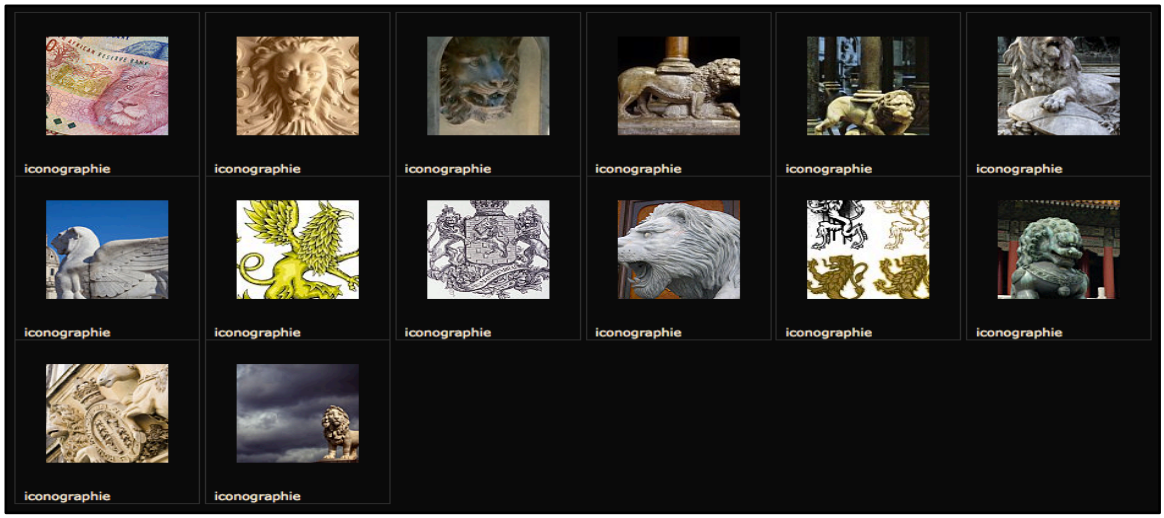


Figure 6.4. Exemple des images trouvées pour le concept *lion* dans le domaine *iconographie*

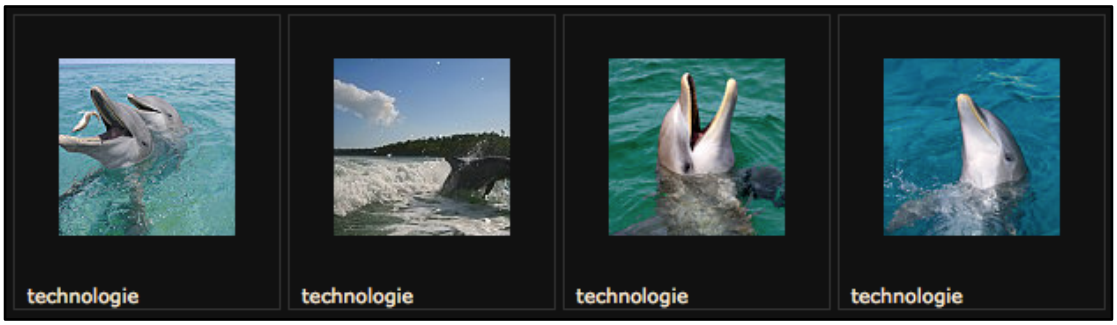


Figure 6.5. Exemple des images trouvées pour le concept *dauphin* dans le domaine *technologie*

Un autre problème concerne le domaine *générique* pour les vocables polysémiques. À titre d'exemple, pour le vocable TULIPE*,* dans le *TLFi*, il existe un lexème dont la définition pour le domaine *botanique* est « Plante bulbeuse de la famille des Liliacées, caractérisée par une haute tige droite et lisse, de longues feuilles lancéolées, engainantes, de grandes fleurs terminales, généralement solitaires, évasées en urne, formées de six divisions vivement colorées, au fond panaché de noir, de bleu, garni d'un gros pistil et d'étamines rayonnantes, et qui est très appréciée pour l'ornementation », et des lexèmes dont les définitions tiennent au domaine *générique* comme « Fleur de cette plante, utilisée pour sa valeur ornementale ». Ainsi, lors de l'indexation d'images, en fonction des relations d'hyperonymie trouvées dans les hiérarchies sémantiques entre les mots-clés de l'image comme *tulipe-fleur* ou *tulipe-plante,* certaines images ont été indexées au domaine *générique* et *botanique*. Toutefois, entre les images appartenant à ces deux domaines, il n'existe pas de différences visuelles majeures (cf. figure 6.6.). Même si le *TLFi* précise bien la différence entre *tulipe-plante* et *tulipe-fleur*, elle n'est toutefois pas perçue par les annotateurs des images, ce qui, ensuite, ne permet pas l'indexation correcte des images à ces deux domaines.

Le même cas se présente pour le concept *désert* qui peut appartenir au domaine *générique* et *géographie*, mais la différence visuelle est difficilement saisissable.

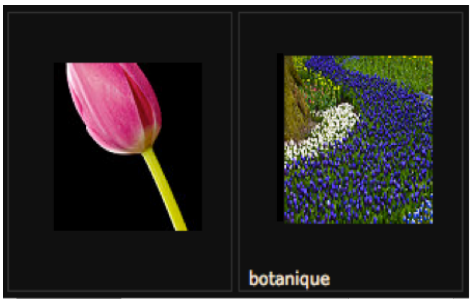


Figure 6.6. Résultats de la recherche pour le concept *tulipe*

6.3.2.2. Relations d'association

Lors de l'évaluation de l'indexation (cf. §6.2.1.), nous avions constaté que les relations d'association contribuaient à l'augmentation du rappel, mais n'amélioraient pas la précision. Il en est de même pour la recherche d'images. Ainsi, pour le concept *homme,* le prototype de recherche nous propose, entre autres, une image d'un *étang*, car

dans les hiérarchies sémantiques construites à partir du *TLFi* le terme *étang*[80] est associé au terme *homme*. Le concept *tomate* se trouve dans le même cas. En effet, le prototype de recherche trouve des images qui représentent des *spaghettis sauce tomate à l'italienne.* Cela vient du fait que, dans les hiérarchies sémantiques construites, au terme *italien*[81] du domaine *art culinaire* est associé le terme *tomate*. Les images trouvées pour le concept *tomate* ont une précision et un rappel égaux à 0, car l'évaluateur a estimé que les images trouvées ne correspondaient pas au concept de la requête, le domaine *art culinaire* n'étant pas représentatif du concept. Toutefois, dans le cas où plusieurs domaines auraient été trouvés pour ce concept, les images du domaine *art culinaire* auraient alors pu être évaluées positivement, et cela montrerait la diversité des domaines auxquels le concept donné peut appartenir.

6.3.3. Avantages de notre approche de recherche d'images

Dans les sections précédentes, nous nous sommes intéressée aux possibilités d'exploitation des hiérarchies sémantiques construites automatiquement à partir du *TLFi* dans le domaine d'indexation et de recherche d'images. Nous allons maintenant présenter ci-dessous les principaux avantages de l'approche proposée dans ce domaine d'application.

6.3.3.1. Structuration des résultats de recherche selon les domaines

La structuration des images trouvées selon les domaines du concept recherché permet de montrer à l'utilisateur les différents sens du concept et d'affiner sa recherche en fonction de ses besoins d'information. Alors que la plupart des systèmes de recherche proposent les résultats sous forme d'une simple liste ordonnée selon la pertinence des images pour la requête et que ce type d'affichage ne permet pas d'avoir une vision de la diversité d'images sans parcourir l'ensemble des images (cf. §2.2.3.).

[80] La définition du lexème ÉTANG dans le *TLFi* est « Étendue d'eau généralement stagnante, d'une faible profondeur, située dans une cuvette naturelle ou creusée par l'homme ».
[81] La définition du lexème ITALIEN dans le *TLFi* pour le domaine art culinaire est « Accommodé à l'huile d'olive, à la sauce tomate et au parmesan ».

La figure 6.7. présente un exemple des résultats de recherche pour le concept *lion*, en utilisant notre prototype de recherche. Cela permet à l'utilisateur de parcourir dans un second temps uniquement les images du domaine qui l'intéresse.

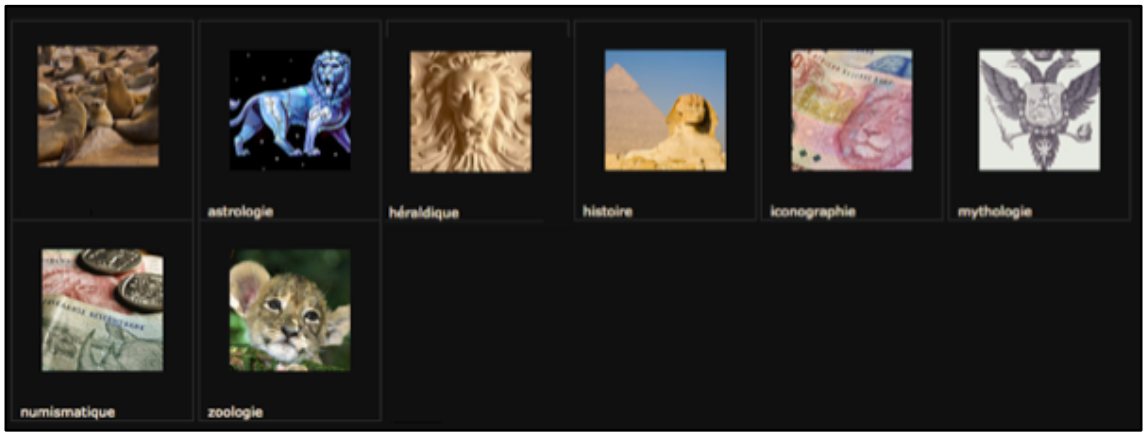


Figure 6.7. Exemple des images trouvées par notre système de recherche pour le concept *lion*

Voici un autre exemple qui présente les résultats de recherche pour le concept *avocat* (cf. figure 6.8.).

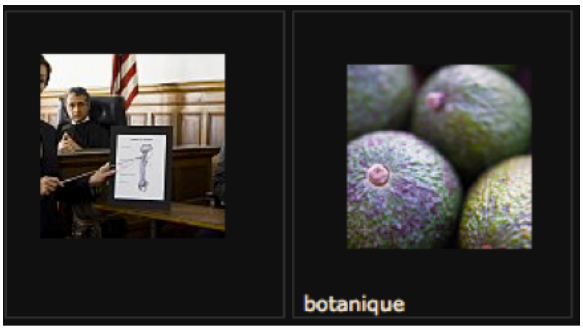


Figure 6.8. Résultats de la recherche pour le concept *avocat*

En fonction du domaine, l'utilisateur peut visionner la liste des images trouvées (cf. figures 6.9., 6.10.).

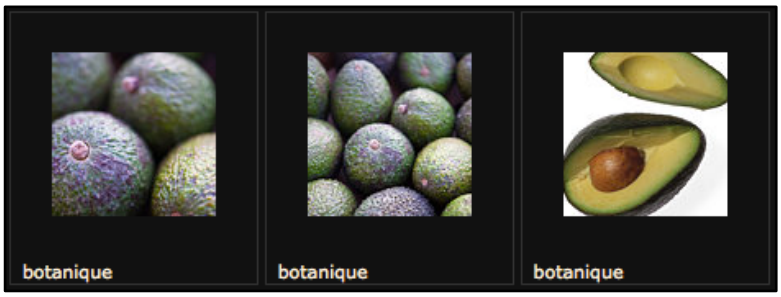


Figure 6.9. Images trouvées pour le concept *avocat* dans le domaine *botanique*

Figure 6.10. Images trouvées pour le concept *avocat* dans le domaine *générique*

6.3.3.2. Recherche associative d'images

Nous avons vu, lors de l'évaluation, que les relations d'association contribuent à la baisse de la précision de la recherche. Toutefois, les relations d'association de lieu (ex. *Afrique, Asie, Europe,* etc.) peuvent être exploitées dans la recherche d'images afin d'assurer une recherche associative. On sait que l'homme pense de manière associative, en combinant certains concepts selon sa propre expérience (ex. *hiver-neige, été-soleil*). Ainsi, une recherche associative permet de trouver, pour un concept donné, toutes les images qui sont liées par une relation d'association de lieu. Par exemple, pour le concept *Afrique* (cf. figure 6.11.), on retrouve des images d'oiseaux (*tisserin*), d'animaux (*lion*), de fleurs (*lotus*) d'Afrique (cf. figure 6.12.). Cela permet à l'utilisateur de se faire une idée générale de la flore et la faune de l'Afrique, dans le cas où il n'a pas une idée précise de ce qu'il recherche.

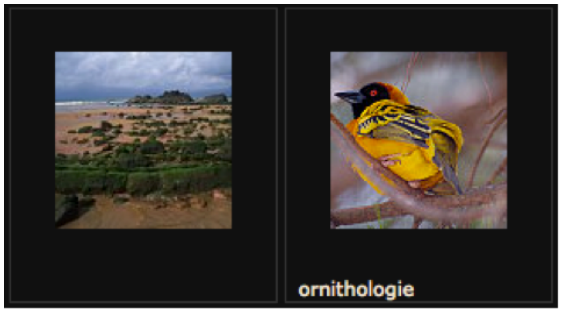


Figure 6.11. Résultats de la recherche pour le concept *Afrique*

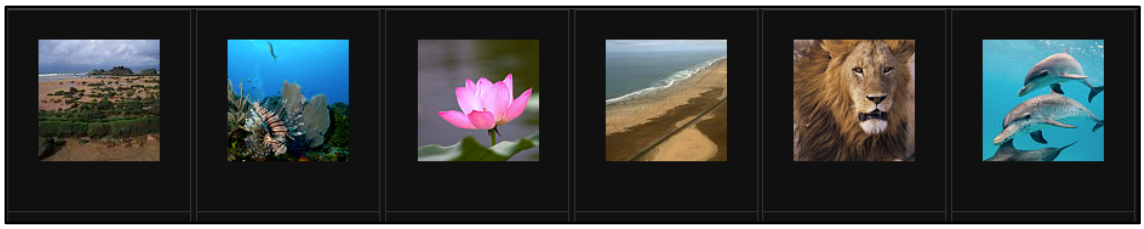


Figure 6.12. Résultats de la recherche pour le concept *Afrique* dans le domaine *générique*

6.3.3.3. Requêtes complexes

Notre prototype de recherche permet aussi une recherche complexe en combinant plusieurs concepts. Il propose les images qui ont été indexées par tous les concepts de la requête, ce qui ne le différencie pas des autres systèmes. Mais, dans le cas des requêtes dont les mots-clés sont liés par la relation d'hyperonymie comme *fleur-tulipe,* le système ne propose que les images correspondant au concept le plus spécifique *tulipe* (cf. figure 6.13.).

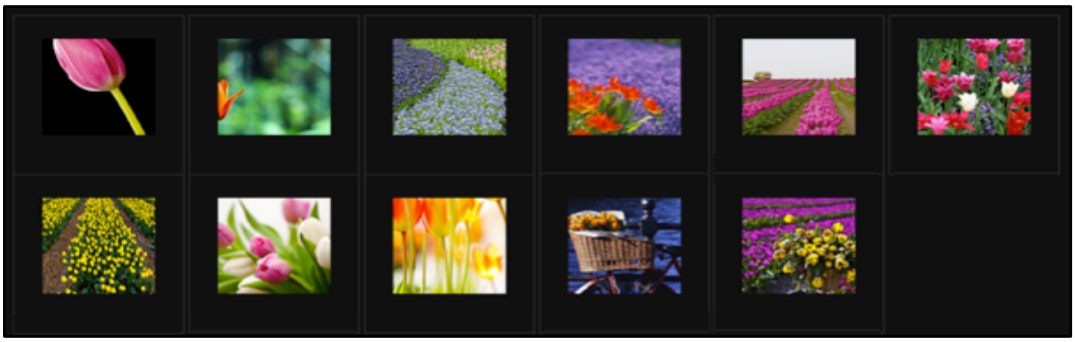


Figure 6.13. Résultats de la recherche pour la requête *fleur tulipe* dans le domaine *générique*

6.3.4. Bilan sur la recherche d'images

Dans cette section, nous avons proposé un algorithme de recherche d'images, qui, en exploitant les arbres hiérarchiques construits à partir du *TLFi* structure les réponses (images trouvées) selon les domaines auxquels le concept recherché peut appartenir. Cela offre la possibilité d'introduire un niveau d'interaction avec l'utilisateur qui permet à ce dernier de ne parcourir que les images du domaine qui l'intéresse. De plus, ce type de structuration des résultats de recherche utilisant les connaissances du *TLFi* permet aussi à l'utilisateur de se faire immédiatement une idée générale de la diversité des images, particulièrement dans le cas de concepts ambigus. Les résultats obtenus lors de l'évaluation de l'algorithme proposé à travers le prototype de recherche mis en place sont assez satisfaisants. Le prototype retrouve les images avec une précision élevée.

Un autre avantage de notre prototype de recherche est que, grâce aux relations d'association, comme le lieu, il est capable de retrouver des images qui n'ont pas été forcement indexées par le concept recherché (ex. *lion-Afrique*), même si cela peut conduire à faire baisser la précision.

Au terme de cette étude, nous pouvons donc affirmer que les domaines du *TLFi* sont applicables à la recherche d'images et apportent un nouveau moyen de structuration des résultats de recherche.

6.4. Conclusion

Dans ce chapitre, nous avons présenté une approche d'application des connaissances extraites à partir d'un dictionnaire de langue, tel le *TLFi*, au domaine d'indexation et de recherche d'images en mettant en place un prototype d'indexation et de recherche. Dans un premier temps, notre approche a consisté à projeter les descriptions textuelles des images sur des hiérarchies sémantiques obtenues à partir du *TLFi* afin d'identifier les termes d'indexation qui représentent au mieux le contenu de l'image en utilisant les relations *is-a* et les relations d'association. Dans un second temps, nous avons proposé une méthode de recherche d'images qui s'appuie sur l'indexation réalisée en structurant les images selon les domaines d'utilisation du concept recherché. Nous avons constaté que les relations de type *is-a* contribuent à l'augmentation de la précision tandis que les relations d'association augmentent le rappel. Les résultats de l'évaluation ont montré

que les connaissances issues du *TLFi* sont applicables au domaine de recherche d'images, pour l'indexation textuelle automatique d'images et pour la présentation des résultats de façon structurée selon les domaines auxquels les concepts de la requête de recherche peuvent faire référence.

Conclusion générale et perspectives

Contributions

Le travail proposé dans cette thèse porte sur l'utilisation de ressources lexicales telles qu'un dictionnaire de langue dans le domaine de l'indexation et de la recherche d'images. Notre objectif était de montrer que l'exploitation d'informations sémantiques contenues dans une telle ressource pouvait améliorer le processus d'indexation et de recherche d'images. Dans ce cadre, nous avons proposé plusieurs contributions que nous rappelons dans les sections suivantes.

Structuration automatique des connaissances du *TLFi*

Le *TLF* est un *dictionnaire de langue.* Par opposition à une visée encyclopédique, le *TLF* s'attache donc à définir chaque lexème de la langue française par ses caractéristiques linguistiques : sa forme, son sens, ses emplois stylistiques et syntaxiques. Son informatisation, qui a permis la consultation de l'ensemble de l'ouvrage sur l'Internet, ne l'a hélas pas rendu directement utilisable par les applications du traitement automatique des langues (TAL). En effet, les connaissances du *TLFi*, comme dans tout dictionnaire de langue, demeurent encodées majoritairement de manière implicite. Pour pouvoir les expliciter, nous avons dû, dans un premier temps, modéliser les données de la microstructure du *TLFi*. L'analyse de cette information nous a permis, dans un deuxième temps d'implémenter une formule de pondération des noms des définitions du *TLFi*, basée sur l'hypothèse que le nom de poids maximal représente le candidat hyperonyme d'un lexème, puis de les hiérarchiser en construisant ainsi automatiquement pour tous les lexèmes d'un vocable donné ses propres hiérarchies sémantiques. La représentation des connaissances du *TLFi* sous la forme d'arbres hiérarchiques a permis, d'une part de définir une méthode d'enrichissement du thésaurus actuel de l'entreprise construit manuellement, et d'autre part de l'utiliser lors de l'indexation automatique et la recherche d'images.

La base de données qui contient l'ensemble des hiérarchies sémantiques construites pour chaque vocable du *TLFi* sera mise à la disposition de la communauté scientifique.

Indexation automatique et recherche conceptuelle d'images

Afin d'automatiser le processus d'indexation des images dans le cadre de l'entreprise, nous avons mis en place un algorithme d'indexation automatique des images en utilisant leurs descriptions textuelles. Les hiérarchies sémantiques obtenues à partir du *TLFi* nous ont permis d'analyser les relations sémantiques entre les mots-clés et par conséquent de déterminer automatiquement les bons termes d'indexation pour une image donnée. Nous avons montré que les relations d'hyperonymie permettent de déterminer des termes d'indexation avec une précision très élevée, alors que les relations d'association contribuent à l'augmentation du rappel et en conséquence hélas, à la baisse de la précision.

Par ailleurs, certains des systèmes actuels de recherche d'images proposent les résultats de la recherche sous forme d'une simple liste d'images. Or, cette présentation n'est pas toujours pratique, surtout lorsque les utilisateurs n'ont pas une idée précise de ce qu'ils recherchent, car ils sont alors amenés à parcourir plusieurs pages avant de trouver les images désirées. Pour notre part, nous avons proposé un prototype du système de recherche d'images qui exploite les hiérarchies sémantiques construites à partir du *Trésor de la langue française informatisé* (*TLFi*). Nous avons montré que l'exploitation d'une telle ressource dans le domaine de recherche d'images améliore la précision de la recherche en structurant les résultats selon les domaines auxquels les concepts de la requête de recherche peuvent faire référence. En outre, les hiérarchies sémantiques du *TLFi* permettent de réaliser une recherche associative d'images.

Actuellement, le prototype d'indexation et de recherche d'images développé est opérationnel et sera prochainement intégré dans le système de l'entreprise.

Création du thésaurus des domaines du *TLFi*

Afin de pouvoir exploiter les domaines du *TLFi*, nous les avons normalisés manuellement et hiérarchisés en créant ainsi une base de données. Cela nous a permis de réduire considérablement le nombre des domaines de 7 786 à 758. Le besoin d'une telle ressource avait déjà été soulevé dans plusieurs travaux scientifiques, elle pourra être ainsi facilement réutilisable, par exemple pour la désambiguïsation des sens des lexèmes.

Limites et perspectives

Malgré ses apports, le travail que nous avons réalisé présente certaines limites ; néanmoins, un certain nombre de perspectives se dégagent pour l'avenir.

Incomplétude des connaissances du *TLFi*

Nous avons montré que l'exploitation d'un dictionnaire de langue, tel le *TLFi*, permettait d'améliorer l'indexation et la recherche d'images, toutefois les connaissances du *TLFi* ne sont pas suffisantes. Par rapport à d'autres dictionnaires comme *Le Petit Robert,* le *TLFi* n'a pas évolué avec le temps et ne tient pas compte de l'actualité. Ainsi, dans le *TLFi* nous ne trouvons pas de lexèmes qui sont entrés récemment dans le langage par le biais des avancées technologiques, par exemple *internet, tchat, blog, sms,* etc. Pour cela, de nouvelles ressources peuvent être utilisées. Compte tenu du fait que notre approche de construction automatique de hiérarchies sémantiques à partir du *TLFi* peut être appliqué à n'importe quel dictionnaire de langue, nous proposons d'utiliser pour cette tâche Wiktionnaire[82] qui est un dictionnaire collaboratif multilingue (libre et gratuit) disponible en ligne.

Pour ce que concerne la reconnaissance des noms propres, le dictionnaire relationnel multilingue de noms propres Prolexbase[83] (Tran & Maurel, 2006) pourrait être utilisé. L'avantage de cette ressource est qu'elle permet de déterminer non seulement les noms propres des villes, mais aussi ceux des célébrités ou des monuments historiques. De plus, dans Prolexbaxe les noms propres sont mis en relation avec les adjectifs et noms relationnels correspondants (ex. *Paris-parisien*).

Par ailleurs, la normalisation des locutions du *TLFi*, en cours d'achèvement dans le cadre d'un projet interne de l'ATILF, permettra de les considérer aussi comme des termes d'indexation et ainsi de résoudre les problèmes rencontrés lors des évaluations de nos méthodes.

[82] http://fr.wiktionary.org
[83] http://www.cnrtl.fr/lexiques/prolex/

Recherche d'images par la négation

Dans nos recherches, nous avons uniquement utilisé les relations d'hyperonymie et d'association. Toutefois, dans les hiérarchies sémantiques construites, il existe aussi une relation de type *is_not_TA* qui indique que le terme associé n'est pas caractéristique pour un lexème (ex. *arme-judo*) et une relation de type *not-is-a* qui indique que le terme n'est pas un hyperonyme d'un lexème donné. L'exploitation de ces relations devrait permettre de réduire le champ de recherche et d'améliorer les résultats, particulièrement dans le cas d'une recherche itérative qui, comme s'est le cas dans le moteur de recherche de Xilopix, permet à l'utilisateur d'invalider certaines images proposées par le système pour affiner sa recherche.

Aspects multilingues

En interne, le prototype travaille sur une seule langue, ici le français, par contre les descriptions textuelles des images et les requêtes des recherches peuvent être formulées en plusieurs langues. Dans ce cas, il conviendra seulement d'utiliser les systèmes de traduction afin de traduire les mots-clés. L'exploitation de nouvelles ressources linguistiques multilingues n'est donc pas nécessaire pour généraliser nos propositions à d'autres langues.

Vers une représentation graphique du *TLFi*

Actuellement, sur le Web, il existe plusieurs thésaurus[84] et réseaux sémantiques visuels[85] qui donnent une image globale d'un vocable du dictionnaire, explicitée par les relations sémantiques entre les lexèmes. L'application Prox proposée par Gaume (2004) fait partie de cette catégorie : elle permet une modélisation géométrique du sens sous forme de graphes de type « petit monde » qui représentent des réseaux sémantiques souvent trop complexes. Ainsi, la méthode de construction automatique de hiérarchies sémantiques pour les vocables du *TLFi* que nous avons proposée pourrait être appliquée pour permettre une représentation graphique des connaissances du *TLFi*. Ce fait facilitera le parcours dans les réseaux sémantiques et permettra ainsi de mieux structurer la présentation du sens d'un lexème, de sélectionner sa définition en fonction du

[84] http://www.visualthesaurus.com
[85] http://www.lexipedia.com

domaine d'usage, et de structurer la visualisation des sens des lexèmes. Ce type de représentation des connaissances d'un dictionnaire pourrait être très utile lors de l'apprentissage d'une langue par exemple.

Liste des figures

Figure 1.1. Exemple d'organisation générale de macro- et microstructure des vocables AVOCAT1 et AVOCAT2 dans le *TLFi* 25
Figure 1.2. Taux des vocables du *TLFi* appartenant aux différentes catégories grammaticales 26
Figure 1.3. Microstructure du *TLFi* 27
Figure 1.4. Exemple complexe d'une entrée cachée sous la définition de l'élément formant -PATHIE dans le *TLFi* 29
Figure 1.5. Modèle de données du *TLFi* 30
Figure 1.6. Exemple de résultats de recherche dans le *TLFi* avec une recherche assistée du type X: code_grammatical(substantif) ; 32
Figure 1.7. Exemple d'une liste de fleurs trouvées dans le *TLFi* avec une recherche complexe du type X: code_grammatical(substantif); 33
Figure 1.8. Exemple d'une liste de meubles trouvés dans le *TLFi* avec une recherche complexe du type X: code_grammatical(substantif); 34
Figure 1.9. Extrait de l'article du vocable AVOCAT dans SEMEME 42
Figure 1.10. Schéma d'organisation de l'information dans SEMEME 43
Figure 1.11. Exemple des définitions structurées du *TLFi* dans le projet Definiens 46
Figure 1.12. Exemple d'une structure ambigüe du *TLFi* (à gauche) et structure de WOLF (à droite) (Eckard, Barque, Nasr, & Sagot, 2012) 47
Figure 2.1. Schéma d'un système classique de recherche d'information 55
Figure 2.2. Image de *lion* dans Flickr 63
Figure 2.3. Résultats de la recherche pour la requête *avocat* dans Google Images (octobre 2012) 66
Figure 2.4. Résultats de la recherche pour la requête *lion* dans Google Images 66
Figure 2.5. Résultats de la recherche pour la requête *lion* dans Getty Images 67
Figure 2.6. Exemple des images trouvées dans Google Images pour le concept *lion* (octobre 2012) 68
Figure 2.7. Exemple d'utilisation du système QBIC 70
Figure 2.8. Exemple de courbe rappel-précision 75
Figure 3.1. Exemple des différents sens du LION dans le WordNet anglais 82
Figure 3.2. Exemple de deux clusters déterminés pour *grizzly* (Popescu, Grefenstette, & Moëllic, 2007) 84
Figure 3.3. Exemple des résultats obtenus pour *dog (chien)* dans Olive (a) et Google Images (b) (Popescu & Grefenstette, 2008) 85
Figure 3.4. Un exemple d'un extrait du réseau sémantique du ConceptNet (Liu & Singh, 2004) 87
Figure 3.5. Exemple d'un extrait de thésaurus Garnier sur le sujet *thèmes* 91

Figure 3.6. Représentation des relations et des termes dans un thésaurus de genres cinématographiques 93
Figure 3.7. Modèle de thésaurus hiérarchique à arborescence précise..... 94
Figure 3.8. Modèle de thésaurus hiérarchique à arborescence simple 94
Figure 3.9. Exemple de schéma fléché du domaine *Traitement* du thésaurus *Vieillissement* (Baddas & Labarde, 2003)..... 95
Figure 3.10. Exemple de terminogramme pour le domaine *Traitement* du thésaurus *Vieillissement* (Baddas & Labarde, 2003)..... 96
Figure 3.11. Exemple de types d'ontologies (Uschold & Gruninger, 2004)..... 98
Figure 3.12. Différents types d'ontologies selon leur niveau d'abstraction..... 99
Figure 3.13. Exemple de catégorisation d'un concept générique dans différentes ontologies (Chandrasekaran, Josephson, & Benjamins, 1999)..... 100
Figure 3.14. Exemple de l'interface de l'utilisateur pour l'annotation de l'image (Kong, Hwang, Na, & Kim, 2005)..... 101
Figure 3.15. Exemple de structure hyperonymique..... 107
Figure 3.16. Hiérarchies extraites à partir des dictionnaires CED et LDOCE 110
Figure 3.17. Exemple de polyhiérarchie..... 111
Figure 3.18. Hiérarchie problématique..... 111
Figure 3.19. Taxonomie avec des boucles..... 112
Figure 4.1. Modélisation réalisée à partir des données issues du projet SEMEME 118
Figure 4.2. Répartition des lemmes dans les définitions du *TLFi* 120
Figure 4.3. Taux des lemmes de différentes catégories grammaticales dans les définitions du *TLFi* 121
Figure 4.4. Répartition des lemmes des définitions du *TLFi* entre les différentes catégories grammaticales 121
Figure 4.5. Précision pour des noms de poids maximal obtenue avec la formule de pondération TF-DF proposée et TD-IDF 138
Figure 4.6. Précision pour des noms de poids maximal obtenue avec la formule de pondération TF-DF-ITPOS et TD-IDF-ITPOS..... 139
Figure 4.7. Précision pour des noms de poids maximal obtenue avec la formule de pondération ITPOS 140
Figure 5.1. Vue globale de l'algorithme de construction de hiérarchies sémantiques 150
Figure 5.2. Exemple du processus 1 de l'algorithme 152
Figure 5.3. Exemple de détermination des nœuds pères pour les nœuds lexèmes du vocable ANANAS..... 153
Figure 5.4. Exemple de détermination des nœuds fils pour des nœuds pères des nœuds lexèmes du vocable ANANAS 155
Figure 5.5. Exemple de transformation des nœuds fils en nœuds pères des nœuds lexèmes du vocable ANANAS 157

Figure 5.6. Exemple de fusion du thésaurus des domaines aux hiérarchies sémantiques construites automatiquement 158
Figure 5.7. Exemple des hiérarchies sémantiques construites pour le vocable ANANAS 159
Figure 5.8. Exemple des hiérarchies sémantiques construites pour le vocable RAQUETTE 161
Figure 5.9. Résultats de l'évaluation manuelle des relations hyperonymiques 162
Figure 5.10. Résultats de l'évaluation manuelle des relations hyperonymiques du domaine *flore* et *faune* 167
Figure 5.11. Coïncidence des structures hyperonymiques construites manuellement et celles construites automatiquement 171
Figure 6.1. Vue globale de l'algorithme d'indexation automatique d'images 178
Figure 6.2. Évaluation des niveaux de représentation des termes d'indexation ... 185
Figure 6.3. Vue globale de l'algorithme de recherche d'images 193
Figure 6.4. Exemple des images trouvées pour le concept *lion* dans le domaine *iconographie* 199
Figure 6.5. Exemple des images trouvées pour le concept *dauphin* dans le domaine *technologie* 199
Figure 6.6. Résultats de la recherche pour le concept *tulipe* 200
Figure 6.7. Exemple des images trouvées par notre système de recherche pour le concept *lion* 202
Figure 6.8. Résultats de la recherche pour le concept *avocat* 202
Figure 6.9. Images trouvées pour le concept *avocat* dans le domaine *botanique* 203
Figure 6.10. Images trouvées pour le concept *avocat* dans le domaine *générique* 203
Figure 6.11. Résultats de la recherche pour le concept *Afrique* 204
Figure 6.12. Résultats de la recherche pour le concept *Afrique* dans le domaine *générique* 204
Figure 6.13. Résultats de la recherche pour la requête *fleur tulipe* dans le domaine *générique* 204

Liste des tableaux

Tableau 2.1. Formules de calcul des mesures de l'exactitude de l'indexation..... 72
Tableau 2.2. Formules de calcul des mesures de la consistance de l'indexation ... 73
Tableau 4.1. Information sur des données du corpus de travail 119
Tableau 4.2. Expressions métalinguistiques de la classe d'opposition ou de négation 125
Tableau 4.3. Expressions métalinguistiques de la classe générique..... 126
Tableau 4.4. Information sur le corpus de travail..... 136
Tableau 4.5. Information sur le corpus de référence 137
Tableau 5.1. Exemples de relations identifiées comme hyperonymiques..... 163
Tableau 5.2. Exemples des relations identifiées comme non hyperonymiques ... 164
Tableau 5.3. Exemples des relations identifiées comme hyperonymiques dans le domaine *flore* 168
Tableau 5.4. Exemples des relations identifiées comme hyperonymiques dans le domaine *faune* 169
Tableau 5.5. Exemples des relations identifiées comme non hyperonymiques dans le domaine *faune* 170
Tableau 5.6. Exemple des relations hyperonymiques existantes dans les deux hiérarchies 171
Tableau 6.1. Résultats du processus d'indexation intégrant des relations is-a..... 181
Tableau 6.2. Exemples d'images avec les termes d'indexation par lesquels elles ont été indexées 184
Tableau 6.3. Exemples des images indexées par les termes d'indexation incorrects 187
Tableau 6.4. Résultats de l'évaluation du processus d'indexation intégrant les relations d'association..... 189
Tableau 6.5. Exemples des termes d'indexation déterminés par l'intégration de relations is-a (termes I) et la prise en compte de relations d'association (termes II)..... 191
Tableau 6.6. Formule de calcul de la précision et du rappel de la recherche..... 195
Tableau 6.7. Évaluation des résultats de recherche 198

Index des notions

A

antonymie, 82
article, 23, 24, 26, 28, 29, 30, 31, 37, 41, 42, 56, 71, 75, 112, 113, 119, 122, 124
attributs, 97

B

bouclage de pertinence, 54, 68
bruit, 64, 74, 90

C

caractéristiques spécifiques, 106, 128, 135
circularité, 15, 36, 112, 113
classes, 97, 229
complétude de l'indexation, 72, 181, 188, 189, 192
concept, 14, 36, 68, 82, 83, 84, 88, 89, 92, 96, 99, 100, 180, 198, 199, 200, 201, 202, 203, 204, 205
consistance, 72, 73
courbe rappel-précision, 75

D

définition approximative, 38
définition dérivative, 36
définition hyperonymique, 35, 36, 38
définition méréonymique, 37
définition synonymique, 36
descripteur, 88, 89, 90, 93
dictionnaire, 15, 16, 21, 23, 24, 26, 28, 30, 31, 40, 48, 49, 53, 78, 86, 106, 108, 109, 111, 112, 113, 114, 133, 138, 139, 145, 160, 205, 207, 209, 210
domaine de définition, 14, 15, 16, 17, 18, 22, 23, 27, 32, 33, 40, 41, 60, 62, 65, 68, 71, 77, 78, 79, 81, 83, 88, 90, 91, 92, 94, 95, 96, 97, 98, 99, 100, 102, 105, 107, 108, 113, 115, 117, 119, 122, 123, 125, 127, 128, 130, 131, 133, 134, 138, 139, 140, 145, 146, 147, 149, 150, 151, 152, 153, 154, 156, 157, 158, 160, 161, 166, 167, 168, 169, 170, 171, 172, 175, 177, 178, 180, 182, 185, 192, 193, 194, 195, 196, 198, 199, 200, 201, 202, 203, 204, 205, 207, 208, 211

E

entrée, 22, 24, 28, 29, 32, 48, 84, 112, 122, 123, 136, 140, 213
exactitude de l'indexation, 39, 71, 72, 128

F

facette, 61, 92, 97
F-mesure, 74

G

genre prochain, 35, 36, 44, 47, 104, 106, 114, 134, 136, 164

H

hiérarchie sémantique, 145, 155
hyperonyme absent, 110
hyperonyme direct, 156
hyperonyme indirect, 156
hyperonymes conjoints, 111
hyperonymie, 34, 38, 39, 47, 48, 65, 78, 82, 83, 104, 108, 114, 133, 142, 160, 161, 163, 166, 185, 188, 192, 200, 204, 208, 210

I

indexation, 13, 14, 15, 16, 17, 18, 21, 49, 51, 53, 54, 55, 56, 58, 59, 60, 61, 62, 63, 64, 65, 68, 71, 72, 73, 76, 77, 78, 79, 81, 83, 88, 89, 90, 91, 100, 102, 108, 113, 114, 118, 131, 133, 141, 145, 173, 175, 177, 178, 179, 180, 181, 182, 184, 185, 186, 187,

188, 189, 190, 191, 192, 193, 196, 198, 199, 200, 201, 205, 207, 208, 209
indexation automatique d'images, 63, 175, 177, 188, 206
indexation collaborative, 62
indexation d'images par le contenu, 64, 68
indexation manuelle d'images, 61

J

justesse de l'indexation, 72, 181, 189

L

lemmatisation, 24, 31, 56, 63, 117, 119, 165
lexème, 25, 32, 34, 35, 36, 37, 38, 39, 41, 44, 46, 55, 56, 82, 83, 87, 88, 106, 107, 113, 118, 119, 120, 121, 122, 124, 125, 127, 128, 129, 130, 132, 135, 136, 139, 141, 145, 146, 147, 148, 149, 150, 151, 152, 153, 155, 156, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 186, 194, 199, 200, 201, 207
lexie, 24, 28, 30, 41, 44, 120, 122
locution, 25, 41, 55, 88, 115, 119, 120, 122, 124, 129, 165, 186

M

macrostructure, 17, 21, 24, 25, 26, 30, 48
méronymie, 82, 83, 110
microstructure, 17, 19, 21, 24, 25, 26, 30, 41, 43, 48, 49, 150, 207, 213
modèle booléen, 59, 60
modèle probabiliste, 60
modèle vectoriel, 59, 60
mot, 22, 25, 29, 31, 49, 56, 60, 62, 77, 88, 127, 150
mot vide, 56
mot-clé, 13, 14, 21, 25, 49, 53, 54, 61, 62, 63, 64, 65, 68, 71, 77, 78, 105, 178, 179, 180, 181, 182, 185, 186, 188, 189, 190, 192, 193, 194, 199, 200, 204, 208, 210
mot-forme, 25, 31, 55, 56, 58, 64, 107, 113, 119, 132, 147
mot-vedette, 24, 26, 30

N

nœud, 47, 94, 148, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 167, 179, 188, 194
nœud domaine, 151, 155, 157, 158
nœud lemme, 151, 153, 154, 155, 156, 157, 158
nœud lexème, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 214
non-descripteur, 89
normalisation, 18, 56, 58, 117, 119, 123, 124, 130, 131, 133, 209

O

ontologie, 19, 47, 86, 88, 96, 97, 99, 100, 101, 105, 108
ontologie d'application, 99
ontologie de domaine, 98
ontologie de haut niveau, 98
ontologie de tâche, 98
ontologie légère, 97
ontologie lourde, 98
ontologie d'application, 99
ontologie de domaine, 98
ontologie de haut niveau, 98
ontologie de tâche, 98
ontologie légère, 97

P

polyhiérarchie, 110
pondération, 16, 17, 55, 56, 57, 58, 75, 77, 115, 117, 125, 126, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 145, 153, 164, 173, 177, 207
pondération globale, 57, 133, 134, 139
pondération locale, 56, 57, 58, 133, 137
pondération par position, 133, 135, 138, 139, 140, 141
précision, 16, 48, 57, 73, 74, 75, 83, 84, 85, 87, 97, 100, 101, 103, 105, 106, 107, 108, 137, 138, 139, 140, 141, 167, 172, 181, 188, 189, 192, 195, 196, 198, 200, 203, 205, 208

pureté de l'indexation, 72

R

rappel, 57, 62, 73, 74, 75, 83, 181, 188, 192, 195, 196, 198, 200, 205, 208
recherche, 13, 14, 15, 16, 17, 18, 21, 23, 24, 28, 31, 32, 33, 34, 40, 45, 46, 47, 48, 49, 51, 53, 54, 55, 58, 59, 60, 62, 64, 65, 66, 67, 68, 69, 70, 71, 73, 74, 75, 76, 77, 78, 79, 81, 83, 84, 85, 86, 87, 88, 89, 90, 91, 93, 98, 100, 102, 107, 108, 113, 114, 118, 133, 141, 145, 158, 165, 173, 175, 177, 180, 181, 182, 192, 193, 195, 196, 198, 200, 201, 202, 203, 204, 205, 207, 208, 209, 210
recherche d'images par le contenu, 51, 68, 69, 70, 71, 77
recherche textuelle d'images, 65
règles d'inclusion, 146
règles d'association, 11, 143, 148
règles d'inclusion, 11, 143, 146
règles de hiérarchisation, 11, 143, 148
relation d'association, 92, 120, 159, 199, 203
relation d'équivalence, 92
relation hiérarchique, 92
relation in_domain, 158, 160
relation is_not_TA, 159
relation is-a, 159, 179, 188, 194
relation TA, 159, 179, 188, 194
relation TA_location, 159

S

schémas fléchés, 94
silence, 74, 90
structure arborescente, 93
synonymie, 14, 39, 60, 65, 78, 82, 90, 92
syntagme figé, 124
syntagme hyperonymique, 35, 39, 163

T

terme, 23, 25, 29, 39, 56, 57, 58, 59, 60, 62, 64, 72, 73, 84, 87, 88, 89, 92, 93, 104, 109, 111, 117, 135, 159, 172, 180, 181, 182, 185, 186, 199, 201, 205
terme associé, 92
terme équivalent, 92
terme générique, 92
terme spécifique, 92
termes associés, 13, 92, 179
termes équivalents, 92, 95
termes génériques, 92, 94, 111
termes spécifiques, 92, 93, 94
terminogrammes, 95
thésaurus, 14, 16, 18, 21, 54, 67, 68, 70, 77, 78, 79, 88, 89, 90, 91, 92, 93, 94, 95, 96, 101, 102, 105, 108, 114, 118, 119, 130, 131, 133, 141, 143, 145, 157, 158, 159, 160, 161, 166, 170, 171, 172, 173, 178, 207, 208, 210
thésaurus hiérarchique à arborescence précise, 94
thésaurus hiérarchique à arborescence simple, 94
token, 55
tokenisation, 55, 63
troponymie, 82

V

vocable, 18, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 37, 41, 42, 113, 119, 122, 123, 124, 125, 127, 128, 138, 140, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 164, 165, 166, 167, 170, 171, 172, 177, 200, 207, 210
vocable homonyme, 24, 123
vocable monosémique, 25
vocable polysémique, 25, 200

Références bibliographiques

Agirre, E., Olatz, A., Arregi, X., Artola, X., Snchez, A., Lersundi, M., et al. (2000). Extraction of semantic relations from a basque monolingual dictionary using constraint grammar. *Proceedings of Euralex.*

Aitchison, J., & Gilchrist, A. (1992). *Construire un thésaurus.* (A.D.B.S, Éd.) Paris.

Alshawi, H. (1987). Processing Dictionary Definitions with Phrasal Pattern Hierarchie. *Computational linguistics*, 13 (3/4), 195-202.

Amsler, R. (1980). The structure of the merriam-webster pocket dictionary. Ph. d. dissertation, University of Texas at Austin, Austin, Texas.

Aspura, Y. I., Khalid, M., Noah, S. A., & Abdullah, S. N. (2011). Towards a multimodality ontology image retrieval. *Proceedings of the Second international conference on Visual informatics: sustaining research and innovations*. 382-393.

Assadi, H., & Bourigault, D. (2000). Analyses syntaxique et statistique pour la construction d'ontologies à partir de textes. Dans J. Charlet, M. Zackland , G. Kassel, & D. Bourigault, *Ingénierie des connaissances - Evolutions récentes et nouveaux défis.* Eyrolles.

ATILF (ouvrage collectif publié sous le nom du laboratoire) (2004). *Trésor de la langue française informatisée.* CNRS Editions, Livre d'accompagnement 591 p. et CD du texte intégral, Version PC, ISBN 2-271-06273-X, novembre 2004, Version Mac OX X, ISBN 2-271- 06365-5, septembre 2005.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., & Ives , Z. (2007). Dbpedia: A nucleus for a web of open data. *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, 722-735. Busan, Korea.

Auger, A., & Barrière, C. (2008). Pattern-based approaches to semantic relation extraction : A state of- the-art. *Special Issue on Pattern-based Approaches to Semantic Relation Extraction, Terminology, 14* (1), 1-19.

Baccini, A., Dejean, S., Kompaore, D., & Mothe, J. (2010). Analyse des critères d'évaluation des systèmes de recherche d'information. *Technique et science informatiques. 29* (3), 289-308.

Bachimont, B. (2000). Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en Ingénierie des connaissances. *Ingénierie des Connaissances: Evolutions récentes et nouveaux défis,1*, 1-16.

Barque, L., & Polguère, A. (2012). *Guide des annotateurs pour le balisage des définitions du TLFi Projet Definiens - version 2*. ATILF CNRS, Nancy.

Barque, L., Polguère, A., & Nasr, A. (2010). From the Definitions of the Trésor de la Langue Française to a Semantic Database of the French Language. *Proceedings of the 14th EURALEX International Congress.* Leewarden.

Barque, L., & Polguère, A. (2009). Structuration et balisage sémantique des définitions du Trésor de la Langue Française informatisé (TLFi). *Fourth International Conference on Meaning-Text Theory.* Montréal.

Baddas, N., & Labarde, G. (2003). *Mini-thésaurus : Vieillissement humain aspect physiologique.* Université de Versailles Saint Quentin en Yvelines.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007). Open information extraction from the web. *JCAI*, 2670-2676.

Baneyx, A., Malaisé, V., Charlet, J., Zweigenbaum, P., & Bachimont, B. (2005). Synergie entre analyse distributionnelle et patrons lexico-syntaxiques pour la construction d'ontologies différentielles. *Conférence TIA-2005.* Rouen.

Baryla, C. (1985). Thésaurus iconographique : Système descriptif des représentations. *Bulletin des Bibliothèques de France, 2*.

Bernet, C., & Pierrel, J.-M. (2005). Histoire de Frantext : constitution d'une base textuelle (1964-2002) et perspectives. Dans J. C. Arnould, & É. Champion (Éd.), *L'édition électronique en littérature et dictionnairique : évaluation et bilan.* Presses Universitaires de Rouen.

Boulogne, A. (2004). *Vocabulaire de la documentation.* Paris: ADBS.

Boujemaa, N., Fauqueur , J., Ferecatu , M., Fleuret , F., Gouet , V., Lesaux, B., et al. (2001). Ikona: Interactive specific and generic image retrieval. *International workshop on Multimedia ContentBased Indexing and Retrieval (MMCBIR'2001).*

Bourigault, D., Aussenac-Gilles, N., & Charlet, J. (2004). Construction de ressources termi- nologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle (RIA) – Techniques Informatiques et structuration de terminologiques , 18* (1), 87-110.

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

Bourigault, D., & Fabre, C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaires, 25*, 131-151.

Bourigault, D. (1994). Extraction et structuration automatiques de terminologie pour l'aide à l'acquisition de connaissances à partir de textes. *Actes du 9e congrès Reconnaissance des Formes et Intelligence Artificielle - AFCET.* Paris.

Byrd, R., Calzolari, N., Chodorow, M., Klavans, J., Neff, M., & Rizk, O. (1987). Tools and methods for computational linguistics. *Computational Linguistics*, 13 (3/4), 219-240.

Calzolari, N. (1984). Detecting patterns in a lexical data base. *10th International Conference on Computational Linguistics (COLING-1984)*, 170-173.

Cardaci, M., Gesu, V. D., Petrou, M., & Tabacchi, M. E. (2006). On the Evaluation of Images Complexity : A Fuzzy Approach. *Fuzzy logic and applications : 6th international work- shop, WILF 2005*, *3849*, 305-311. Crema, Italy.

Castillo, M., Real, F., Asterias, J., & Rigau, G. (2004). The TALP systems for disambiguating WordNet glosses. *Proceedings of ACL 2004 SENSEVAL-3 Workshop*, 93-96. Barcelona, Spain.

Cederberg, S., & Widdows, D. (2003). Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, *4*, 111-118.

Chai, J. Y., Zhang, C., & Jin, R. (2007). An Empirical Investigation of User Term Feedback in Text-based Targeted Image Search. *ACM Transactions on Information Systems (TOIS), 25* (1).

Chandrasekaran, B., Josephson, J. R., & Benjamins, R. V. (1999). What Are Ontologies, and Why Do We Need Them? *Journal IEEE Intelligent Systems, 14* (1), 20-26.

Chen, H. -L., & Rasmussen, E. M. (1999). Intellectual access to images. *Library Trends, 48* (2), 291-302.

Chodorow, M., Byrd, R., & Heidorn, G. (1985). Extracting semantic hierarchies from a large on-line dictionary. *23rd Annual Meeting of the Association for Computational Linguistics (ACL-1985)*, 299-304.

Choi-Jonin, I., & Delhay, C. (2008). *Introduction à la méthodologie en linguistique.* Presses Universitaires de Strasbourg.

Clark , M., Kim , Y., Kruschwitz , U., Song , D., Albakour , D., Dignum , S., et al. (2012). Automatically structuring domain knowledge from text : a review of current research. *Information Processing and Management : an International Journal , 48* (3), 552–568.

Clausi, D., & Jernigan, M. (2000). Designing gabor filters for optimal texture separability. *Pattern Recognition, 33*, 1835-1849.

Claveau, V., Moreau, F., & Sébillot, P. (2007). Description des textes. Dans P. Gros, *L'indexation multimédia,* 163-190.

Coelho, T. S., Calado, P. P., Souza, L. V., Ribeiro-Neto, B., & Muntz, R. (2004). Image retrieval using multiple evidence ranking. *IEEE Transactions on Knowledge and Data Engineering*, 16 (4), 408-417.

Copestake, A. (1990). An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary. *Proceedings of the First*

International Workshop on Inheritance in Natural Language Processing, 19-29. Tilburg.

Dendien, J., & Pierrel, J. -M. (2003). Le Trésor de la Langue Française Informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence. *TAL (Traitement Automatique des Langues), 44* (2), 11-37.

Doerr, M. (2003). The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine, 24* (3).

Dolan, W., Vanderwende, L., & Richardson, S. D. (2000). Polysemy in a Broad-Coverage Natural Language Processing System. Dans Y. Ravin , & C. Leacock (Éds.), *Polysemy: Theoretical and Computational Approaches*, 178-204. Oxford University Press.

Eckard, E., Barque, L., Nasr, A., & Sagot, B. (2012). Dictionary-Ontology Cross-Enrichment Using TLFi and WOLF to enrich one another. *COLING Workshop on Cognitive Aspects of the Lexicon.*

Enser, P. G., Sandom, C. J., Lewis, P. H., & Hare, J. S. (2006). The reality of the semantic gap in image retrieval. *Proceedings of the 1st International Conference on Semantic and Digital Media Technologiess.*

Enser, P., & McGregor, C. (1993). *Analysis of visual information retrieval queries.* British Library Research and Development Report, 6104.

Etzioni, O., Cafarella, M. J., Downey, D., Kok, S., Popescu, A. -M., Shaked, T., et al. (2004). Web-Scale information extraction in KnowitAll. *Proceedings of the 13th International World Wide Web Conference*, New York, USA, 100-110.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Felber, H. L. (1987). *Manuel de terminologie.* Paris: Unesco.

Ferecatu, M., Boujemaa, N., & Crucianu, M. (2008). Semantic interactive image retrieval combining visual and conceptual content description. *ACM multimedia systems Journal, 13* (5-6), 309-322.

Ferecatu, M., Crucianu, M., & Boujemaa, N. (2005). *Active SVM-based Relevance Feedback with Hybrid Visual and representation.* INRIA - Rocquencourt, IMEDIA.

Flickner, M., Sawhney , H. S., Ashley , J., Huang , Q., Dom , B., Gorkani , M., et al. (1995). Query by image and video content : The QBIC system. *IEEE Computer, 28* (9), 22-32.

Furnas, G. W., Landauer , T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACMr, 30* (11), 964-971.

Gaudin, F., & Guespin, L. (2000). *Initiation à la lexicologie française - de la néologie aux dictionnaires*. Bruxelles: Duculot.

Gaume, B. (2004). Balades Aléatoires dans les Petits Mondes Lexicaux. I3: Information Interaction Intelligence.

Gevers, T., & Smeulders, A. (2000). Pictoseek : combining shape and color invariant features for image retrieval. *IEEE Transactions on Image Processing, 9* (1), 102-119.

Gruber, T. R. (1995). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal Human-Computer Studies, 43* (5-6), 907-928.

Gruber, T. R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition, 5* (2), 199-220.

Guarino, N. (1998). Formal Ontology and Information Systems. *Formal Ontology in Information Systems: Proceedings of the 1st International Conference*, 3-15. Trento, Italy.

Guha, R. V., & Lenat, D. B. (1990). Cyc : a mid-term report. *Journal AI Magazine, 11* (3), 32-59.

Hathout, N. (1996). Pour la construction d'une base de connaissances lexicologiques à partir du Trésor de la Langue Française. Les marqueurs superficiels dans les définitions spécialisées. Cahiers de Lexicologie, 68 (1), 137-173.

Harris, Z. (1990). La genèse de l'analyse des transformations et de la métalangue. *Langage , 25* (99), 9-20.

Hayes, P., & Warren, M. (2010). A Lightweight Ontology for Describing Images. *AAAI Spring Symposium: Linked Data Meets Artificial Intelligence'10.*

Hearst, M. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING)*, 539-545. Nantes, France.

Hollink, L., Schreiber, G., & Wielinga, B. (2007). Patterns of semantic relations to improve image content search. *Journal Web Semantics: Science, Services and Agents on the World Wide Web, 5* (3), 195-203.

Hsu, M. -H., Tsai, M. -F., & Chen, H. -H. (2008). Combining WordNet and ConceptNet for Automatic Query Expansion: A Learning Approach. *Proceedings of the 4th Asia information retrieval conference on Information retrieval technology*, 213-224.

Hsu, M. -H., & Chen, H. -H. (2006). Information Retrieval with Commonsense Knowledge. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 651-652. Seattle, WA, USA.

Hyvönen, E., Saarela, S., & Viljanen, K. (2004). Application of ontology techniques to view- based semantic search and browsing. *Proceedings of The Semantic Web: Research and Applications, First European Semantic Web Symposium.*

Hyvönen, E., Styrman, A., & Saarela, S. (2002). Ontology-Based Image Retrieval. *Proceedings of the XML Finland 2002 Conference*, 15-27.

Iakovidis, D. K., Schober, D., Boeker, M., & Schulz, S. (2009). An Ontology of Image Representations for Medical Image Mining. *Information Technology and Applications in Biomedicine.*

Ide, N., & Véronis, J. (1995). Knowledge Extraction from Machine-readable Dictionaries: An Evaluation. (P. Steffens, Éd.) Machine Translation and the Lexicon, 898, 19-34.

Ide, N., & Véronis, J. (1994). Refining Taxonomies Extracted from Machine Readable Dictionaries. (S. Hockey, & N. Ide, Éds.) Research in Humanities Computing 2, 145-170.

Ide, N., & Véronis, J. (1993). Extracting knowledge-bases from machine-readable dictionaries: Have we wasted our time? *Proceedings of KB&KS'93 Workshop*, 257-266. Tokyo.

Imbs, P. (1971). *Trésor de la Langue Française* (Vol. t.1). Préface.

Imbs, P., & Quemada, B. (1971-1994). *Le Trésor de la Langue Française.*

Jensen, K., & Binot, J. -L. (1987). Disambiguating prepositional phrase attachments by using on-line dictionary definitions. *Computational Linguistics, 13* (3-4), 251-260.

Kefi, L., Berrut, C., & Gaussier, E. (2005). Indexation Complexe de documents: vers une vérification qualitative. *Inforsid*, 521-538. Grenoble.

Kekre, H. B., Mukherjee, P., Kakaiya, M., Wadhwa, S., Singh, S., & Thepade, S. (2010). Image Retrieval with Shape Features Extracted using Morphological Operators with BTC. *International Journal of Computer Applications, 6* (8), 28-33.

Khalid, Y. I., & Noah, S. A. (2011). A framework for integrating DBpedia in a multi-modality ontology news image retrieval system. *International Conference on Semantic Technology and Information Retrieval (STAIR)*, 144-149.

Kong, H., Hwang, M., & Kim, P. (2006). The study on the semantic image retrieval based on the personalized ontology. *International Journal of Information Technology , 12* (2).

Kong, H., Hwang, M., Na, K., & Kim, P. (2005). The Study on the Semantic Image Retrieval Using the Cognitive Spatial Relationships in the Semantic Web. *IFIP — The International Federation for Information Processing , 188*, 101-111.

Lenat, D. B. (1995). CYC: A Large-Scale Investment in Knowledge Infrastructure. *Magazine Communications of the ACM, 38* (11), 33-38.

Litkowski, K. C. (2005). Computational Lexicons and Dictionaries. (O. Elsevier Publishers, Éd.) *Encyclopedia of Language and Linguistics (2nd ed.)*.

Liu, H., & Singh, P. (2004). Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal , 22* (4), 211-226.

Loiseau, S., Gréa, F., & Magué, J.-P. (2011). Dictionnaires, théorie des graphes et structures lexicales. *Revue de Sémantique et de Pragmatique, 27*, 51-78.

Lux-Pogodalla, V., & Polguère, A. (2011). Construction of a French Lexical Network: Methodological Issues. *First InternationalWorkshop on Lexical Resources, WoLeR 2011*. Ljubljana, Slovénie.

Lv, Y., & Zhai, C. (2009). Adaptive relevance feedback in information retrieval. *Proceedings of the 18th ACM conference on Information and knowledge management*, 255-264.

Ma, W. -Y., & Manjunath, B. (1999). NeTra : A Toolbox for Navigating Large Image Databases. *Mutimedia Systems*, 7 (3), 184-198.

Macgregor, G., & McCulloch, E. (2006). Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review*, 55 (5), 291-300.

Manser, M. (2012). État de l'art sur l'acquisition de relations sémantiques entre termes : contextualisation des relations de synonymie. *RECITAL'2012, 14e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues* . Grenoble, France.

Marchesotti, L., Perronnin, F., Larlus, D., Csurka, G., & Michallon, L. (2012). Évaluation automatique de la qualité esthétique des photographies à l'aide de descripteurs d'images génériques. *RFIA 2012 (Reconnaissance des Formes et Intelligence Artificielle)*. Lyon.

Maron, M. E., & Kuhns, J. L. (1960). On relevance, probabilistic indexing, and information retrieval. *Journal of the Association for Computing Machinery, 7* (3), 216-244.

Martin, R. (2001). *Sémantique et automate.* Paris: PUF.

Mel'cuk, I., Clas, A., & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire.* Paris/Louvain-la-Neuve: Duculot.

Mezaris, V., Kompatsiaris, I., & Strintzis, M. (2003). An ontology approach to object-based image retrieval. *Proceedings of the IEEE International Conference on Image Processing, ICIP03*, *2*, 511-514.

Michiels, A., & Noël, J. (1982). Approaches to thesaurus construction. *Proceedings of the 9th conference on Computational linguistics (COLING '82*), 1, 227-232.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM, 38* (11), 39-41.

Montemagni, S. (1992). Tailoring a broad-coverage grammar for the analysis of dictionary definitions. *EURALEX '92: Papers submitted to the 5th International Euralex Congress on Lexicography.*

Montemagni, S., & Vanderwende, L. (1992). Structural Patterns vs. String Palterns for Extracting Semantic Information from Dictionaries. *Proceedings of COLING'92*, 546-552.

Murala, S., Gonde, A. B., & Maheshwari, R. P. (2009). Color and Texture Features for Image Indexing and Retrieval. *IEEE International Advance Computing Conference (IACC 2009) Patiala*, 1411-1416. India.

Nakamura, J., & Nagao, M. (1988). Extraction of semantic information from an ordinary english dictionary and its evaluation. *2th International Conference on Computational Linguistics (COLING-1988)*, 459-464.

Nasr, A., Béchet, F., Rey, J. -F., Favre, B., & Le Roux, J. (2011). Macaon: An nlp tool suite for processing word lattices. *The 49th Annual Meeting of the Association for Computational Linguistics.*

Navigli, R., & Velardi, P. (2010). Learning Word-Class Lattices for Definition and Hypernym Extraction. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, 1318-1327. Uppsala, Sweden.

Navigli, R. (2009). Using cycles and quasi-cycles to disambiguate dictionary glosses. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL'09*, 594-602. Athens, Greece.

Navigli, R., & Velardi, P. (2008). From Glossaries to Ontologies: Extracting Semantic Structure from Textual Definitions. *Proceedings of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, 71-87.

Panofsky, E. (1955). *Meaning in the Visual Arts : papers in and on art history.* New York, Garden City: Doubleday.

Pantel, P., & Ravichandran, D. (2004). Automatically Labeling Semantic Classes. *Proceedings of Conference of HLT / North American Chapter of the Association for Computational Linguistics,* 321-328.

Pasca, M. (2008). Turning Web Text and Search Queries into Factual Knowledge: Hierarchical Class Attribute Extraction. *AAAI*.

Perreira Da Silva, M., & Courboulay, V. (2011). Une nouvelle mesure de complexité pour les images basée sur l'attention visuelle. *XXIIIème coloque GRETSI.*

Pierrel, J.-M. (2011). Recherche et valorisation en lexicographie française : les ressources informatisées du laboratoire ATILF. Dans *Recherches, didactiques,*

politiques linguistiques : perspectives pour l'enseignement du français en Italie, M.-C. Jullio, D. Londei et P. Puccini Eds., Francoangeli, Collana Il punto, 2011, 165-180.

Polguère, A. (2009). Lexical systems : graph models of natural language lexicons. *Language Resources and Evaluation, 43* (1), 41-55.

Polguère, A. (2008). *Lexicologie et sémantique lexicale : notions fondamentales.* Les Presses de l'université de Montréal.

Ponzetto, S. P., & Strube, M. (2007). Deriving a Large Scale Taxonomy from Wikipedia. *Proceeding AAAI'07 Proceedings of the 22nd national conference on Artificial intelligence* , *2*, 1440-1445.

Ponzetto, S. M., & Navigli, R. (2009). Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia. *roceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, 2083-2088. Pasadena, CA.

Ponzetto, S. P., & Strube, M. (2011). Taxonomy induction based on a collaboratively built knowledge repository. *Artificial Intelligence , 175*, 1737-1756.

Popescu, A., & Grefenstette, G. (2008). A Conceptual Approach to Web Image Retrieval. *LREC 2008.* Marrakech, Morocco.

Popescu, A., Grefenstette, G., & Moëllic, P. A. (2007). Improving Image Retrieval Using Semantic Resources. *Springer Studies in Computational Intelligence, 93*.

Pruvost, J. (2002). *Les dictionnaires de langue française.* (P.u.f. France, Éd.) Paris.

Qi, X., & Han, Y. (2007). Incorporating multiple SVMs for automatic image annotation. *Pattern Recognition , 40* (2), 728-741.

Raghavan, V. V., & Wong, S. K. (1986). A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science*, 37 (5), 279-287.

Ravin, Y. (1990). Disambiguating and interpreting verb definitions. *Proceedings of the 28rd Annual Meeting of the ACL*, 260-267.

Ritter, A., Soderland, S., & Etzioni, O. (2009). What is this, anyway: Automatic hypernym discovery. *Proceedings of the 2009 AAAI Spring Symposium on Learning by Reading and Learning to Read*, 88-93.

Robertson, S. E., & Spärck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, 27* (3), 129-146.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. (Prentice-Hall, Éd.) *The SMART Retrieval System: Experiments in Automatic Document Processing*, 313-323.

Sagot, B., & Fišer, D. (2012). Automatic Extension of WOLF. *GWC2012 - 6th International Global Wordnet Conference.* Matsue, Japon.

Salton, G. (1989). *Automatic Text Processing* (Addison-Wesley Ed.).

Salton, G., & Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management, 24* (5), 513-523.

Salton, G., & McGill, M. (1983). *Introduction to Modern Information Retrieval.* New York: McGraw-Hill.

Salton, G., Yang, C., & Yu, C. (1975). A Theory of Term Importance in Automatic Text Analysis. *Journal of the American Society for Information Science, 26* (1), 33-44.

Salton, G. (1969). A comparison between manual and automatic indexing methods. *Journal of American Documentation, 20* (1), 61-71.

Shatford, S. (1986). Analyzing the subject of a picture : a theoretical approach. *Cataloging and Classification Quarterly, 6* (3), 39-62.

Singh, P., & Barry, B. (2003). Collecting commonsense experiences. *Proceedings of the 2nd international conference on Knowledge capture*, 154-161.

Snow, R., Jurafsky, D., & Ng, A. Y. (2005). Learning syntactic patterns for automatic hypernym discovery. *NIPS 2005.*

Soergel, D. (1994). Indexing and Retrieval performance : the logical evidence. *Journal of the American Society for Information Science, 45* (8), 589-599.

Solli, M., & Lenz, R. (2010). Color semantics for image indexing. *Conference on Colour in Graphics, Imaging, and Vision*, 353-358.

Sparck Jones, K. (1986). *Synonymy and semantic classification.* (E. U. Press, Éd.) Edinburgh, England.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2008). YAGO: A Large Ontology from Wikipedia and WordNet. *Web Semantics: Science, Services and Agents on the World Wide Web, 6* (3), 203-217.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago: a core of semantic knowledge. *Proceedings of the 16th international conference on World Wide Web*, 697-706.

Tesnière, L. (1959). *Éléments de syntaxe structurale.* (Klinksieck, Éd.).

Touratier, C. (2000). *La sémantique.* Paris: Armand Colin.

Tran, M., & Maurel, D. (2006). Prolexbase — Un dictionnaire relationnel multilingue de noms propres. *TAL*, 47 (3), 115-139.

Velardi, P., Faralli, S., & Navigli, R. (2013). OntoLearn Reloaded: A Graph-Based Algorithm for Taxonomy Induction. *Computational Linguistics, 39* (3), 665-707.

Uschold, M., & Gruninger, M. (2004). Ontologies and Semantics for Seamless Connectivity. *ACM SIGMOD Record, 33* (4), 58-64.

Voorhees, E. M. (1993). Using WordNet to Disambiguate Word Senses for Text Retrieval. *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, 171-180. NY, USA.

Wang, C., Zhang, L., & Zhang, H. -J. (2008). Learning to Reduce the Semantic Gap in Web Image Retrieval and Annotation. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 355-362. Singapore.

Wang, X. J., Ma, W. Y., & Li, X. (2004). Data-driven approach for bridging the cognitive gap in image retrieval. *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo*, *3*, 2231-2234. Taiwan.

Weber, N., & Buitelaar, P. (2006). Web-based ontology learning with ISOLDE. *Proceedings of ISWC2006 Workshop on Web Content Mining with Human Language Technologies.*

Wong, R. F., & Leung, C. C. (2008). Automatic semantic annotation of real-world web images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30* (11), 1933-1944.

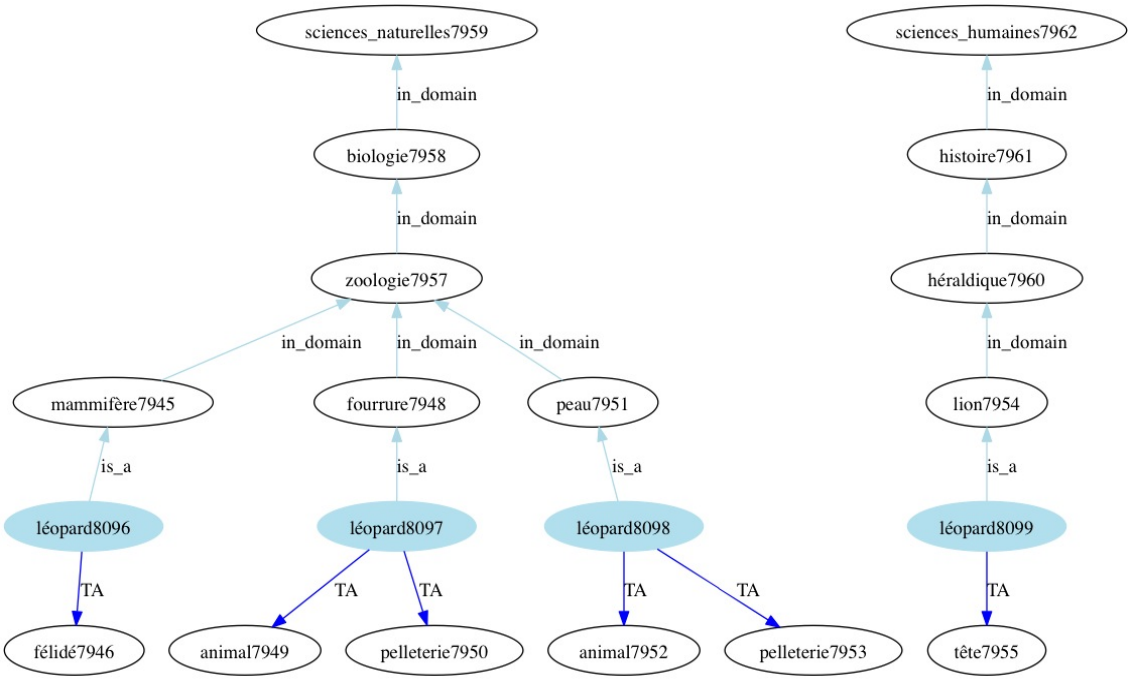
Wu, W., Li, H., Wang, H., & Zhu, K. (2011). *Towards a Probabilistic Taxonomy of Many Concepts.* Microsoft Technical Report.

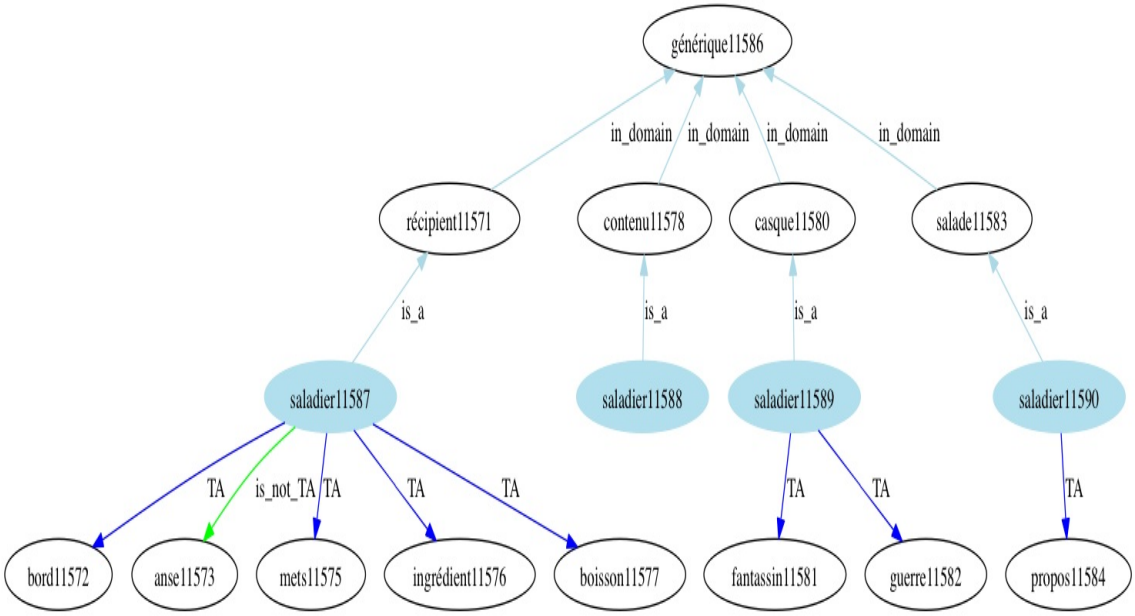
Xu, Y., Jones , G. J., & Wang , B. (2009). Query dependent pseudo-relevance feedback based on wikipedia. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 59-66.

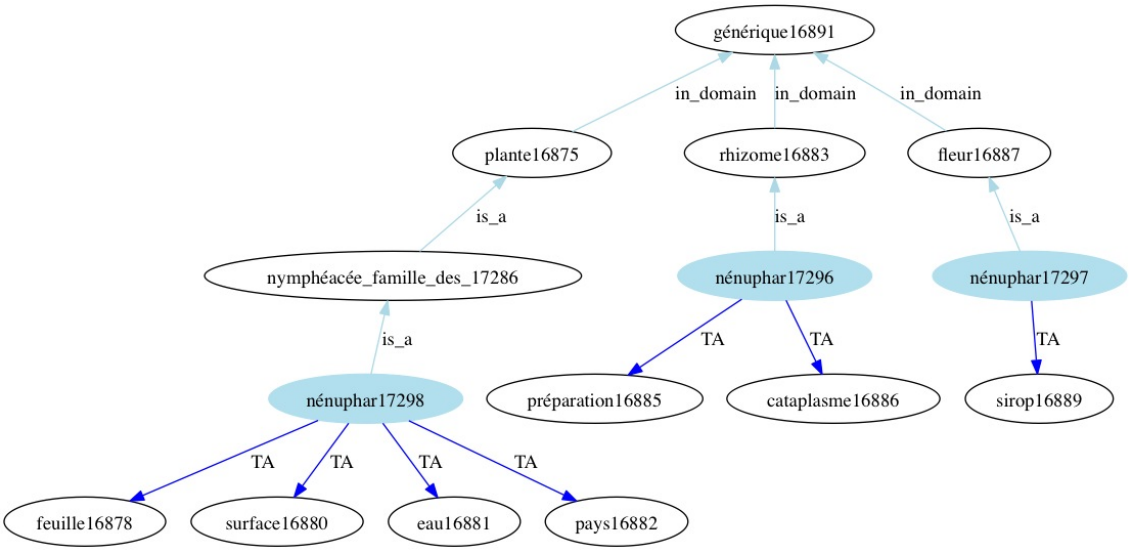
Yang, H., & Callan, J. (2009). A metric-based framework for automatic taxonomy induction. *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL)*, 271-279. Suntec.

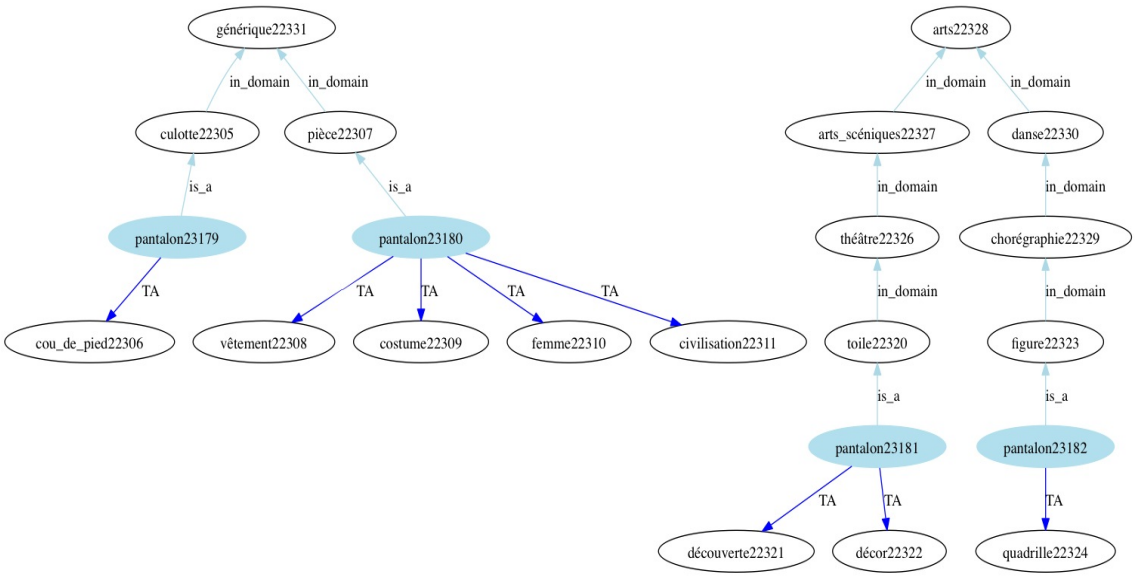
Zhang, C., Chai, J. Y., & Jin, R. (2005). User term feedback in interactive text-based image retrieval. *Proceedings SIGIR'05*, 51-58.

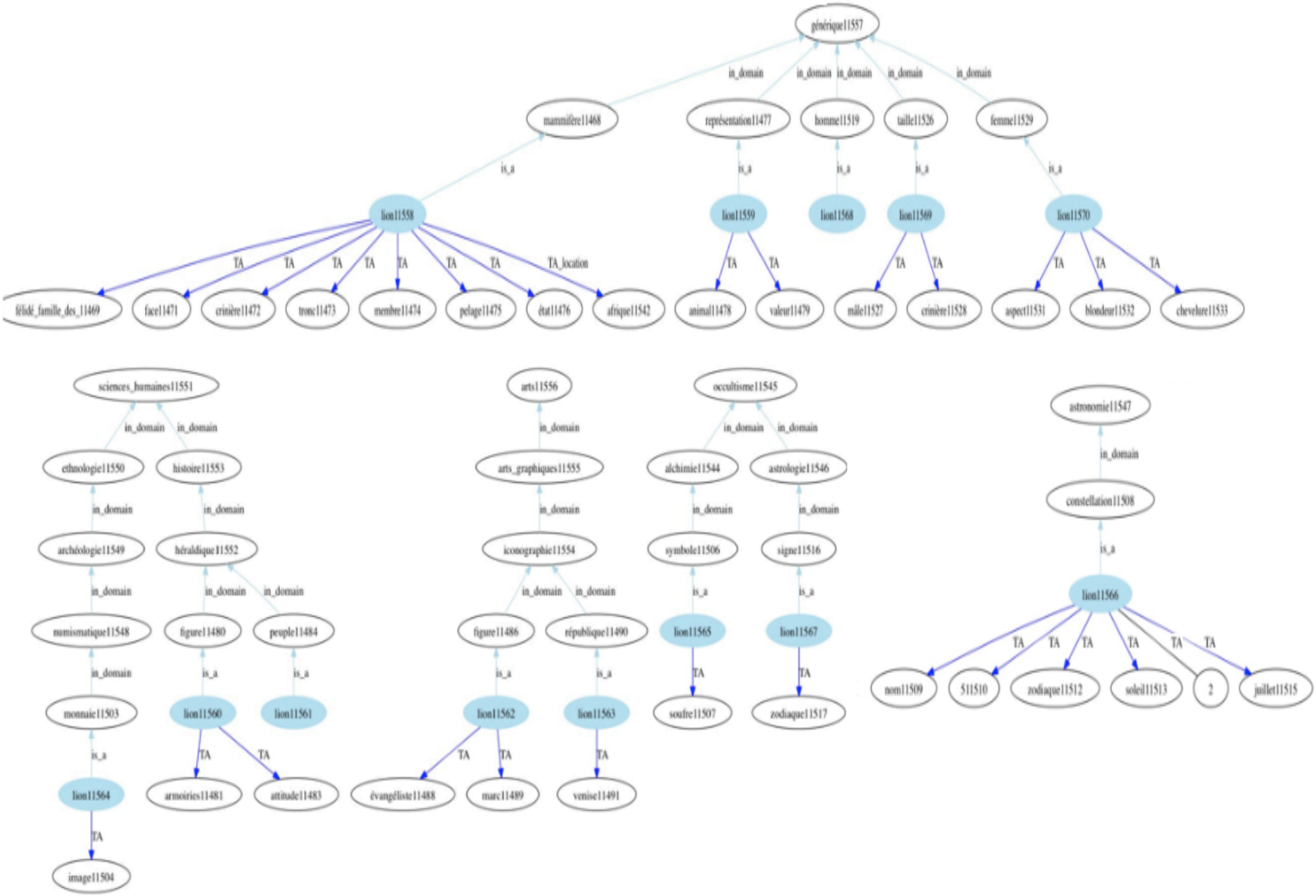
Annexe 1. Exemples de hiérarchies sémantiques

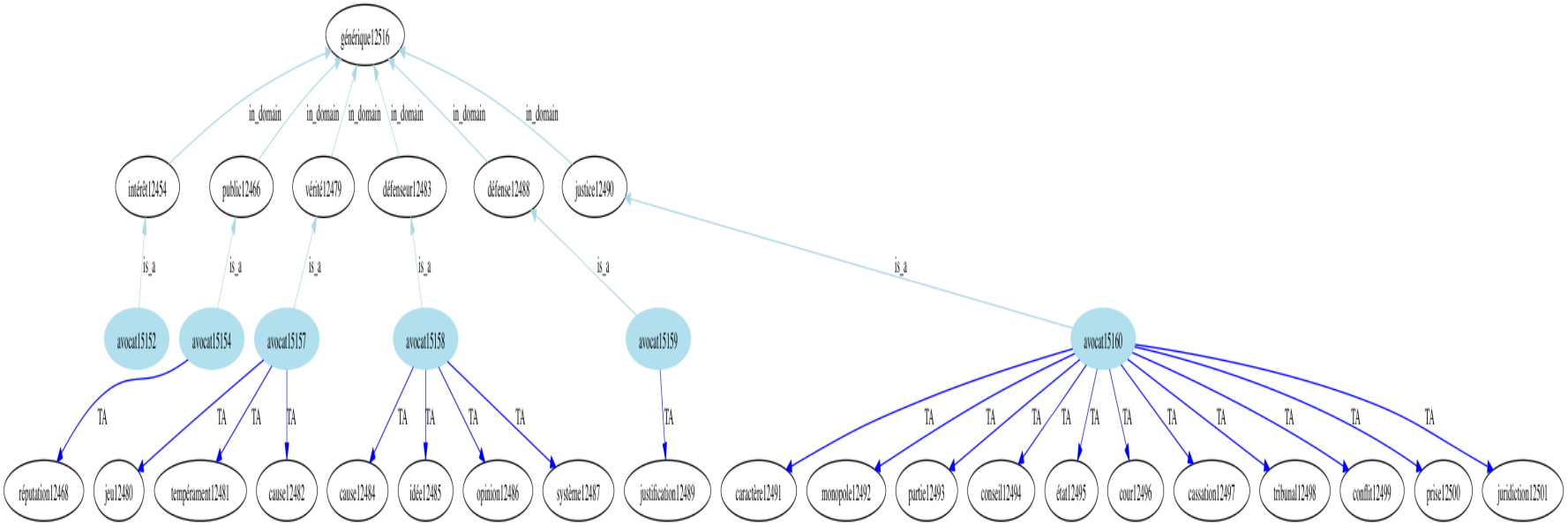


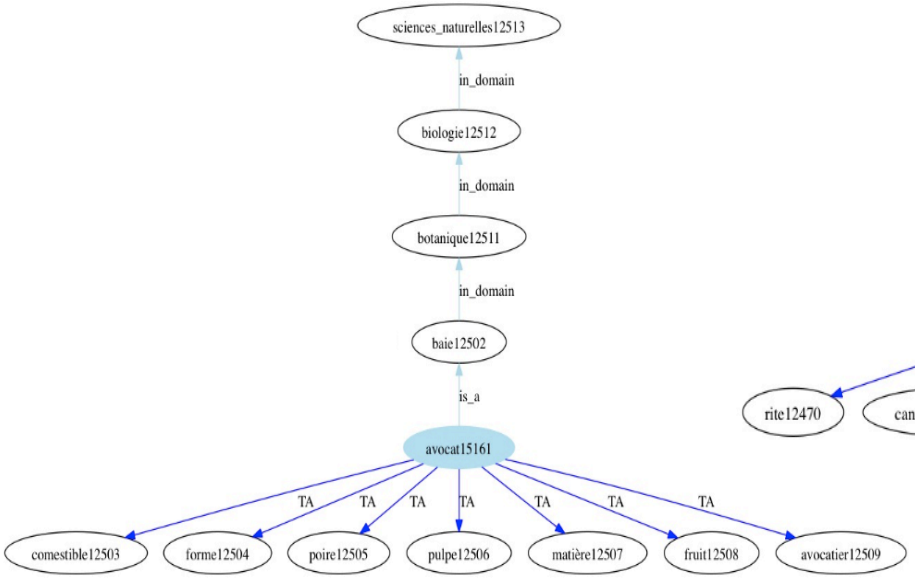


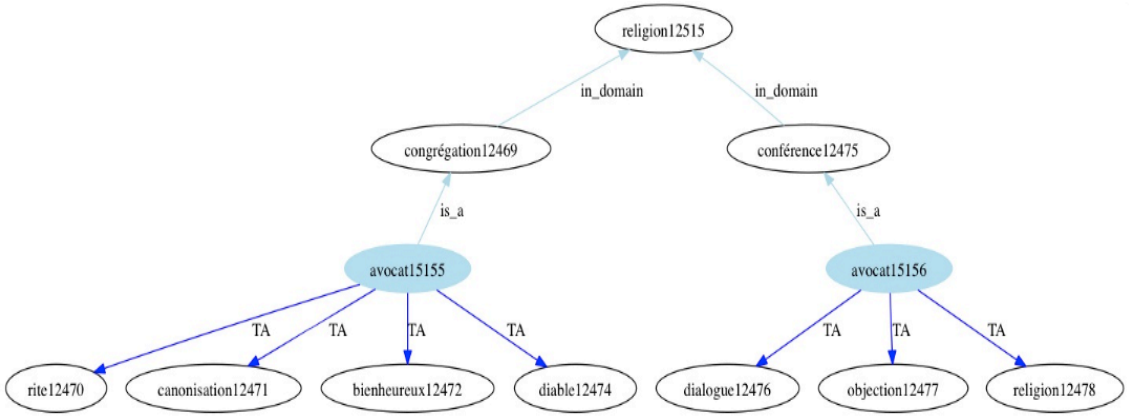

générique16891
in_domain
in_domain
in_domain
plante16875
rhizome16883
fleur16887
is_a
is_a
is_a
nymphéacée_famille_des_17286
nénuphar17296
nénuphar17297
is_a
TA
TA
TA
nénuphar17298
préparation16885
cataplasme16886
sirop16889
TA
TA
TA
TA
feuille16878
surface16880
eau16881
pays16882

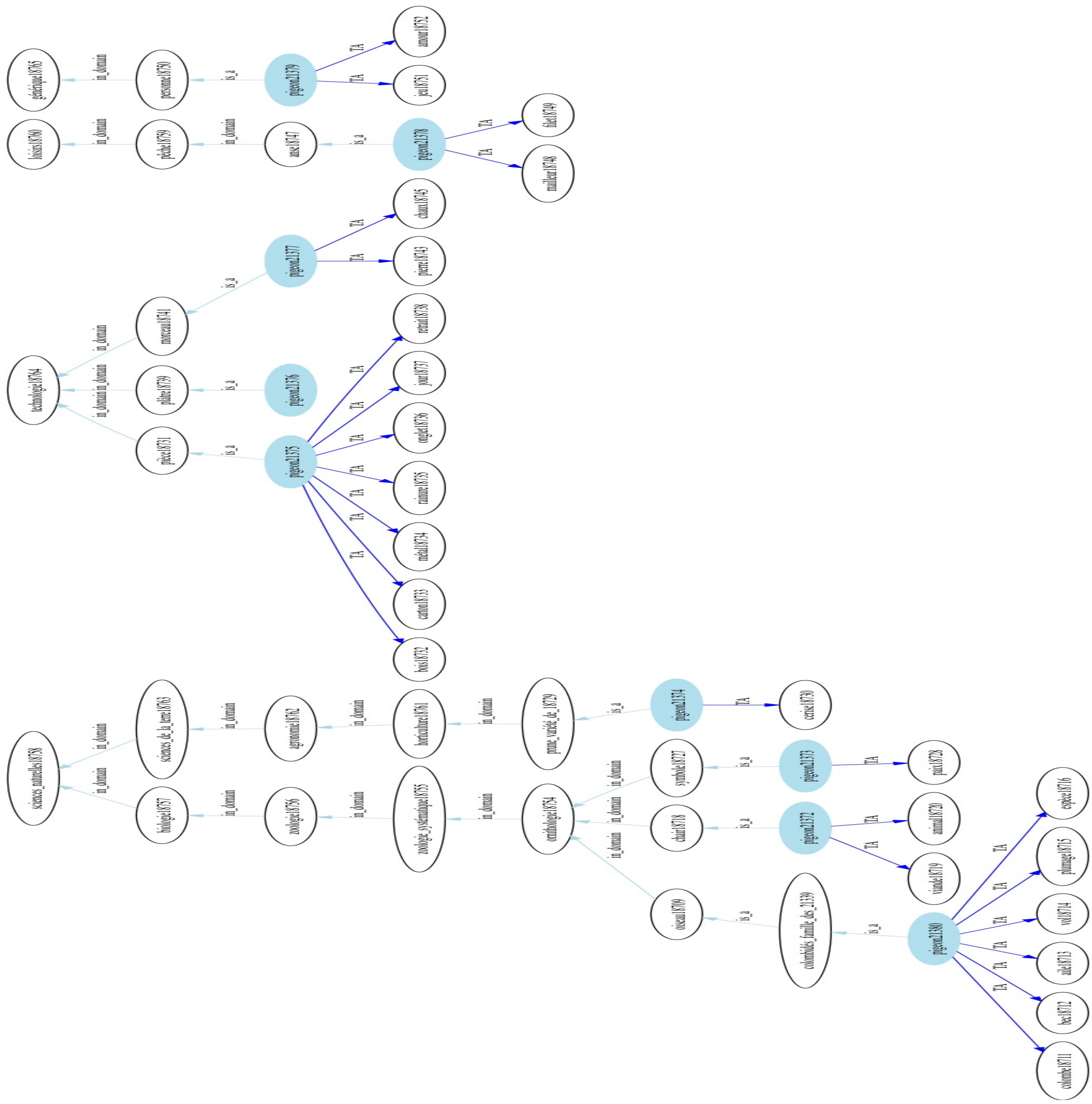

générique22331
in_domain
in_domain
culotte22305
pièce22307
is_a
is_a
pantalon23179
pantalon23180
TA
TA
TA
TA
TA
cou_de_pied22306
vêtement22308
costume22309
femme22310
civilisation22311
arts22328
in_domain
in_domain
arts_scéniques22327
danse22330
in_domain
in_domain
théâtre22326
chorégraphie22329
in_domain
in_domain
toile22320
figure22323
is_a
is_a
pantalon23181
pantalon23182
TA
TA
TA
découverte22321
décor22322
quadrille22324


générique11557
in_domain
mammifère11468
représentation11477
homme11519
taille11526
femme11529
is_a
lion11558
lion11559
lion11568
lion11569
lion11570
TA
TA_location
félidé_famille_des_11469
face11471
crinière11472
tronc11473
membre11474
pelage11475
état11476
afrique11542
animal11478
valeur11479
mâle11527
crinière11528
aspect11531
blondeur11532
chevelure11533
sciences_humaines11551
ethnologie11550
histoire11553
archéologie11549
héraldique11552
numismatique11548
figure11480
peuple11484
monnaie11503
lion11560
lion11561
lion11564
armoiries11481
attitude11483
image11504
arts11556
arts_graphiques11555
iconographie11554
figure11486
république11490
lion11562
lion11563
évangéliste11488
marc11489
venise11491
occultisme11545
alchimie11544
astrologie11546
symbole11506
signe11516
lion11565
lion11567
soufre11507
zodiaque11517
astronomie11547
constellation11508
lion11566
nom11509
511510
zodiaque11512
soleil11513
2
juillet11515


générique12516
in_domain
intérêt12454
public12466
vérité12479
défenseur12483
défense12488
justice12490
is_a
avocat15152
avocat15154
avocat15157
avocat15158
avocat15159
avocat15160
TA
réputation12468
jeu12480
tempérament12481
cause12482
cause12484
idée12485
opinion12486
système12487
justification12489
caractère12491
monopole12492
partie12493
conseil12494
état12495
cour12496
cassation12497
tribunal12498
conflit12499
prise12500
juridiction12501


sciences_naturelles12513
in_domain
biologie12512
botanique12511
baie12502
is_a
avocat15161
TA
comestible12503
forme12504
poire12505
pulpe12506
matière12507
fruit12508
avocatier12509


religion12515
in_domain
congrégation12469
conférence12475
is_a
avocat15155
avocat15156
TA
rite12470
canonisation12471
bienheureux12472
diable12474
dialogue12476
objection12477
religion12478


générique18765
loisirs18760
personne18750
pêche18759
pigeon21379
anse18747
amour18752
jeu18751
pigeon21378
filet18749
mailleur18748
chaux18745
pierre18743
pigeon21377
morceau18741
technologie18764
plâtre18739
pigeon21376
pièce18731
pigeon21375
retrait18738
jour18737
onglet18736
rainure18735
métal18734
carton18733
bois18732
sciences_naturelles18758
sciences_de_la_terre18763
biologie18757
agronomie18762
zoologie18756
horticulture18761
zoologie_systématique18755
prune_variété_de_18729
ornithologie18754
pigeon21374
cerise18730
symbole18727
pigeon21373
paix18728
chair18718
pigeon21372
animal18720
viande18719
oiseau18709
colombidés_famille_des_21339
pigeon21380
espèce18716
plumage18715
vol18714
aile18713
bec18712
colombe18711