Méthodes et modèles de construction automatisée d'ontologies pour des domaines spécialisés

Olena Goncharova

N° d'ordre NNT : 2017LYSE2018

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON
en cotutelle avec
L'UNIVERSITE POLYTECHNIQUE DE KHARKIV
(UKRAINE)

Opérée au sein de

L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 512 Informatique et Mathématiques

Discipline : Informatique

Soutenue publiquement le 23 février 2017, par :

Olena GONCHAROVA

Méthodes et modèles de construction automatisée d'ontologies pour des domaines spécialisés

Devant le jury composé de :

Marc BEZE, Professeur des universités, Université d'Avignon, Président

Sylvie DESPRES, Professeure des universités, Université Paris 13, Rapporteure

Christophe ROCHE, Professeur des universités, Université de Savoie, Rapporteur

Thierry HAMON, Maître de conférences, Université Paris 13, Examinateur

Jean-Hugues CHAUCHAT, Professeur émérite des universités, Université Lumière Lyon 2, Directeur de thèse

Natalia SHARONOVA, Professeure d'université, Co-directrice de thèse

Contrat de diffusion

UNIVERSITÉ LUMIÈRE LYON 2

Thèse présentée pour obtenir le grade de
Docteur de l'Université Lumière Lyon 2

École Doctorale Informatique et Mathématiques (ED 512)
Laboratoire ERIC (EA 3083)
Discipline : Informatique

Méthodes et modèles de construction automatisée d'ontologies pour des domaines spécialisés

Par : Olena Orobinska

Présentée et soutenue publiquement le 23 février 2017, devant un jury composé de :

Jean-Hugues Chauchat,	Professeur émérite des Universités, Universitè Lyon 2,	*Directeur*
Natalia Sharonova,	Professeur des Universités, UNT Technique "KhPI" de Kharkov,	*Directrice*
Sylvie DESPRES,	Professeur des Universités, Université Paris 13	*Rapporteur*
Christophe ROCHE,	Professeur des Universités, Université de Savoie	*Rapporteur*
Marc EL BEZE,	Professeur des Universités, Université d'Avignon	*Examinateur*
Thierry HAMON,	Maître de Conférences, Université Paris 13	*Examinateur*
Sylvie SZULMAN,	Maître de Conférences Honoraire, Université Paris 13	*Examinateur*

à ma mère, à sa mémoire
et à Andrey

Abstract

Abstract of the thesis of Ms Olena OROBINSKA, born GONCHAROVA

The thesis has been prepared within a co-supervision agreement with the Professors Jean-Hugues Chauchat (ERIC-Lyon2) and N.V. Charonova (National Polytechnic University of Kharkov in Ukrainia).

The results obtained can be summarized as follows :

1. State of the art :
   - Retrospective of theoretical foundations concerning the formalization of knowledge and natural language as precursors of ontology engineering.
   - Update of the state of the art on general approaches in the field of ontology learning, and on methods for extracting terms and semantic relations.
   - Overview of platforms and tools for ontology construction and learning ; list of lexical resources available online able to support ontology learning (concept learning and relationship).
2. Methodological proposals :
   - Learning morphosyntactic patterns and implementing partial taxonomies of terms.
   - Finding semantic classes representing concepts and relationships for the field of radiological safety.
   - Building a frame for the various stages of the work leading to the construction of the ontology in the field of radiological safety.
3. Implementation and experiments :
   - Loading of two corpuses specialized in radiological protection, in French and Russian, with 1,500,000 and 600,000 lexical units respectively.
   - Implementation of the three previous methods and analysis of the results obtained.

The results have been published in 13 national and international journals and proceedings, between 2010 and 2016, including IMS-2012, TIA-2013, TOTH-2014, Bionica Intellecta (Бионика интеллекта) , Herald of the NTU " KhPI " (Вестник НТУ " ХПИ ").

Key words. ontology learning, text processing, semantic analysis, terms extraction.

Résumé de la thèse de Mme Olena OROBINSKA, née GONCHAROVA

La thèse est préparée dans le cadre d'une convention de cotutelle sous la direction des Professeurs Jean-Hugues Chauchat (ERIC-Lyon2) et N.V. Charonova (Université Nationale Polytechnique de Kharkov en Ukraine).

Les résultats obtenus peuvent se résumer ainsi :

1. État de l'art :

— Rétrospective des fondations théoriques sur la formalisation des connaissances et langue naturelle en tant que précurseurs de l'ingénierie des ontologies.
— Actualisation de l'état de l'art sur les approches générales dans le domaine de l'apprentissage d'ontologie, et sur les méthodes d'extraction des termes et des relations sémantiques.
— Panorama des plateformes et outils de construction et d'apprentissage des ontologies ; répertoire des ressources lexicales disponibles en ligne et susceptibles d'appuyer l'apprentissage d'ontologie (apprentissage des concepts et relation).

2. Propositions méthodologiques :

— Une méthode d'apprentissage des patrons morphosyntaxiques et d'installation de taxonomies partielles de termes.
— Une méthode de formation de classes sémantiques représentant les concepts et les relations pour le domaine de la sécurité radiologique.
— Un cadre (famework) d'organisation des étapes de travaux menant à la construction de l'ontologie du domaine de la sécurité radiologique.

3. Implémentation et expérimentations :

— Installation de deux corpus spécialisés dans le domaine de la protection radiologique, en français et en russe, comprenant respectivement 1 500 000 et 600 000 unités lexicales.
— Implémentation des trois méthodes proposées et analyse des résultats obtenus. Les résultats ont été présentés dans 13 publications, revues et actes de conférences nationales et internationales, entre 2010 et 2016, notamment IMS-2012, TIA-2013, TOTH-2014, Eastern-European Journal of Eenterprise Technologies, Bionica Intellecta (Бионика интеллекта), Herald of the NTU « KhPI » (Вестник НТУ « ХПИ »).

Mots-clès. apprentissage des ontologies, traitement de texte, analyse sémantique, extraction de termes, extraction de relations.

Remerciements

Dans la vie de chacun d'entre nous, il y a des événements et les personnalités qui changent notre destin, en ouvrant les horizons nouveaux. Parmi ces événements, j'estime avoir eu la chance de rencontrer le professeur Jean-Hugues Chauchat, mon directeur de thèse du côté de l'Université Lyon 2, et d'avoir pu venir en France, grâce à son aide, pour mener mes travaux. Son style d'organisation des recherches et de travail, sa façon de penser et voir les choses, resteront pour moi les meilleurs exemples et une expérience inestimable.

Ce travail n'aurait pu être possible sans de nombreuses personnes auxquelles je tiens aussi à exprimer ma profonde gratitude.

L'encadrant du côté de l'Université National Technique « Institut Polytechnique de Kharkiv », La Professeure Natalia Sharonova, qui avait dirigé mes premiers pas en recherche dans le domaine d'intelligence artificielle et la présentation formelle des connaissances.

Je tiens à remercier mes rapporteurs les professeurs Christophe Roche et Sylvie Després, ainsi mes examinateurs les docteurs Sylvie Szulman et Thierry Hamon, pour me faire l'honneur d'accepter de juger ce travail. J'aimerais noter ici que les travaux scientifiques de chacun d'eux ont été pour moi des ressources importantes tout au long de mes recherches.

Je tiens à remercier le docteur Annie Morin qui a gracieusement accepté de faire un important travail de première révision, technique et substantielle ; son estimation positive m'a beaucoup encouragée.

Je remercie professeur Marc EL Beze pour ses observations détaillées qui m'ont permis d'améliorer la présentation des résultats et l'argumentation en faveur de nos méthodes et du choix des outils adaptés à leur réalisation.

Je remercie ensuite tous les membres de laboratoire ERIC avec lesquels j'ai partagé l'ambiance de travail, toujours intéressant et consciencieux et avec un bon esprit d'équipe et, en particulier, le professeur Jérôme Darmont, le directeur du laboratoire pour l'organisation des meilleures conditions de travail, la professeure Sabine Loudcher et les docteurs Julien Ah-Pin, Cécile Favre, Fabien Rico et Julien Velcin pour leurs nombreux conseils constructifs au cours de travaux de recherche.

Adien Guille, Pavel Soriano et Xinyu Wang pour l'aide technique lors de la réalisation des expérimentations et la présentation des résultats.

Je remercie les nombreuses personnes dont j'ai la fierté compter parmi ses amis et qui m'ont ouverte sur le vrai esprit de la France ; et tout particulièrement Julien Crevel dont le coup de main garantissait toujours le bon fonctionnement des logiciels et de l'ordinateur, Anne-Marie et Didier Augy, Delphine Rykhtik, Philippe Henry, Valérie Pietroforte et Max Beligne.

Enfin, je tiens à exprimer ma gratitude envers mon mari et toute ma famille pour le soutien constant, la compréhension et la patience avec lesquels mes proches ont accepté ma longe absence d'Ukraine.

Table des matières

Abstract **iii**

Remerciements **v**

1 Introduction **1**

1.1 Contexte de travail 1

1.2 Problématique et objectifs de la thèse 2

1.3 Contribution et publications 3

1.3.1 Contribution 3

1.4 Organisation de la thèse 5

2 Ontologie, connaissance, langue naturelle **6**

2.1 De l'Ontologie métaphysique aux ontologies appliquées 6

2.1.1 Profil historique du sujet 6

2.1.2 Du point vue de l'informatique 7

2.2 Domaines d'application des ontologies 9

2.2.1 Web Sémantique 9

2.2.2 Recherche d'information (Information Retrieval, IR) 9

2.2.3 Systèmes du type Question-Réponse 10

2.2.4 Intégration de bases de données hétérogènes 10

2.2.5 Ingénierie logicielle 10

2.3 Formalisation des connaissances : les aspects du problème 10

2.3.1 Interdépendance entre la connaissance et la langue 11

2.3.2 Les notions principales : données, information, connaissances, concepts . . . 12
2.4 Modèles de représentation des connaissances . . . 14
2.4.1 Modèles procéduraux . . . 16
2.4.2 Modèles déclaratifs . . . 16
2.5 Modélisation de la langue naturelle . . . 21
2.5.1 La linguistique structurale et le modèle fonctionnel . . . 21
2.5.2 Modèle génératif . . . 22
2.5.3 Sémantique vs syntaxe . . . 24
2.5.4 Les verbes dans les modèles linguistiques . . . 24
2.6 Les ressources lexicales . . . 25
2.6.1 FrameNet . . . 26
2.6.2 WordNet . . . 28
2.6.3 PyTez (RuTez) . . . 30
2.6.4 BabelNet . . . 31
2.6.5 VerbNet et VerbOcean . . . 32
2.7 Conclusion de chapitre 2 . . . 33

3 Apprentissage des ontologies : l'état de l'art général 34
3.1 Domaine de l'apprentissage des ontologies . . . 34
3.2 Panorama des outils . . . 35
3.3 Classification des ontologies . . . 36
3.3.1 Classification des ontologies selon leurs objectifs . . . 37
3.3.2 Classification des ontologies par degré de formalisation . . . 40
3.4 Méthodologies générales de construction des ontologies . . . 41
3.4.1 Methontology . . . 42
3.4.2 L'ingénierie des ontologies basée sur l'apprentissage (learning-driven ontology engineering process) . . . 43
3.4.3 Approche du Modèle d'Entreprise (Enterprise Model Approach) . . . 44
3.5 Projets et plateformes ontologiques . . . 45
3.5.1 TERMINAE . . . 45
3.5.2 text2Onto . . . 46

3.5.3 Lemon (Lexicon Model for Ontologies) 48
3.6 Méthodes pour l'apprentissage d'ontologie 49
3.6.1 L'Analyse de Concepts Formels, FCA 49
3.6.2 Méthode de l'identification comparative 55
3.7 Conclusion du chapitre 3 59

4 L'apprentissage des ontologies : méthodes et techniques 61
4.1 Spécificité de la tâche 61
4.2 La terminologie, une sous-langue spécialisée 62
4.3 Extraction des termes 63
4.3.1 Méthodes basées sur la fréquence 64
4.3.2 Méthodes basées sur les corpus contrastés 66
4.3.3 Méthodes basées sur la mesure d'association entre les mots 67
4.3.4 Méthodes basées sur le contexte 71
4.3.5 Méthodes basées sur les thématiques (topic-based) 73
4.4 Extraction des relations 74
4.4.1 Extraction des relations – une tâche à plusieurs niveaux 74
4.4.2 Classification des méthodes d'extraction de relations 75
4.4.3 Classification des relations 76
4.4.4 Les méthodes d'extraction des relations 78
4.5 Conceptualisation 80
4.6 Conclusion du chapitre 4 82

5 Nos expérimentations 83
5.1 Introduction 83
5.2 Domaine d'application : la radioprotection 84
5.2.1 Le choix du domaine 84
5.2.2 Les principaux aspects de la radioprotection 85
5.2.3 L'organisation de la collaboration avec un spécialiste du domaine . 86
5.3 Notre approche : cadre général 87
5.4 Acquisition des ressources 89
5.4.1 Installation des deux corpus 89

5.4.2 Étiquetage des textes 90
5.4.3 Sélection des ressources lexicales 92
5.5 Méthodes et Résultats 93
5.5.1 Modélisation sémantique et linguistique du noyau d'ontologie . . . 93
5.5.2 Cadre prédicatif 101
5.5.3 Méthode des patrons terminologiques 109
5.5.4 Règles de reconnaissance 115
5.5.5 Synthèse des résultats 116
5.6 Conclusion du chapitre 5 117

6 Conclusion et discussion **120**
6.1 Résumé du travail 120
6.2 Perspectives de travail 121

Annexe A **123**

Annexe B **127**

Annexe C **131**

Table des figures

1.1 L'univers des ontologies. 2

2.1 Arbre de Porphyre (selon Petrus Hispanus). 8
2.2 Le triangle sémiotique générique. 12
2.3 La pyramide de la sagesse DIKW. 12
2.4 Les modèles de représentation des connaissances. 15
2.5 Exemple de présentation des connaissances en réseau. 18
2.6 Les grandes périodes de la sémantique lexicale. 22
2.7 Visualisation des principales catégories du cadre de RISQUE. 27
2.8 Fragment du modèle de domaine. 29
2.9 Un aperçu du fonctionnement de BabelNet. 31

3.1 Distribution des outils employés dans l'ingénierie des ontologies. 36
3.2 Classification des ontologies selon leurs objectifs. 38
3.3 L'ontologie de représentation d'OWL. 38
3.4 « Diamant » de SOWA. 39
3.5 Spectre des ontologies. 40
3.6 Méthodologie généraliste selon *Simperl et al.* (2010). 43
3.7 Fonctionnalité de la plateforme TERMINAE. 46
3.8 Schema modulaire de text2Onto. 47
3.9 Schéma de Lemon. 49
3.10 Module syntaxique de Lemon. 49
3.11 L'ensemble des étapes de la construction d'une ontologie avec FCA. 50
3.12 Le treillis correspondant au contexte formel donné. 52

3.13 Transformation du treillis en l'ontologie et définition des concepts. 54
3.14 L'ontologie complétée par les relations associatives. 55
3.15 Hiérarchie des notions. 59

4.1 Ontology learning "layer cake". 61
4.2 Enchaînement des tâches, indication des techniques adoptées pour chacune et éléments d'ontologie produits. 62

5.1 Éléments du système de gestion des risques radiologiques. 85
5.2 La construction du noyau de l'ontologie. 88
5.3 Flot de travail pour la construction de l'ontologie de domaine. 89
5.4 Schéma de la méthode d'installation du noyau d'ontologie. 94
5.5 Corpus français : les verbes associés au concept *dommage*, rangés selon le coefficient d'association ***K*** = 1,3E-4 (la formule 5.1) 97
5.6 Corpus russe : les verbes associés au concept *ущерб* (*dommage*), rangés selon le coefficient d'association ***K*** = 1,3E-4 (la formule 5.1) 98
5.7 Quantification de la similarité entre le verbe *protéger* et ses synonymes. . . 108
5.8 Taxonomie formée par « déploiement » autour du mot *dose*. 109
5.9 Schéma de notre système d'enrichissement de l'ontologie par la terminologie dérivée de la liste des termes génériques. 110
5.10 Taxonomie avec la racine *NOM+ADJ*. 111
5.11 Taxonomie avec la racine *NOM+PP*. 111
5.12 Taxonomie avec la racine *NOM+PRP+NOM*. 112

6.1 Triplet de RDF. 133
6.2 La hiérarchie historique des langages de représentation des ontologies. . . 135
6.3 Classe « `Protection` ». 135

Liste des tableaux

2.1 Les définitions des notions générales. 13

2.2 Les propriétés permettant de distinguer la donnée, l'information et la connaissance. 14

2.3 Comparaison des modèles de présentation des connaissances. 21

2.4 La correspondance entre le cadre RISQUE et notre modèle. 28

3.1 Catégories des outils liés au développement, maintenance, alignement et évaluation des ontologies (selon leur fonction). 36

3.2 Distribution des outils selon les langages de programmation et accessibilité au code. 37

3.3 Exemple du contexte formel représentant les propriétés reliant les organes et les pathologies cancéreuses. 53

3.4 Formalisation du passage du treillis à l'ontologie. 53

3.5 Définition des concepts de l'ontologie en Logique Descriptive. 55

3.6 Le fragment du contexte formel pour les radionucléides. 55

3.7 Treillis des relations entre les concepts. 56

4.1 Récapitulatif des caractéristiques fréquentielles pour l'extraction des termes. 65

4.2 Types de relations entre les termes. 78

4.3 Correspondance entre les relations ontologiques et les prédicats. 79

5.1 Liste réduite de nos 17 balises morpho-syntactiques pour le français, regroupant celles de TreeTagger. 91

5.2 Liste réduite de nos 8 balises morpho-syntactiques pour le russe, regroupant celles de TreeTagger. 91

5.3 Nombre de synonymes du terme *dommage* dans les différents dictionnaires. 92

5.4 Informations agrégées sur les deux corpus. 97

5.5 Résumé des résultats de labellisation de termes généraux en français. . . . 98
5.6 Résumé des résultats de la labellisation des termes généraux en russe. . . . 99
5.7 Résumé des résultats de la labellisation des termes généraux en français. . 99
5.8 Résumé des résultats de labellisation des termes généraux en russe. 100
5.9 Correspondance entre les résultats de labellisation de termes généraux en français et en russe. 100
5.10 Récapitulatif sur l'extraction des triplets SVO à partir des concepts initiaux. 104
5.11 Les catégories de relations sémantiques entre les concepts du domaine de la Sécurité Radiologique. 104
5.12 Exemples des verbes liant les concepts du noyau de l'ontologie. 105
5.13 Les noms et les verbes du corpus reconnus par CRISCO. 105
5.14 Distribution des nombre de synonymes par mot dans le corpus. 105
5.15 Sélection des éléments des classes sémantiques. 107
5.16 Liste des verbes spécifiques au domaine, présents dans le corpus et absents du dictionnaire. 108
5.17 Liste des noms spécifiques au domaine, présents dans le corpus et absents du dictionnaire. 108
5.18 Score de validations des patrons terminologiques formés automatiquement. 112
5.19 Patrons terminologiques formés à partir du glossaire. 113
5.20 Liste finale de nos patrons terminologiques avec des exemples. 114
5.21 Résultats de l'extraction des candidats-termes extraits au moyen des patrons terminologiques. 118
5.22 Exemples de taxonomies de termes dérivés. 119
5.23 Nombres d'inductions de taxonomies partielles à partir de trois racines syntaxiques. 119
5.24 Exemples de règles de reconnaissance. 119

Chapitre 1

Introduction

1.1 Contexte de travail

L'accroissement constant des connaissances humaines (*Vernadski* (2012)) induit une demande de méthodes et d'outils pour les présenter, les stocker et les rendre opérationnelles. La communauté scientifique cherche à développer des agents logiciels capables de traiter les informations de manière intelligente.

Les ontologies actuelles en tant qu'outils prometteurs de présentation et de partage opérationnel des connaissances ont deux objectifs. Tout d'abord, les ontologies définissent les unités conceptuelles qui comprennent les concepts eux-mêmes (interprétés souvent comme un vocabulaire univoque partagé par la communauté) et les relations qui les rassemblent. Deuxièmement, elles doivent permettre d'enregistrer les propriétés pragmatiques et formelles de ces unités.

L'apprentissage des ontologies, et toute l'ingénierie d'ontologie dont l'apprentissage est une partie, visent à réaliser ces deux objectifs : le premier par la réalisation d'un modèles sémantique indépendant de son implémentation, et le deuxième par l'élaboration de langages à la fois riches en l'expressivité et rigoureux, permettant le raisonnement non contradictoire.

De nombreux ouvrages exposent les principes généraux pour le développement de logiciels de construction d'ontologies et d'autres systèmes dit intelligents, notamment *Hailpern et Tarr* (2006), *Uschold* (2008a), *Nicola et al.* (2008), *Cimiano et al.* (2006), *Paquette* (1996b).

Pour les humains, les connaissances s'expriment sous de nombreuses formes (paroles, musique, dessin, construction, outils, geste, mimiques...) ; parmi elles, le discours, oral ou écrit, joue un rôle majeur. Les théories de la langue naturelle sont nécessaires pour modéliser son fonctionnement et travailler avec les ordinateurs car, pour les machines, les connaissances ne se manifestent qu'à travers des symboles, des signes qui peuvent être transmis, reçus, stockés et interprétés.

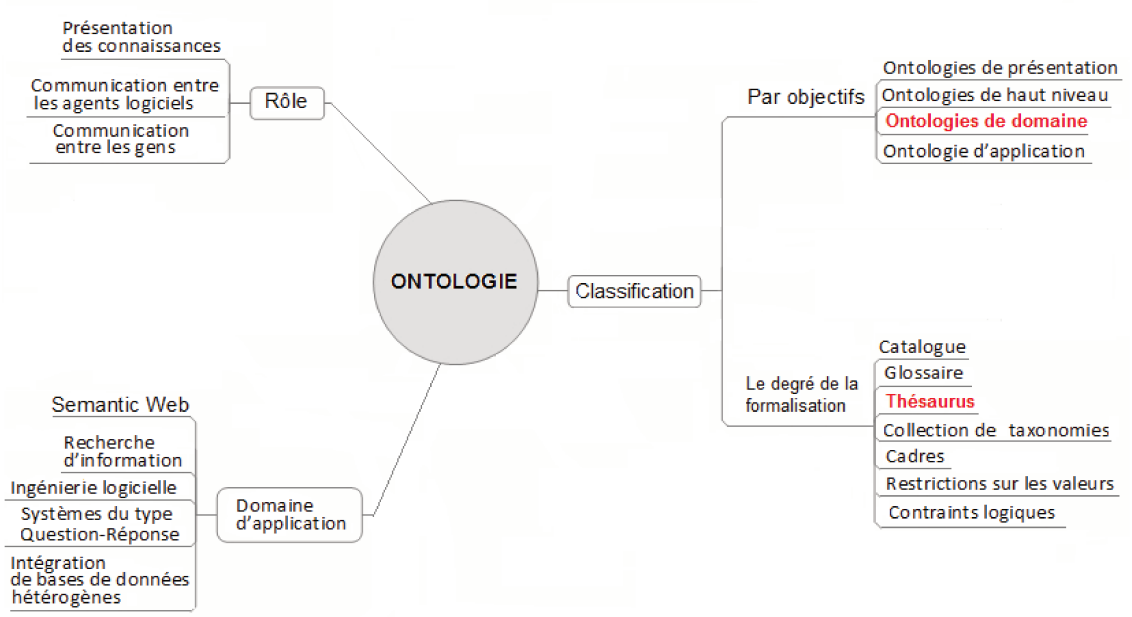

Figure 1.1: L'univers des ontologies.

Pour l'apprentissage des ontologies, nous utilisons comme sources de connaissances des corpus de textes écrits en langue naturelle et des outils en ligne conçus par des linguistes : parseurs et dictionnaires de synonymes. Notre démarche part de l'interprétation formelle des connaissances et de l'interprétation systémique de la langue, puis des méthodes concrètes de détection des éléments du texte portant la sémantique spécifique du domaine.

1.2 Problématique et objectifs de la thèse

Dans le cadre de cette thèse, nous nous focalisons sur la dimension sémantique des ontologies : nous développons une approche qui amorce la construction d'une ontologie de domaine à partir d'une liste de termes génériques et d'un cadre prédicatif limitant l'interprétation de ces derniers. Nous cherchons à trouver un équilibre entre :

- les méthodes statistiques utilisées pour détecter le lexique de domaine (les termes) dans un grand corpus ;
- une analyse syntaxique de profondeur suffisante pour définir les interdépendances sémantiques entre les notions auxquelles se réfèrent ces termes ;
- la nécessité d'utiliser les compétences des experts.

En résumé : en fonction de la classification des ontologies présentée sur la figure 1.1 nous cherchons à créer une ressource terminologique sous la forme d'un thésaurus enrichi par des relations non-taxonomiques entre les concepts.

Ici, les concepts comme les relations seront présentés sous forme de classes sémantiques de synonymes : ensembles de termes-synonymes pour les concepts, et ensemble de verbes-synonymes pour les relations. Au départ, chaque concept est repéré par un seul nom, complété ensuite par un ensemble de termes synonymes ; au total, 1850 termes ont été retenus. De même, chaque relation est définie par un ensemble de verbes-synonymes. Pour construire et enrichir progressivement ces ensembles, on utilisera les corpus de textes du domaine, des analyseurs syntaxiques, des dictionnaires en lignes et, à la fin de chaque étape, l'expertise d'un spécialiste de la sécurité radiologique.

Au cours du travail, on peut distinguer deux catégories de problèmes à résoudre : premièrement « les problèmes appliqués », liés directement à l'analyse du domaine choisi ; il faut identifier la nature des informations portant des connaissances sur le domaine : visuelles, sonores, enregistrées sous forme d'autres signaux, etc. Bien que nous ne prenions en considération que les informations textuelles, elles sont tout de même très hétérogènes : le lexique spécifique du domaine, les entités nommées, les grandeurs physiques et leurs unités de mesure, les symboles d'éléments chimiques, des sigles, etc.

La deuxième catégorie de problèmes est le choix d'une structure modulaire de système d'information (SI), des ressources auxiliaires et de l'organisation des flux de travaux qui vont permettre l'analyse des textes complets à différents niveaux de granularité.

Les problèmes de modélisation d'une ontologie ont été analysés par *Noy et McGuinness* (2001), *Noy et al.* (2006), *Maedche et Staab* (2000), *Buitelaar et Cimiano* (2008), *Simperl et al.* (2008) ; les auteurs concluent que, quelle que soit la stratégie de construction de l'ontologie (top-down ou bottom-up), il est impossible d'élaborer d'une seule traite le modèle de domaine et de prévoir à l'avance toutes ses classes et leurs relations ; par ailleurs, les critères de choix des concepts sont assez subjectifs ; il faut donc faciliter le travail des experts en leur proposant les listes, les plus complètes possibles, d'entités potentiellement pertinentes extraites automatiquement du corpus.

Notre premier objectif est de trouver une approche d'analyse des textes en langue naturelle, puis de mettre en œuvre les méthodes d'extraction de connaissances du domaine.

Notre deuxième objectif est d'établir la structure du système qui appliquera ces méthodes.

Pour cela, nous avons analysé et comparé les modèles de représentations des connaissances et les approches de description et d'analyse des langues naturelles.

1.3 Contribution et publications

1.3.1 Contribution

Deux grands corpus spécialisés, en français et en russe, ont été mis en place ; ces corpus, en partie parallèles, sont composés de textes sur la sécurité radiologique.

Un modèle conceptuel du domaine de sécurité radiologique a été élaboré à l'aide d'un expert du domaine. Sa validité a été confirmée par les résultats expérimentaux.

Au cours de ce travail de thèse, nous avons proposé trois méthodes pour extraire des termes, installer un noyau d'ontologie, puis l'enrichir par la labelisation des concepts et par la définition des relations qui les lient.

La première méthode permet d'établir le noyau d'ontologie sous forme de l'ensemble des classes sémantiques des noms représentant les concepts de l'ontologie sous-jacente.

La deuxième méthode permet d'établir les relations associatives entre les concepts de l'ontologie sous-jacente (noyau d'ontologie) au moyen des classes sémantiques des verbes (nous appelons « cadre prédicatif » l'ensemble de toutes ces classes).

La troisième méthode se base sur l'utilisation des patrons morpho-syntaxiques (nous les appelons « patrons terminologiques »). Elle permet de compléter et d'enrichir l'ontologie sous-jacente par des termes multi-mots. La nouveauté de la méthode consiste en ce que les patrons sont installés de manière automatique et qu'ils permettent de former les taxonomies des classes d'ontologie.

Enfin, nous présentons la structure du système d'information qui généralise notre approche pour la construction de l'ontologie de domaine.

Les résultats de nos travaux ont été présentés dans 13 publications, revues et actes de conférences nationales et internationales, entre 2010 et 2016, notamment IMS-2012, TIA-2013, TOTH-2014, Eastern-European Journal of Eenterprise Technologies, Bionica Intellecta (Бионика интеллекта), Herald of the NTU « KhPI » (Вестник НТУ « ХПИ »), Herald of Kherson National Technical University (Вестник Херсонского Национального Техничского Университета).

Une partie des publications est mentionnée ci-dessous :

1. Orobinska O., Chauchat J-H., Sharonova N. 2016. Formation semi-automatique de classes sémantiques couvrantes pour enrichir une ontologie de domaine. Dans les actes en ligne du 13-éme atelier sur la Fouille de Données Complexes (FDC) de la conférence Extraction et Gestion des Connaissances (EGC 2016), France, 2016, Caen, pp.51-63 ;
2. Orobinska O., Chauchat J-H., Charonova N. Application de ressources linguistiques à grande échelle pour le peuplement d'une ontologie de domaine. TOTH-2014 , VII Iconférence internationale Terminology & Ontology : Theories and applications, University of Savoie, Chambéry, France, 2014, June 12-13 ;
3. Orobinska O. Chauchat J-H., Charonova N. Enrichissement d'une ontologie de domaine par extension des relations taxonomiques à partir de corpus. In proceeding of 10th International Conference on Terminology and Artificial Intelligence, 2013 Octobre 25-26, Paris-13 ;

4. Automatic Method of Domain Ontology Construction based on Characteristics of Corpora POS-Analysis. In Proceedings of the XV International Conference IMS-2012, St-Petersburg, 2012. October 10 – 12. P.209-212 ;
5. Generalizing Framework for ontology learning. In the first Ukrainian Conference on Intelligent Systems and Applied Linguistics – 2012. – March 15-16 ;
6. Orobinska O., Sharonova N. 2011. Ontology construction from text's corpus with FCA. In journal Bionics of Intelligence : Sci. Mag. 2011, 2 (76) p. 129-135.

1.4 Organisation de la thèse

Ce mémoire est organisé comme suit. Le chapitre 2 aborde les fondements théoriques des ontologies de l'Antiquité à nos jours : - la représentation formalisée des connaissances, et - l'évolution de la manière de concevoir la langue naturelle et ses théories ; ces théories générales fondent les travaux informatiques contemporains de construction des ontologies. La section 2.6 du chapitre 2 décrit les ressources lexicales construites pour la Recherche d'Information (Information retrieval, IR), notamment l'impact de FrameNet sur la formation initiale du modèle du domaine de la sécurité radiologique (SR).

Le chapitre 3 commence par l'état actuel du domaine d'application des ontologies (section 3.1), suivi du panorama des outils (section 3.2) et de la classification des ontologies (section 3.2). La section 3.4 présente les stratégies mises en œuvre pour suivre le cycle de vie complet d'une ontologie. La section 3.5 décrit en détail la logique de plusieurs plateformes actuelles d'apprentissage des ontologies (l'annexe A présente une liste plus complète de projets avec leur caractéristiques).

Le chapitre 4 présente un état de l'art approfondi des méthodes d'extraction des termes et des relations (sections 4.3 et 4.4), préfiguré par la spécification des taches à résoudre au cours de la construction d'une ontologie de domaine (section 4.1) ainsi que par l'analyse des propriètès principales de la terminologie (section 4.2).

Dans le chapitre 5 nous présentons le domaine de la sécurité radiologique, nos méthodes et les résultats de nos expérimentations. Nous présentons trois méthodes pour la construction complète d'une ontologie de domaine et leur application à la sécurité radiologique. Nous plaidons pour l'importance d'une analyse linguistique approfondie (morphologique, syntaxique et sémantique) du corpus des textes. Le chapitre se conclut par une discussion des problèmes ouverts pour l'apprentissage d'ontologie.

La diversité des langages actuels de représentation des ontologies, et leurs avantages et inconvénients sont discutés brièvement dans l'annexe C.

La conclusion de chaque chapitre décrit les progrès et les problèmes du sujet abordé.

Chapitre 2

Ontologie, connaissance, langue naturelle

2.1 De l'Ontologie métaphysique aux ontologies appliquées

Le terme *ontologie* est fortement polysémique et désigne un ensemble de notions reliées mais non identiques qui s'inscrivent dans les registres scientifiques distincts. Envisagé dans l'acception de l'ingénierie d'ontologie, il devient fédérateur dans la mesure où il correspond à une visée descriptive commune des objets que toutes les disciplines concernées par la présentation formelle des connaissances essaient de réaliser. Dans ce chapitre nous allons évoquer brièvement trois axes de recherches liés au développement des ontologies en informatique : la philosophie, l'ingénierie des connaissances et la linguistique.

2.1.1 Profil historique du sujet

Le mot *ontologie* a des racines grecques antiques où la syllabe *on* signifie l'être et *logos* fait référence aux notions telles que *le raisonnement*, *le mode de pensée* mais également à *l'enseignement* et *l'apprentissage*. Jusqu'à la fin du vingtième siècle l'ontologie faisait partie de la philosophie qui analyse le monde à travers des catégories générales. Un autre aspect de l'ontologie est la tentative de définir l'essence de l'homme et à prouver la correspondance entre le monde et les capacités de l'homme à le comprendre et le décrire.

En Europe, les premiers essais de construction du paysage ontologique du monde ont été entrepris par les Grecs anciens dont les noms célèbres sont si nombreux qu'il est impossible d'en dresser une liste exhaustive. De l'antiquité au 20-ème siècle, citons Socrate, Platon, Aristote, Emmanuel Kant, François-René de Chateaubriand, Edmund Husserl, Martin Heidegger, Ludwig Wittgenstein, Vladimir Vernadski, etc.

Le terme *ontologie* lui-même a été inventé par le philosophe allemand Jacob Lorhard en 1606 ; 9 ans plus tard, en 1613, Rudolf Goclenius, un autre philosophe scolastique allemand, en a donné la définition explicite : l'étude de nature de l'être en tant que tel (ou de la réalité) et ses catégories de base et leurs relations, *Øhrstrøm et al.* (2008). Pour les philosophes scolastiques médiévaux l'Ontologie était la théorie et la méthodologie de la pensée.

Au milieu du XX-ème siècle, la pensée scientifique se déplace vers la recherche d'outils pour formaliser la description et la manipulation des connaissances accumulées. On développe alors des langues artificielles et des grammaires formelles, telles que les algèbres de logique en mathématique, et des théories linguistiques. Ces idées commencent à être mises en œuvre à partir des années 1980 : les systèmes experts sont devenus les premiers artefacts d'ingénieurs basés sur les connaissances. Ils ont donné naissance à de nouvelles disciplines telles que l'intelligence artificielle (AI), l'ingénierie des connaissances, puis l'ingénierie d'ontologie.

De l'Ontologie philosophique (ou métaphysique) l'ingénierie d'ontologie a hérité du principe universel de *catégorisation*, élaboré par les Grecs dans l'Antiquité. Il est exposé dans les ouvrages d'Aristote réunis sous le titre « *La métaphysique* », notamment les ouvrages « *La substance* » et « *Les catégories* », où Aristote propose d'examiner toutes les manifestations du monde observables à travers deux types de caractéristiques : essentielles, qui perdurent au fils du temps, et transitoires qui changent facilement. Ainsi, il distingue dix catégories : l'action, l'état, le lieu, la passion, la qualité, la quantité, la relation, la situation, la substance, le temps. *Roche* (2005).

L'ouvrage de Porphyre « *Introduction aux Catégories d'Aristote* » (*Isagogè*) prolonge l'idée aristotélicienne : on y trouve, pour la première fois, la hiérarchie des objets nommée Arbre de Porphyre. Le concept placé en tête de l'arbre est caractérisé par une paire de propriétés distinctes qui permettent de spécifier deux nouvelles catégories d'objets possédant, chacune, l'une de ces propriétés. Pour chacune d'elles, de même, on trouve une nouvelle paire de propriétés qui définissent une spécification plus étroite. On continue jusqu'à ce que la division devienne impossible ou inopportune. La figure 2.1 présente un exemple de l'arbre de Porphyre selon Petrus Hispanus (Pierre d'Espagne).

2.1.2 Du point vue de l'informatique

En informatique, le terme *ontologie* apparaît au début des années 1990 lorsque la DARPA (Defense Advanced Research Projects Agency), agence du département de la Défense des États-Unis, a sponsorisé le consortium KSE (Knowledge Sharing Effort). L'objectif principal du KSE est l'élaboration de méthodes et techniques d'acquisition et de réutilisation des connaissances par (et pour) des systèmes d'informations (SI) à base de connaissances (SBC). Ils doivent garantir que l'interprétation des informations transmises d'un système à l'autre demeure saine, même si les agents logiciels utilisent des noms différents pour les désigner. Pour cela, les SBC associent de façon modulaire, deux types de connaissances : les connaissances explicitement spécifiées sur le domaine en

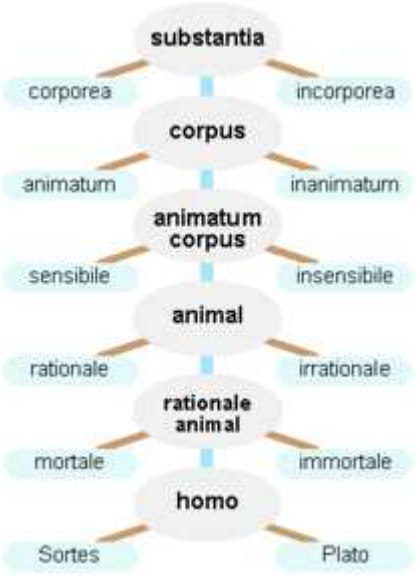

Figure 2.1: Arbre de Porphyre (selon Petrus Hispanus).

question, on les appelle « *les connaissances déclaratives* », et les règles de leur utilisation, i.e. « *les connaissances procédurales* ».

L'utilisation du terme *ontologie* pour la dénomination du module déclaratif a été proposée par Robert Neches, l'un des fondateurs de KSE, *Neches et al.* (1991).

Citons d'autres définitions qui font écho à l'interprétation philosophique. Ainsi, Ch. Roche définit une ontologie comme la « *représentation d'une modélisation d'un domaine sous la forme d'un ensemble de concepts définis par intension de relations et de propriétés logiques* », *Roche* (2005). B. Bachimont définit une ontologie comme la « *tâche de modélisation menée à partir de l'expression linguistique des connaissances* », *Bachimont* (2000).

Dans *Kister et al.* (2011) les auteurs envisagent l'ontologie dans un contexte de partage des connaissances où elle « *constitue un outil de structuration des connaissances et nécessite la prise en compte de relations sémantiques* ».

Une des premières définitions d'ontologie en tant qu'artefact d'ingénieur se trouve dans *Gruber* (1995) : « *an ontology is an explicit specification of a conceptualization* ». D'autres définitions comparables sont données dans *DeNicola et al.* (2005), *Horridge et al.* (2004).

2.2 Domaines d'application des ontologies

Dans cette section nous allons aborder brièvement les domaines où sont requises les ontologies, en tant que telles, et les domaines appliquant les principes méthodologiques fondant leur construction. La diversité des applications possibles montre l'actualité du problème de la construction automatisée des ontologies.

2.2.1 Web Sémantique

Aujourd'hui, le Web Sémantique est le plus grand domaine d'application des technologies sémantiques. Son idée fondamentale, formulée par Tim Berners-Lee en 2001, est d'accompagner l'extension et la croissance à long terme du Web actuel dans le cadre des recommandations de W3C [1]. Aujourd'hui, afin de garantir l'interopérabilité sur la toile on vise à compléter les ressources textuelles par des informations permettant leur interprétation univoque par des personnes et des ordinateurs. Depuis 2001 sont apparus de nombreux outils, techniques et langages de description des ontologies (cf. le chapitre 3).

2.2.2 Recherche d'information (Information Retrieval, IR)

Ce domaine couvre des activités telles que :

— Recherche des documents qui contiennent les informations pertinentes, correspondant aux requêtes de l'utilisateur. La plupart des moteurs de recherche réalisent l'indexation des textes à l'aide du modèle vectoriel où chaque texte est présenté comme « sac de mots » (*bag of words*). Les principaux inconvénients de cette approche sont :

  1. la redondance des index, les mêmes notions étant dénommées par les mots différents ;
  2. les mots d'un document sont considérés comme indépendants, ce qui ne correspond évidemment pas à la réalité ;
  3. les mots sont polysémiques, ce qui induit des ambiguïtés et aboutit à résultats non pertinents pour l'utilisateur. On peut pallier ces inconvénients par l'utilisation d'une indexation conceptuelle à l'aide des ontologies de domaines ; les concepts sont alors associés aux termes correspondants et liés par des relations prédéfinies.

— Classification de documents, i.e. l'attribution de chaque document à l'une des catégories prédéfinies.

— Regroupement sémantique, i.e. rassemblement des documents dont les sujets sont proches. Dans ce cas, le travail principal est la définition des rubriques autour

1. World Wide Web Consortium

desquelles les documents doivent être réunis. L'extraction des rubriques (*topic extraction*) est une des tâches actuelle de l'ingénierie des connaissances ayant des méthodes communes avec l'apprentissage des ontologies *Rizoiu et Velcin* (2011).
— Production de résumés automatiques.

2.2.3 Systèmes du type Question-Réponse

Dans ce cas, il s'agit du développement de systèmes interactifs du type Question-Réponse pour que l'utilisateur obtienne un résultat concret, et pas seulement une liste de références aux documents correspondant plus ou moins à sa requête-question.

2.2.4 Intégration de bases de données hétérogènes

L'intégration de bases de données hétérogènes est un problème complexe qui est devenu crucial pour fournir aux utilisateurs une interface unifiée permettant l'accès (par des requêtes) à des ressources hétérogènes. Dans ce cas, les ontologies sont utilisées pour spécifier le contenu des ressources hétérogènes.

2.2.5 Ingénierie logicielle

Le principe de conceptualisation et de distinction des objets selon leurs propriétés est universel ; il est aussi utilisé en ingénierie logicielle (SE – Software Engineering). Depuis au moins 20 ans, la tendance est à l'unification et la spécification des processus pendant tout le cycle de vie d'un Système d'Information. Il s'agit de la stratégie MDD – Model-Driven Development – et de l'architecture de construction de logiciel basée sur modèles, MDA – Model Driven Architecture. La stratégie MDD permet d'économiser du temps et des ressources, de réduire les charges et de garantir la flexibilité des processus de mise en œuvre, de la maintenance, des tests et simulation et l'interopérabilité des SI grâce aux dessins de modules et modèles de données lisibles automatiquement *Happel et Seedorf* (2006), *Miller et Mukerji* (2003).

2.3 Formalisation des connaissances : les aspects du problème

Le concept de *connaissance* est étudié par plusieurs disciplines : psychologie, philosophie, informatique, pédagogie, linguistique... Voici quelques travaux classiques qui abordent le problème de la formalisation des connaissances dans les domaines mentionnés : *Ackoff* (1989), *Popper* (1979), *Minsky* (1980), *Apresjan* (1973), *Yourdon* (1989), *Winograd* (1972), *Chomsky et Braudeau* (1969), *Wittgenstein* (2005), *Paquette* (1996b), *Kayser* (1997) et beaucoup d'autres que nous allons citer au fur et à mesure.

La formalisation des représentations de connaissances s'est développée durant les années 1980-1990, poussée par le développent des systèmes experts, et des bases de connaissances dans le cadre de l'intelligence artificielle. Dès le début, l'utilisation de la théorie des réseaux et des graphes a induit de nouvelles formes de représentation des connaissances, et notamment des ontologies.

Dans un cadre plus général, l'interprétation des connaissances comprend l'élaboration des conventions agrégées, basées sur les faits : « *knowledge consists of beliefs with the right objective connection to facts* » *Dretske* (2000), car on ne peut pas définir les connaissances sans évoquer les procédés par lesquels elles sont exprimées dans la langue naturelle. Pendant longtemps, les essais de modélisation des connaissances et les études sur la langue naturelle ont été abordés de façon dispersée par différentes sciences, telles que la philosophie, la psychologie, les mathématiques, la linguistique, avant que leurs apports respectifs ne soient réunis dans le cadre de nouvelles disciplines synthétiques telles que l'intelligence artificielle, la linguistique sémantique, la linguistique computationnelle.

Les notions de « donnée », d'« information », de « connaissance », de « concepts » sont très générales mais elles ont les significations différentes selon les disciplines. Elles font également partie du lexique de l'ingénierie d'ontologie et, en particulier, du domaine apprentissage des ontologies. Pour cette raison, nous allons présenter une brève synthèse de leurs définitions dans la section 2.3.2.

2.3.1 Interdépendance entre la connaissance et la langue

L'étude scientifique de sujets tels que la langue naturelle et la formalisation des connaissances sont difficiles car ils se présentent à la fois comme objet et comme instrument d'investigation. C'est dans le cadre de la sémiotique qu'on a analysé, pour la première fois, la corrélation entre l'univers de l'existant et celui de nos pensées où la langue, en tant que système de signes, sert d'intermédiaire matériel.

Ch. Pearce, G. Frege, G. de Saussure ont contribué à la compréhension de ces catégories. Ch.Ogden et I. Richard ont proposé de visualiser le modèle sous forme du « triangle sémiotique [2] » (cf. la figure 2.2).

Une ontologie, en tant que système formel de signes, s'inscrit parfaitement dans cette présentation où elle « joue le rôle » de langue naturelle dans les systèmes d'information. Ainsi le triangle sémiotique peut être facilement transformé en triangle « ontologique » : les sommets (*signifié, signifiant, référent*) sont interprétés comme (*concept, terme, instance*), respectivement. Les relations entre les sommets ne changent pas.

En informatique les sommets du triangle sémiotique peuvent être associés aux notions de *données*, d'*information* et de *connaissances*. Ici leur connexité se présente autrement, sous forme de la « pyramide de sagesse » (cf. figure 2.3), dit DIKW (Data-

2. Dans la littérature russophone on utilise souvent la nomination « triangle sémiotique de Frege » pour souligner son apport.

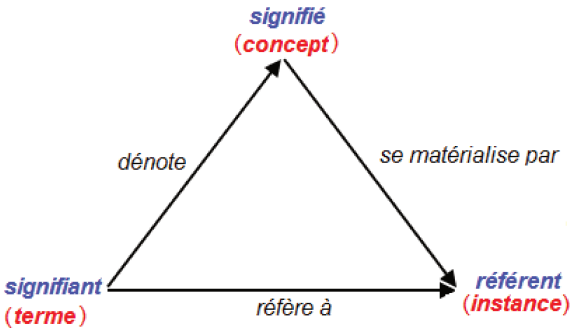

Figure 2.2: Le triangle sémiotique générique.

Information-Knowledge-Wisdom), initialement spécifié par R.L. Ackoff en 1988, cité dans *Rowley* (2007).

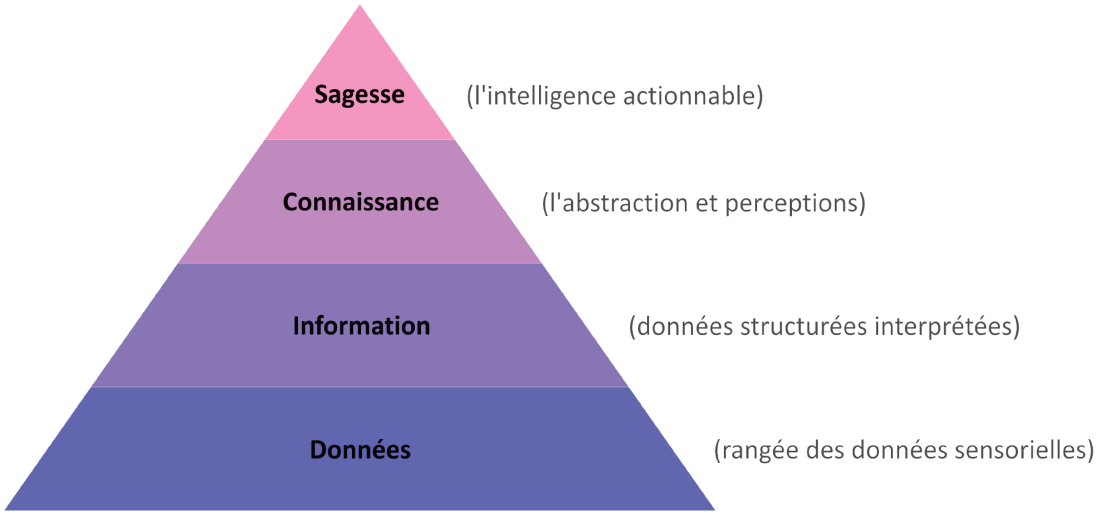

Figure 2.3: La pyramide de la sagesse DIKW.

2.3.2 Les notions principales : données, information, connaissances, concepts

Les travaux abordant les voies possibles de transformation des données en connaissances sont nombreux, ainsi que les essais pour présenter des définitions exhaustives de ces notions ce qui, à notre avis, n'est pas toujours possible. On trouve des exemples dans *Ackoff* (1989), *Shannon* (1948), *Popper* (1979), *Roche* (2005) et *Zins* (2007).

Dans la table 2.1, nous présentons l'ensemble des définitions synthétisées issues des références ci-dessus, en opposant le point de vue des sciences exactes, où ces notions sont des objets concrets avec des paramètres mesurables, et celui de la philosophie, où les définitions ont plutôt un caractère d'interprétations générales.

	En informatique et sciences exactes	**En philosophie et sciences sociales**
Données	Tous les types de signaux pouvant être perçus par l'homme ou par un système d'information.	La notion est floue ; elle est souvent interprétée comme le synonyme d'information ou de faits.
Information	La grandeur quantitative qui caractérise la réduction de l'ambiguïté sur le choix de l'état de l'objet parmi plusieurs variantes possibles.	La caractéristique fondamentale de l'être ; la notion axiomatique telle que la matière, l'énergie, l'espace, le temps ; elle ne peut pas être définie par le biais des autres catégories. Souvent l'information est confondue avec des faits ou des données, et parfois avec des connaissances.
Connaissance	Le résultat de l'accumulation de compétences et de savoir-faire à travers des artefacts multiples (selon K. Popper, *Popper* (1998)) ; il peut être régi, de façon fiable, par les systèmes d'informations.	Le cadre cognitif qui permet à l'homme d'utiliser l'information. L'information valide, en cohérence avec les autres vérités acceptées.
Concept	Correspond à la classe des objets dont les propriétés sont restreintes de façon explicite par l'imposition de contraintes. « Un concept correspond à une définition intentionnelle de la classe (ensemble) de ses référents », *Roche* (2005).	« L'élément de pensée portant sur une pluralité de choses distinctes répondant à une même loi », *Roche* (2005).

Table 2.1: Les définitions des notions générales.

La table 2.2 permet de visualiser le chevauchement des propriétés partagées par les notions de *données*, d'*information* et de *connaissance*, en s'appuyant sur les définitions présentées dans la table 2.1. Ainsi, les *données* peuvent être enregistrées ou « reconnues » par un système d'information (on dit qu'elles sont *reconnaissables*). L'*information* pertinente peut être distinguée du bruit : elle est *interprétable*. Enfin, la *connaissance*, dans le contexte d'un système d'information, suppose la capacité d'utiliser l'information, d'en

déduire de nouveaux faits non explicitement présents (en ce sens, on dit que la connaissance est *prédictive*).

	Reconnaissable	**Interprétable**	**Prédictive**
Données	×	×	
Information	×	×	
Connaissance		×	×

Table 2.2: Les propriétés permettant de distinguer la donnée, l'information et la connaissance.

2.4 Modèles de représentation des connaissances

Dans cette section, nous discuterons brièvement les approches de présentation formalisée des connaissances, parmi lesquelles la représentation sous forme d'ontologie. Les idées sur ce sujet ont évolué au sein de plusieurs courants théoriques de la psychologie, de la philosophie et de nombreuses branches de l'informatique (notamment l'intelligence artificielle), de la pédagogie et, plus récemment, de la linguistique sémantique. Les objectifs de la modélisation des connaissances peuvent être envisagés de différents points de vue :

— Pour élaborer des modèles calculables représentant les processus intellectuels et cognitifs des humains ;

— Pour améliorer la performance des systèmes d'information et pour programmer des tâches dites « intelligentes » ;

— Pour mieux expliquer les phénomènes linguistiques.

Nous nous intéressons ici aux modèles de représentation des connaissances permettant une construction efficace de l'ontologie d'un domaine.

En informatique, à la différence des définitions floues données aux connaissances par des philosophes (cf. la section 2.3.2), en informatique donc la notion de connaissance a une signification plus opérationnelle : c'est une multitude de règles, sémantiques et syntaxiques, explicitement interprétée d'une manière ou d'une autre *Bloch et Maître* (2002). Cette interprétation utilitaire des connaissances nous amène à distinguer entre plusieurs types des connaissances : faits, concepts et procédures *Paquette* (1996a). Pour cette raison, tous les modèles qui tentent de représenter des connaissances peuvent être groupés en deux catégories : ceux qui adoptent l'idée que les lois de l'esprit sont subordonnées à la logique pure ; et ceux qui comprennent l'intelligence comme la combinaison d'associations mentales correspondants aux objets ou aux situations du monde réel *Geeraerts* (2001), *Baader et al.* (2010).

Dès lors, les modèles de représentation des connaissances se fondent soit sur les schémas procéduraux, soit sur les schémas déclaratifs *Paquette* (1996a). Les schémas procéduraux prescrivent les règles récursives de génération de connaissances nouvelles, à partir d'un nombre limité d'axiomes sur les propriétés des entités. Dans les schémas déclaratifs, l'accent est mis sur l'attribution à toute entité de propriétés caractéristiques constatées dans des circonstances données.

Le défaut de la première approche est son incapacité à saisir la diversité des relations et des propriétés des objets matériels ou abstraits du monde réel, à cause de la rigueur de ses théorèmes. A l'opposé, le problème de la deuxième approche est sa faiblesse à garantir la non-contradiction des faits lors de leur description. De nos jours, le défi est de coupler les deux approches en préservant les avantages de chacune et en contournant ses inconvénients.

Les variantes des modèles traditionnels de la représentation des connaissances sont résumées dans la figure 2.4.

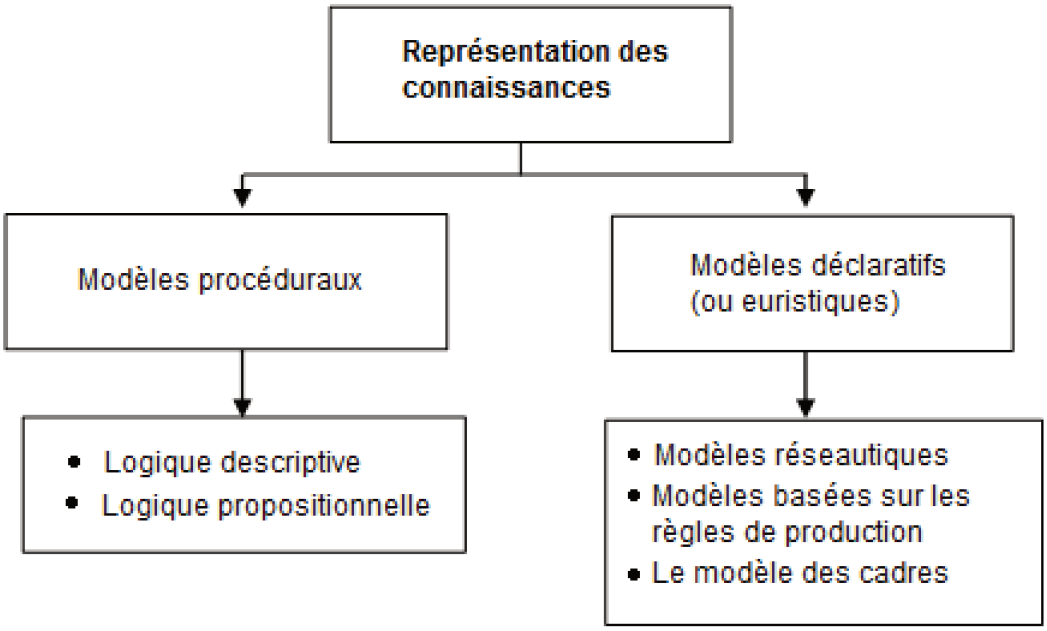

Figure 2.4: Les modèles de représentation des connaissances.

Notons que tous les modèles utilisent l'outil mathématique de la logique propositionnelle d'ordre zéro, qui a évolué vers la famille des algèbres de logique descriptive (DL), et qui permet de définir des propriétés et des relations complexes entre concepts, et de les manipuler *Baader et al.* (2010).

2.4.1 Modèles procéduraux

Les modèles procéduraux partent de l'idée que les calculs des prédicats permettent de capturer sans ambiguïté la sémantique des relations entre les objets. Ces modèles se construisent à l'aide des langages de programmation logique dont l'exemple le plus connu est Prolog. Ce langage utilise le principe d'encapsulation, mécanisme intégrant *les données* et *le code* qui les manipule.

Ces modèles procéduraux comprennent simultanément des *enregistrements* (semblables aux enregistrements dans les bases de données) et des *règles* permettant de les traiter. La combinaison de ces qualités permet l'accumulation des connaissances (données et règles), au cours du travail d'un programme Prolog.

2.4.2 Modèles déclaratifs

Les modèles déclaratifs se fondent sur l'hypothèse que le modèle d'un domaine demeure invariant quels que soient les objectifs de son utilisation. Pour cette raison le modèle comprend deux parties : les structures statiques descriptives, et le mécanisme de raisonnement qui les manipule. Cette approche permet de séparer les aspects syntaxiques et sémantiques de la connaissance représentée.

D'habitude, les modèles déclaratifs représentent une multitude d'affirmations : le modèle de domaine se réalise par la description syntaxique de son état, tandis que le raisonnement se fait par les procédures de recherche dans l'espace des états.

2.4.2.1 Modèle des cadres

Le concept de *frame sémantique*, en français de « cadre », réunit des idées issues de plusieurs grands champs - linguistique, philosophie, psychologie et informatique - pour modéliser les mécanismes de notre compréhension et la saisie du « sens des choses ». Ces idées ont été lancées dès le début des années cinquante. Parmi la pléiade de scientifiques qui ont été à son origine citons les noms de Ch. Fillmore, M. Minsky *Minsky* (1985), A. Newellde, R. Schank, *Schank et Rieger* (1985).

M. Minsky avait travaillé sur la détection des mécanismes de fonctionnement de la mémoire permettant de reconnaître le sens des informations qu'on perçoit ; cela est possible car, selon lui, la mémoire conserve les images composées faisant référence à des objets et circonstances réels. Pour M. Minsky, au cadre de toute nature correspond un minimum d'information structurée permettant d'identifier une certaine classe d'objets dans une situation stéréotypée. On distingue trois catégories de cadres selon le type d'information : les cadres-concepts, les scénarios et les schémas.

Une des propriétés importantes des « cadres », dans la théorie des réseaux sémantiques, est l'héritage des propriétés par les éléments du cadre (Frame Element, FE). La relation

principale par laquelle un élément du cadre évoque un cadre supérieur est la relation du type « ***kind-of*** ».

L'implémentation des systèmes basés sur les cadres a conduit au développement de langages spécifiques, appartenant à la famille des langages de représentation des connaissances. Ils sont présentés dans le chapitre 6.2. On peut mentionner plusieurs systèmes basés sur les cadres où des langages spécifiques ont été implémentés. Ce sont, par exemple, Frame Representation Language (FRL), *Rich* (1983) ; Knowledge Representation Language (KRL), *D.G. et Winograd* (1985) ; KL-ONE, *Woods et Schmolze* (1992) ; LOOPS, *Bobrow et al.* (1983).

Le principe des cadres s'est avéré efficace pour l'élaboration des ressources lexicales liées aux méthodes de fouille de textes. Nous décrivons plusieurs ressources dans la section 2.6.

2.4.2.2 Modèles en réseau

À la base de tout réseau sémantique, il y a un graphe orienté, réalisé de manière déclarative comme présenté dans la figure 2.5. Les sommets (ou nœuds) du graphe correspondent aux objets [3], tandis que les arcs présentent les relations qui les lient. Dans *Sowa* (1991), l'auteur décrit six types de réseaux où la sémantique des relations est désignée par des outils de logique algébrique. La plupart des notations qu'on utilise aujourd'hui ont été introduites en 1909, il y a plus d'un siècle, par Charles Sanders Peirce. Le premier système d'information fondé sur un réseau sémantique a été réalisé par Richard Richens en 1956 au Centre de langues de Cambridge, dans le cadre du projet sur la traduction automatique.

Bien qu'en théorie le nombre des relations pouvant être définies dans un réseau soit illimité, en pratique on n'utilise qu'un petit nombre de relations (les plus répandues) telles que la relation hiérarchique du type ***« is-a »*** ou ***« kind-of »***, la relation d'implication, la relation du type ***« part-of »*** entre les objets, entre les ensembles d'objets et entre les différentes parties des objets. Une classification plus détaillée des relation sémantiques est présentée dans la section 4.4.3 du chapitre 4.

2.4.2.3 Modèle basé sur des règles de production (Les systèmes experts)

Les systèmes experts utilisent souvent des règles de production pour représenter les connaissances d'un domaine. Ils ont connu un grand essor entre les années soixante-dix et le début des années quatre-vingt-dix. Un des exemples de systèmes basés sur les règles de production est le MYCIN ; système conçu pour l'aide au diagnostic des maladies infectieuses et des maladies du sang *Shortliffe* (1976).

3. Bien qu'à l'époque on utilisait le mot « concept » sans lui donner la définition qui lui est attribuée aujourd'hui en ingénierie des ontologies.

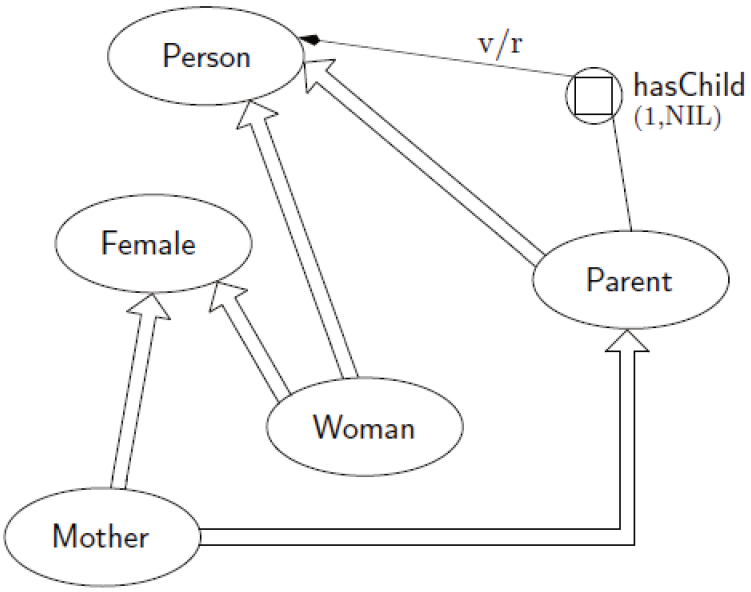

Figure 2.5: Exemple de présentation des connaissances en réseau.

Ces modèles sont basés sur un système de règles qui représentent les connaissances sous forme d'un ensemble de propositions de logique formelle d'ordre zéro, ou logique des prédicats du premier ordre.

Toute règle de production contient une partie conditionnelle et une partie d'action ; sa forme agrégée peut être écrite comme suit :

$$SI \textit{ condition } ALORS \textit{ action}$$

Un système de règles de production est constitué de trois éléments :

— *La base de faits* qui sont les instances des objets, à travers lesquels le domaine d'application peut être représenté.

— *La base de connaissances*, c'est à dire l'ensemble des règles qui lient les faits. La partie *condition* de chaque règle peut être la conjonction de plusieurs expressions logiques ; la partie d'*action* décrit les instructions à exécuter lorsque la règle s'applique sur la base de faits.

— *Le moteur d'inférence* qui exécute les règles.

Par exemple, Jess est un moteur d'inférence (codé par un ensemble de scripts en Java) qui permet de créer des applications interactives [4].

L'inconvénient du modèle basé sur les règles est que, plus elles sont nombreuses, plus elles risquent de se contredire. L'augmentation des contradictions peut être limitée par l'introduction de mécanismes d'exceptions et de retours. Les exceptions sont des règles

4. http ://herzberg.ca.sandia.gov/

strictement spécialisées. Quand le système rencontre une exception, la règle principale n'est pas appliquée. Le mécanisme des retours signifie que la raisonnement logique peut se poursuivre, même si on rencontre une contradiction à une certaine étape ; mais, pour cela, il est nécessaire de rejeter une des propositions acceptées et de faire un retour à l'état précédent.

Le récapitulatif de la comparaison des différents modèles de présentation de la connaissance est présenté dans le tableau 2.3.

Modèle	***Forme de la présentation des connaissances***	***Les opérations principales***	***Les avantages***	***Les inconvénients***
Modèle logique (ou procédural)	ensemble de formules	raisonnement logique (preuve de théorème)	combinaison de la modularité des connaissances et possibilité d'expliquer leurs propriétés ; facilité de détection des contradictions dans les données et de contrôle d'intégrité	difficulté pour structurer les connaissances ; souvent les formules sont très longues et difficiles à lire ; lenteur de traitement
Modèle des cadres	l'ensemble de cadres-concepts et d'exemples	recherche d'un cadre ou d'un slot et correction	les connaissances sont bien structurées, claires pour l'utilisateur ; pas de restrictions sur l'ordre de traitement de l'information ; les cadres sont autonomes pour le traitement de l'information	le temps pour effectuer des opérations augmente si les cadres sont nombreux ; les relations entre les cadres sont compliquées pour les connaissances complexes ; s'il faut travailler avec des problèmes complexes, il existe des relations complexes entre les cadres
Réseau sémantique	Un système de réseaux ou un seul réseau commun	recherche d'information par échantillon, remplacement et copie d'information	clarté ; structure compréhensible par les utilisateurs ; simplicité relative de réalisation technique ; applicabilité de la théorie des graphes pour le traitement et l'analyse	le temps de calcul augmente considérablement lorsque le réseau devient grand ; la structure du réseau change en fonction de l'ordre de sa construction

Modèle basé sur les règles de production	système de règles de production	inférence i.e. stratégie de choix des règles	présentation simple et facile à comprendre ; la structure modulaire des règles permet une correction relativement facile	la simplicité mène à l'appauvrissement de la description ; efficacité faible pour la résolution de problèmes complexes ; difficulté de contrôle des contradictions et de la validité de l'ontologie ; pré-requis professionnels élevés pour les ingénieurs des connaissances

Table 2.3: Comparaison des modèles de présentation des connaissances.

2.5 Modélisation de la langue naturelle

Dans les sections suivantes, nous allons décrire les modèles linguistiques selon leur aptitude à capturer le sens des propositions, écrites ou orales, et donc à être utiles pour mettre en œuvre une ontologie. Des essais systématiques de modélisation de différentes langues ont été entrepris au début du vingtième siècle. On peut dire que les idées scientifiques ont évolué d'une vision plutôt structurale de la langue, vers une approche holistique qui domine aujourd'hui et qui voit la langue comme un système dynamique dont les propriétés ne se réduisent ni à la structure morphosyntaxique de ses éléments, ni aux seules significations des mots composant les énoncés : la langue est un système dynamique reflétant les mécanismes cognitifs humain. On accepte maintenant que la langue ne peut pas être décrite par une approche purement algorithmique et qu'il faut prendre en compte le fonctionnement de nos structures cérébrales, de notre capacité à jouer avec la langue ; ainsi, pour améliorer le traitement automatique de la langue naturelle, il faudrait accumuler une grande masse d'observations enregistrées par des psychologues, les spécialistes en neurolinguistique etc., *Geeraerts* (2006), *Geeraerts* (2001).

Une chronologie résumée des théories linguistiques, selon Dirk Geeraertsm est visualisée sur la figure 2.6.

2.5.1 La linguistique structurale et le modèle fonctionnel

La linguistique structurale a été développée dans les années 20 par le Cercle linguistique de Prague (ou École de Prague), sous l'influence des idées de Ferdinand de Saussure

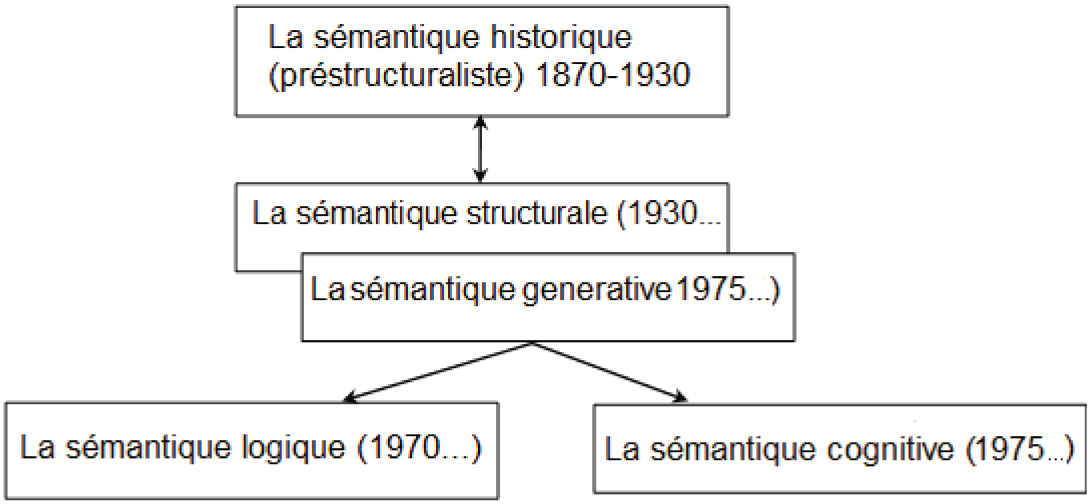

Figure 2.6: Les grandes périodes de la sémantique lexicale.

formulées dans son « Cours de linguistique générale » paru en 1916. D'autres auteurs ont marqué cette époque : Roman Jakobson, Vilém Mathesius, Leontij Kopeckij, Nikolaï Troubetzkoï et beaucoup d'autres, cités par *Léon et al.* (2009). L'aspect révolutionnaire de la pensée saussurienne est de distinguer d'une part la langue comme système de signes et, d'autre part, les idées ou concepts exprimés à l'aide des structures composés de ces signes ; auparavant la linguistique ne distinguait pas les dénominations et les notions dont elles portent le sens.

À partir des années 1920, on considère que les traits essentiels de la langue se réduisent à ses propriétés fonctionnelles, lesquelles se manifestent à travers des structures morphosyntaxiques : le dégagement des «…propriétés structurelles de l'objet en tant que ses propriétés essentielles permet de créer la théorie d'une structure donnée, également applicable à des objets de toute autre nature uniquement parce qu'à leur fond est la même structure. » *Apresjan* (1973), *Léon et al.* (2009), *Bloomfield* (1933). Pour les adeptes de l'approche fonctionnelle, la modélisation de la langue se fait selon les principes des sciences exactes :

— Le modèle constitue une approximation fonctionnelle de l'objet, qui se résume à sa structure.
— Le modèle aboutit à la réduction d'une partie des propriétés de l'objet. Mais cette simplification sert à considérer les cas simples les plus généraux ; il reste possible de les complexifier si nécessaire.

2.5.2 Modèle génératif

À partir des années soixante l'attention des linguistes s'est déplacé du modèle structural vers le modèle génératif proposé par N. Chomsky. La première version de son œuvre

« Les structures syntaxiques » est paru en 1957, mais Chomsky a constamment perfectionné son approche.

N. Chomsky a proposé une théorie déductive de langue, i.e. la possibilité d'obtenir et d'expliquer toute la diversité de phénomènes langagiers à l'aide de règles de transformations. Il a défini la langue comme « *…un ensemble, fini ou infini, de phrases, chacune finie en longueur et construite par concaténation à partir d'un ensemble fini d'éléments…* ».

Son principe de la finitude se formule comme suit : il y a un ensemble fini de règles susceptibles d'engendrer (d'où le nom de Grammaire Générative) l'ensemble infini des éléments (dans notre cas : des mots et des phrases) dit corrects. Ici, on entend par « ensemble fini de règles » à la fois les règles morphologiques de formation des mots et les règles syntaxiques de combinaison des mots dans les phrases.

Le deuxième principe est que chaque niveau structural a ses propres lois qui ne s'appliquent qu'à ce niveau-là. Par exemple, on peut changer l'ordre des mots dans une phrase sans perdre son sens [5], mais on ne peut pas changer l'ordre des syllabes à l'intérieur d'un mot. Par conséquent les méthodes d'analyses doivent être adaptées selon le niveau de granularité où l'on se place.

La vision de la langue comme le système génératif a amené Chomsky au modèle dynamique de langue où il a postulé l'existence d'un nombre limité de structures sous-jacentes, universelles et communes à toutes les langues ; ceci correspond aux schémas abstraits des énoncés corrects et des structures superficielles qui les réalisent en pratique, ces derniers étant différents selon les langues.

Bien que le modèle génératif se soit avéré très efficace, particulièrement pour l'analyse syntaxique des propositions, il lui manque le contenu sémantique parce que ses règles permettent d'engendrer des structures autonomes syntaxiques qui peuvent être vides de sens. Ceci est illustré par l'extrait célèbre d'« Alice au Pays des Merveilles » de L. Carroll (traduction d'Henri Parisot) :

> *« Il était grilheure ; les slictueux toves*
> *Sur l'alloinde gyraient et vriblaient ;*
> *Tout flivoreux étaient les borogoves ;*
> *Les vergons fourgus bourniflaient. »*

Bien que le quatrain résonne correctement en français, il n'y a pas de vrais mots français sauf quelques éléments auxiliaires.

Néanmoins les analyseurs syntaxiques, adaptés à différentes langues, sont basés sur la réversibilité des règles récursives de réduction et des règles de déploiement adoptées dans le modèle génératif.

5. bien que cela dépende des langues

2.5.3 Sémantique vs syntaxe

La syntaxe examine la fonction et la disposition des mots et des propositions dans la phrase [6], tandis que la sémantique s'intéresse aux significations des expressions élémentaires, et à leurs règles de combinaisons produisant des expressions complexes. Au début de l'histoire de la linguistique, la sémantique et la syntaxe s'opposaient ; de nos jours, on cherche à les coupler : on considère qu'à chaque expression élémentaire correspond une (ou plusieurs) signification(s) et que chaque règle syntaxique par laquelle se combinent deux éléments exprime à la fois la sémantique et la combinaison de leur significations. Ainsi la syntaxe et la sémantique sont fortement liées : en partant des significations des expressions élémentaires et des règles de leur combinaison, on arrive à saisir la signification agrégée d'une expression complexe.

2.5.4 Les verbes dans les modèles linguistiques

Puisque l'univocité des significations des concepts d'une ontologie est son principe essentiel, la solution du problème d'ambiguïté des mots est importante dans l'apprentissage d'ontologies. Dans nos expérimentations, nous utilisons les verbes caractéristiques afin de former des contextes univoques lors de l'extraction des termes.

Mais les verbes sont aussi ambigus. Leur désambiguïsation est possible grâce à l'hypothèse d'un lien fort entre les propriétés syntaxiques d'un verbe et sa sémantique : à la similitude des propriétés syntaxiques des verbes correspond, en règle générale, la similitude de leurs caractéristiques sémantiques ; inversement, aux différences de propriétés syntaxiques correspondent des caractéristiques sémantiques distinctes *Apresjan* (1973).

On trouve des idées proches chez L. Tesnière dans sa conception du modèle de dépendance des énoncés, *Tesnière* (1959). Pour lui, la structure linéaire d'une phrase cache les relations subordonnantes de ses constituants, où le rôle régissant est essentiellement joué par le verbe. Tesnière a proposé un schéma visualisant les liens entre les éléments d'un énoncé, avec le verbe en sommet (*stemma*). Sa notion de valence sémantique du verbe comprend le nombre d'arguments (ou d'actants) qui peuvent être régis par un verbe ; cela rapproche le (*stemma*) de L. Tesnière du (*cadre sémantique*) de Ch. Fillmore.

Cette interdépendance est un phénomène universel qui se produit dans toutes les langues normalisées ; l'effet d'analogie mène à l'homogénéisation et à l'unification des propriétés syntaxiques des mots proches sémantiquement. Ces idées sont à la base de la grammaire générative.

La mesure de similitude (et de différence) du sens des verbes selon la similitude et la différence de leurs propriétés syntaxiques peut concerner soit l'ensemble des emplois d'un même verbe (dans ses différents sens), soit la multitude des sens différents qu'expriment tous les verbes du langage. Dans le premier cas, il s'agit de la différentiation des

6. définition issue du Petit Larousse

significations d'un verbe donné ; dans le deuxième cas, il s'agit des classes formées par des verbes proches sémantiquement.

Dans nos recherches, nous nous intéressons à la deuxième option car la définition des groupes de prédicats pertinents pour un domaine est, à notre avis, la démarche clé pour la conceptualisation d'un domaine. On entend ici que le rôle de prédicat est le plus souvent joué par le verbe, *Huddleston et Pullum* (2005).

Afin de trouver les verbes caractéristiques dans le corpus, c'est à dire les verbes qui peuvent servir à indiquer la présence des candidats-termes dans un énoncé, il faut résoudre plusieurs tâches : désambigüer le sens des verbes sélectionnés (*Brown et al.* (2011), *Wagner et al.* (2009)), spécifier le schéma de catégorisation (*Heid* (2007), *Shustova S.* (2015)), définir les rôles sémantiques des arguments (*Gildea et Jurafsky* (2002)).

Comme on l'a dit, c'est Tesnière qui, pour la première fois, a proposé de placer le verbe au sommet (qu'il nomme *stemma*) du schéma visualisant les liens entre les éléments d'un énoncé. Cette approche est utilisée par des analyseurs syntaxiques actuels pour la construction des arbres de dépendances.

Les deux premières tâches sont étroitement liées. Le travail se complique car il n'y a pas, parmi les linguistes, d'accord complet sur le schéma de description des verbes français *Mathieu-Colas* (2006). Par ailleurs, en langue russe, G. Zolotova (*Zolotova* (2011)) a proposé un principe de généralisation des structures syntaxiques minimales qui permet de distinguer le sens des phrases en désambiguïsant le sens des verbes à partir de leur schéma syntaxique. L'auteur définit une unité syntaxique minimale de sens élémentaire, indivisible, dont le fonctionnement est nécessaire et suffisant à la construction des structures plus complexes sans ambiguïté syntaxique.

Les méthodes automatiques de traitement des verbes sont nombreuses et largement utilisées ; nous pensons que l'utilisation des ressources linguistiques est indispensable pour résoudre le problème de l'ambiguïté. La section suivante présente les principaux outils linguistiques disponibles.

2.6 Les ressources lexicales

Trois types des ressources lexicales sont de plus en plus utilisées dans les méthodes de TALN proposant des solutions au problème de l'ambiguïté des mots : les dictionnaires, les corpus annotés et les bases d'information lexicales. Dans cette section, nous allons décrire brièvement les particularités de chaque type d'outil et présenter plusieurs projets à la fois typiques et signifiants : FrameNet (cf. la section 2.6.1), WordNet (cf. la section 2.6.2), RuTes (cf. la section 2.6.3), BabelNet (cf. la section 2.6.4), VerbNet et VerbOcean (cf. la section 2.6.5).

2.6.1 FrameNet

FrameNet est une ressource linguistique en anglais, lisible par la machine et par l'homme. Sa création a débuté en 1997 dans le cadre d'un grand projet de lexique sémantique, créé sous la conduite de Ch. Fillmore *Fillmore et Atkins* (1992) et mettant en œuvre sa conception des cadres sémantiques (cf. la section 2.4.2.1). FrameNet vise à la description de la compatibilité sémantique et syntaxique des mots en fonction de leurs valences qui, à leur tour, dépendent du sens contextuel du mot.

Le contenu de FrameNet ne cesse d'augmenter. On peut consulter sa statistique courante sur le site de Berkeley [7]. Nous présentons plusieurs chiffres récents afin de montrer son ampleur : il y a plus de 1 100 cadres lexicaux hiérarchisés. L'index contient plus de 13 400 unités lexicales (LU) illustrées par des exemples textuels annotés. Au total, il y a plus de 28 000 sets de textes complets annotés et plus de 227 000 de sets textuels.

FrameNet peut être vu comme un exemple d'ontologie linguistique de situations standardisées dont les concepts sont réalisés sous forme de cadres liés par des relations hiérarchiques. FrameNet emploie huit types de relation entre les cadres [8] qui sont réunis en trois groupes : les relations de généralisation, les relations de structure d'événement et les relations « systématiques » *Fillmore et Baker* (2010).

Les relations les plus fréquentes entre les cadres sont :

— La relation du type ***is–a***, qui est la plus stricte ; elle est établie dans le cas où chaque élément d'un cadre parental (FE, frame element) est lié à un élément correspondant d'un cadre subordonné.

— La relation ***Using*** indique le cas où le cadre subordonné utilise le cadre parental comme contexte, par exemple le cadre VITESSE évoque le cadre MOUVEMENT. Dans ce cas il n'est pas obligatoire que tous les FE du cadre parental soient liés avec les éléments du cadre subordonné ;

— La relation ***Subframe*** décrit le cadre subordonné comme sous-événement d'un événement plus large, par exemple, pour le cadre PROCÈS CRIMINEL, les cadres subordonnés sont ARRESTATION, COUR DE JUSTICE, JUGEMENT.

— La relation ***Perspective on*** signifie que le cadre subordonné nuance le point de vue général, non orienté, du cadre parent. Par exemple, les cadres EMBAUCHER et OBTENIR UN TRAVAIL sont des sous-cadres du cadre DATE D'ENTRÉE DANS L'EMPLOI (EMPLOYMENT START) des points de vue respectifs de l'employeur et du travailleur.

Les autres relations utilisées dans FrameNet sont l'antériorité et la causalité. Notons que ces relations sont réalisées en double voix, active et passive, où les rôles des acteurs changent de places. Il existe actuellement des extensions du projet FrameNet dans sept langues mais, pour l'instant, pas pour le français.

7. https ://framenet2.icsi.berkeley.edu/fnReports/data/projectStatus.html

8. Les formes active et passive de la même relation sont distinguées dans FrameNet ; par exemple, la relation qui évoque l'utilisation est réalisée sous forme *Uses* et *Is Used by*

2.6.1.1 Cadre de la notion RISQUE dans FrameNet

Pour nous, le cadre sémantique de RISQUE, tel qu'il est réalisé dans FrameNet, est particulièrement intéressant à cause de la relation étroite existant entre la notion de *risque* et notre domaine d'application qui est *la sécurité radiologique*.

Le schéma initial de notre modèle a été construit en nous appuyant sur le cadre de RISQUE ; c'est pourquoi nous présentons ce projet en détail.

Les concepteurs incluent dans cette notion tous les mots dont la description sémantique fait référence à *la possibilité d'avoir un résultat indésirable*. Il s'agit des cas où l'on voudrait éviter certaines conséquences possibles d'un événement. Les mots qui rentrent dans la famille sont, par exemple, *danger, incertitude, menace, péril*. Le paradigme de RISQUE comprend non seulement des substantifs, mais également des adjectifs, des adverbes, des verbes et des phrases conventionnelles.

Le cadre de *RISQUE* implique nécessairement deux notions, à savoir la Chance (l'aléa) (le hasard) (Occasion), et le Harm (Dommage). Pour le visualiser, les auteurs utilisent un graphe orienté, emprunté à la théorie de la décision. Un des nœuds est carré, il correspond au Choix possible d'une situation. L'autre nœud est un cercle, il correspond à la chance, c'est à dire au résultat de la prise de décision ; chaque nœud peut avoir une ou plusieurs sorties sous forme d'arc (paths). La présence d'un arc liant deux nœuds signifie que la probabilité d'arriver vers tel ou tel état est supérieur à zéro. Ceci est illustré sur la figure 2.7 où A , V et VO correspondent, respectivement, à l'Acteur, à la Victime et à l'Objet évalué ; et D, H, G correspondent à l'Action, au dommage-Harm, et au Gain.

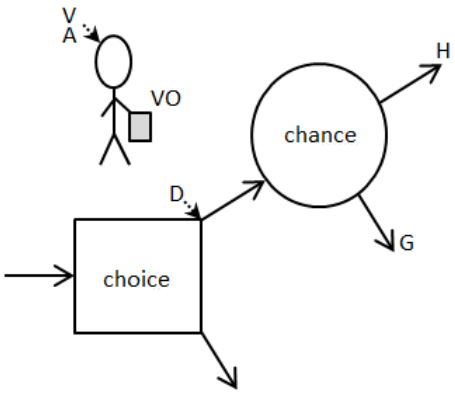

Figure 2.7: Visualisation des principales catégories du cadre de RISQUE.

Notons que dans FrameNet, on trouve également plusieurs cadres concernant le risque ; les éléments de noyau ne sont pas les mêmes bien qu'ils aient des liens, notamment Run_risk, Being_at_risk, Risky_situation etc. Dans FrameNet, les cadres relatifs au risque sont liés par la relation *Is Used by* avec le cadre PROTECTING qui, lui aussi, renvoie à des notions proches de celles utilisées dans le domaine de la sécurité radiologique.

Dans *Fillmore et Atkins* (1992), les auteurs présentent les résultats de recherches sur la lexicalisation de la notion de risque où le mot « risque » joue un rôle prédicatif, soit en tant que le substantif, soit en tant que verbe. Au final, ils proposent 21 patrons syntaxiques permettant détecter les actants d'une situation liée à un risque quelconque.

La correspondance entre les composants de la notion *risque* proposés dans *Fillmore et Atkins* (1992) et ceux qui sont utilisés pour créer le modèle du domaine de la sécurité radiologique est présentée dans la table 2.4.

Dans FrameNet	***Dans notre modèle***
Possibilité (Chance)	–
Dommage (Harm)	Dommage
Victime (Victim)	Population, Personnel
L'acteur (Actor)	Personnel
Motivation	Controle, Protection
L'intention	–
Gain	Sécurité, Sûreté
L'objet évalué (Valued Object)	Source
Bénéficiaire	–
Situation à risque(Risky situatiuon)	Facteurs de risque
L'act, action (Deed)	Accident

Table 2.4: La correspondance entre le cadre RISQUE et notre modèle.

Un fragment du modèle de domaine de la sécurité radiologique que nous avons placé à la base de notre ontologie de noyau est présenté dans la Fig. 2.8.

2.6.2 WordNet

Le thésaurus anglais WordNet *Miller* (1995), *Fellbaum* (1998), *Loukachevitch* (2011) est apparu sur l'Internet en 1995 mais son développement a été lancé à l'Université de Princeton dès 1984 sous la direction du psycholinguiste George Miller. Sa version 3.0 comprend environ 155 000 lexèmes avec les exemples organisées en 117 000 ensembles de synonymes dit *synsets*, pour la langue anglaise. Chaque synset peut être envisagé comme la présentation lexicalisée d'une notion générale, ou concept.

WordNet se construit à partir de la relation de synonymie. Pour les concepteurs du thésaurus, deux expressions sont synonymes si le remplacement de l'une par l'autre ne change pas la valeur de vérité de la proposition. Pour autant, la substituabilité des synonymes dans tous les contextes n'est pas nécessaire : il suffit que les synonymes soient remplaçables dans certains contextes. Cela permet d'admettre qu'un même lexème puisse être associé à plusieurs synsets, ce qui correspond à la flexibilité des langues naturelles. Il y a actuellement plus de 200 000 paires de *lexème – sens* dans WordNet.

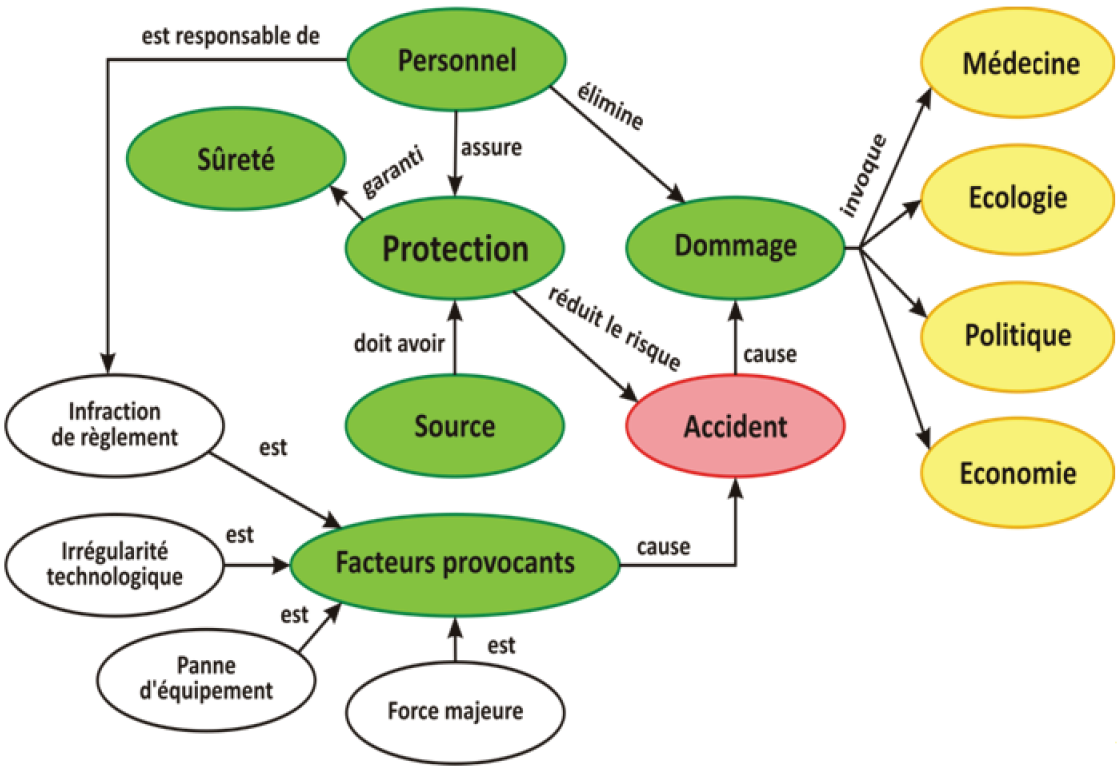

Figure 2.8: Fragment du modèle de domaine.

La définition de la synonymie par la substituabilité a nécessité la division de WordNet selon les parties de discours : les substantifs, les adjectifs, les verbes, les adverbes. La structure de description de chaque partie du discours est différente des autres. Les substantifs sont organisés en système hiérarchique où les propriétés des niveaux supérieurs sont héritées par des niveaux inférieurs. La hiérarchie des noms se concrétise sous forme des relations de trois types : l'*hyperonymie – hyponymie* (soit *is-a*), l'*antonymie*, et la *méronymie – holonymie*(soit *part-of*).

Dans WordNet, les verbes sont distingués les uns des autres selon la notion de champ sémantique (cf. 2.5.3). Il y a trois catégories générales de verbes : les verbes désignant les actions, les verbes des événements et les verbes désignant des états. Le premier groupe de verbes est divisé par 14 champs sémantiques parmi lesquels il y a les verbes de mouvement, de changement, de possession etc. Trois types de relation sont établies entre les verbes dans WordNet : l'implication, la troponymie [9], et la relation de causalité. Néanmoins les auteurs reconnaissent qu'il n'y a pas de délimitation stricte entre les classes de verbes.

9. La relation troponimique correspond au patron : «*verbe-2 décrit de manière plus précise l'action de verbe-1* »

À partir de la version WordNet 2.0, on a introduit les relations entre les synsets qui ont la même racine, i.e. liés sémantiquement, mais qui appartiennent à des parties différentes du discours. Cette option a été introduite pour rendre le modèle de WordNet plus universel et moins dépendant des spécificités des différentes langues.

2.6.3 PyTez (RuTez)

Le thésaurus russe PyTez est construit depuis 1994 au centre de recherches du traitement de l'information de l'Université Lomonosov de Moscou *Dobrov et Lukashevych* (2009). Actuellement il comprend plus de 51 500 concepts (notions générales), plus de 155 millions d'entrées lexicales (mots ou phrases) et plus de 200 000 relations entre les concepts. Au total, compte tenu de la hiérarchie des liens, le thésaurus comprend plus de 2 millions de relations entre les concepts. Les concepts du thésaurus sont également assortis d'entrées lexicales en anglais avec plus de 125 000 mots et phrases.

Dès le début, PyTeza été conçu pour automatiser les recherches d'information par les moteurs de recherche, notamment pour résoudre le problème de l'ambiguïté des requêtes des utilisateurs.

Le principe de son fonctionnement se formule comme suit : parmi les relations potentielles d'un concept, on peut s'appuyer sur les relations qui, dans toutes des observations des entités du concept (ou dans la grande majorité d'entre elles), ne disparaissent pas et ne changent pas. Par exemple, toute forêt se compose d'arbres.

Ce thésaurus utilise plusieurs types de relations. La première est la substitution possédant les propriétés de transitivité et de succession. La deuxième est la méronymie–holonymie ; elle s'applique aux composants d'un objet, et aussi à la description de ses propriétés intrinsèques, et au rôle que joue l'objet dans telle ou telle situation. Une contrainte importante est que chaque méronyme respecte toujours cette relation avec son concept-holonime, et pas avec les autres concepts. Cela permet de garantir la transitivité de la relation. Par exemple, s'il est vrai qu'une branche fait partie d'un arbre et qu'un arbre fait partie d'une forêt, on ne peut pas dire qu'une branche fasse partie d'une forêt.

Un autre type de relation dans le thésaurus PyTezest ce que les auteurs dénomment l'association asymétrique. Il s'agit du cas où un concept n'existerait pas sans un autre concept. Par exemple, le concept « sommet de l'état » requiert l'existence du concept « chef d'état ». Le dernier type de relations est l'association symétrique qui lie des concepts très proches mais que les auteurs n'ont pas osé réunir ensemble dans le même concept. PyTezest conçu pour des applications dans les domaines sociaux et politiques. Mais, selon les auteurs, il garantit une bonne précision des résultats dans la recherche d'information pour une vaste gamme de thèmes plus généraux.

2.6.4 BabelNet

Les concepteurs définissent BabelNet comme un « dictionnaire encyclopédique » fournissant des concepts et des entités-nommées lexicalisés qui sont liées par des relations sémantiques variées *Navigli et Ponzetto* (2012), et ceci en plusieurs langues.

BabelNet encode les connaissances sous forme d'un graphe orienté et labellisé $G = (V, E)$ où V est l'ensemble des nœuds, chacun correspondant à un concept [10] ou à une entité nommée, et $E \subseteq V \times R \times V$ est l'ensemble des arcs liant les paires de concepts. A chaque arc, correspond une relation sémantique qui peut être spécifiée en WordNet, comme, par exemple *is-a*, *part-of*, ou non spécifiée. Chaque nœud $v \subseteq V$ contient la lexicalisation du concept dans les langues différentes. Les auteurs appellent *Babel synsets* ces concepts multilingues. = L'intérêt particulier que présente BabelNet en *ontology learning* tient à ce que cette ressource est construite de façon automatique par l'alignement les synsets de WordNet avec les pages de Wikipedia qui jouent le rôle de contexte désambiguïsé.

La procédure de la construction de BabelNet se fait en plusieurs étapes :

1. La mise en relation de WordNet et de Wikipedia par acquisition automatique du « *mappage* » (de la correspondance) entre les sens définis pour les synsets dans WordNet et les pages de Wikipedia. L'objectif de cette étape est d'éviter la duplication des concepts tout en assurant leur complémentarité.
2. La lexicalisation multilingue des Babel-synsets par les versions des pages de Wikipedia en langues différentes afin d'avoir les liens inter-langues.
3. La validation des relations entre Babel-synsets et les pages multilingues de Wikipedia par le calcul de leur corrélation (avec le coefficient de Dice).

Un exemple illustratif de BabelNet est présenté sur la figure2.9 (tiré de *Navigli et Ponzetto* (2012)) : les arcs non-labellisés sont obtenus à partir de liens dans les Wikipages, tandis que ceux labellisés viennent de WordNet.

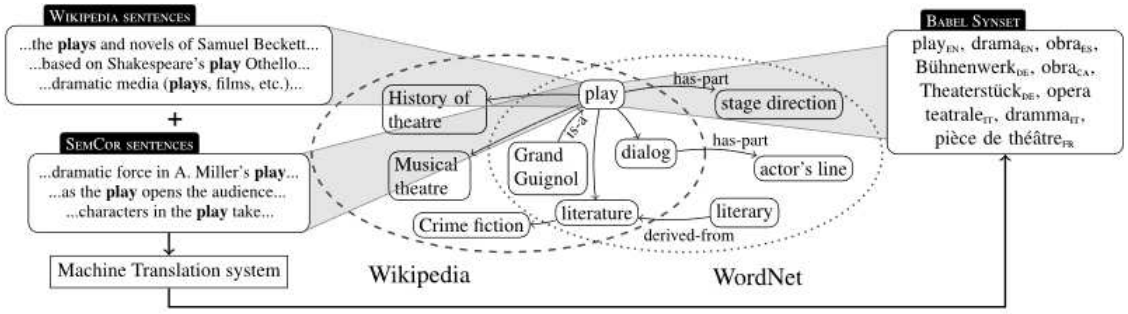

Figure 2.9: Un aperçu du fonctionnement de BabelNet.

10. autrement dit chaque nœud correspond à un synset de WordNet

2.6.5 VerbNet et VerbOcean

VerbNet [11] (*Kipper et al.* (2006), *Kipper et al.* (2007)) et VerbOcean [12] (*Chklovski et Pantel* (2005)) sont deux ressources lexicales dédiées uniquement aux verbes. Elles ont à peu près la même taille : chacune compte à peu près quatre mille lemmes de verbes.

VerbNet est une base lexicale de verbes compatibles avec WordNet et avec FrameNet et construite sur le principe du cadre. VerbeOcean est un réseau sémantique de relations entre les verbes qui sont proches sémantiquement.

VerbNet est construit à partir de la classification des verbes de B.Levin, mais sa classification est plus détaillée. Dans VerbNet les classes initiales de B. Levin ont été réorganisées en sous-classes complémentaires afin de garantir la cohérence sémantique et syntaxique des verbes-membres. Ensuite les classes de verbes peuvent être examinées à différents niveaux. L'hypothèse fondamentale de cette classification est que le comportement syntaxique des verbes, repéré à l'aide d'un schéma actanciel, dépend directement de leur sémantique. VerbNet associe la sémantique du verbe aux cadres syntaxiques en ajoutant des restrictions sémantiques aux éléments syntaxiques des cadres de chaque classe. Les verbes de la même classe doivent avoir le même comportement syntaxique.

Chaque classe de VerbNet contient :

— La liste des verbes-membres de la classe ;

— La liste des cadres syntaxiques définissant leur comportement syntaxique ;

— La liste des rôles thématiques qui conviennent à chaque cadre.

Et chaque cadre syntaxique contient :

— La description syntaxique, qui comprend l'indication des parties du discours et la structure de la proposition correspondant au cadre donné ;

— La liste des rôles thématiques correspondants aux éléments syntaxiques de ce cadre ;

— Les restrictions sémantiques sur les éléments syntaxiques du cadre ;

— Des exemples de propositions correspondant au cadre.

L'interface du VerbNet permet la recherche de classes, la consultation des listes de classes, des cadres et de leurs attributs, etc.

VerbNet propose 5 257 sens de verbes et 3 769 lemmes (divisés en 274 classes de premier niveau), 94 prédicats sémantiques, 23 rôles thématiques et 55 restrictions syntaxiques.

De son côté, VerbeOcean distingue cinq types de relations : la similitude, l'antonymie, l'inclusion, la force et la relation temporelle (qui signifie qu'un événement a eu lieu avant l'autre).

11. http ://verbs.colorado.edu/ mpalmer/projects/verbnet.html
12. http ://demo.patrickpantel.com/demos/verbocean/

2.7 Conclusion de chapitre 2

Actuellement, les ontologies sont devenues le principal moyen de représentation des connaissances, au sens d'un ensemble de faits constants ou dynamiques. Pour l'apprentissage efficace d'une ontologie on doit utiliser trois constituants :

— Un modèle de structuration de l'information permettant de transmettre, sans aberration, des données complexes entre les agents logiciels.

— Un modèle de langue expliquant les principes de génération des phrases correctes.

— Des ressources lexicales ou lexicographiques permettant de résoudre le problème de l'ambiguïté des mots.

Bien que l'apparition du Web Sémantique ait stimulé le développement des ontologies, les précédents modèles de connaissances ont tous contribué à améliorer l'apprentissage automatique des ontologies. Ainsi, les ontologies ont hérité d'avancées telles que la description de faits riches et extensibles, la possibilité du raisonnement logique et le contrôle de la cohérence des données.

La théorie de la représentation des discours et le modèle dépendanciel d'énoncés de Tesnière sont, de notre point de vue, parmi les approches les plus efficaces pour l'analyse des textes parce qu'ils permettent de s'appuyer sur les prédicats pour repérer dans un énoncé la présence de concepts et de leurs attributs.

On dispose maintenant, pour plusieurs langues, de nombreuses ressources lexicales ; mais leur organisation structurelle et leur taille varient beaucoup ; le choix d'une ressource pour la construction d'ontologie est donc une tâche non-triviale.

Chapitre 3

Apprentissage des ontologies : l'état de l'art général

3.1 Domaine de l'apprentissage des ontologies

L'apprentissage des ontologies, désigné ici par OL, pour Ontology Learning, fait partie de l'ingénierie des ontologies et plus largement de l'ingénierie des connaissances (Knowledge Engineering) et de la fouille de texte (Text Mining) *Maedche et Staab* (2000).

L'apprentissage des ontologies comprend deux étapes :

- L'élaboration de méthodes d'analyse des données textuelles et leur présentation formelle ;
- La mise en œuvre de ces méthodes dans le cadre d'environnements logiciels.

L'objectif principal de l'apprentissage des ontologies est d'introduire des méthodes d'apprentissage automatique tout au long du cycle de vie de l'ontologie afin de réduire le coût des ressources, humaines et matérielles, et de réduire les délais. L'apprentissage des ontologies est un domaine pluridisciplinaire qui nécessite les compétences des informaticiens, des linguistes, des managers et des spécialistes des domaines d'application. D'autres domaines voisins, traditionnels et tout récents, où l'apprentissage des ontologies peut, soit puiser, soit, à l'inverse, proposer ses méthodes sont : l'extraction des thématiques (Topic Extraction), *Aussenac-Gilles* (2005), la reconnaissance des entités nommés (Named Entity Recognizing), *Nadeau et Sekine* (2007), *Liu et al.* (2011) l'exploration automatique des opinions (Opinion Mining), *Pang et Lee* (2008), l'extraction d'événements (Event Extraction), *Tannier* (2014), les technologies de résumé automatique de textes (Automatic Text Summarization), *Nenkova et McKeown* (2012). Et cette liste ne cesse de s'allonger.

Dans ce chapitre, nous présentons l'état de l'art selon les axes suivants :

– Les approches méthodologiques sur l'organisation des travaux pour la construction d'une ontologie de domaine ;
– Les méthodes d'apprentissage pour chaque étape ;
– Le panorama des plateformes et des outils disponibles.

3.2 Panorama des outils

Comme il a été montré dans la section 2.2, les ontologies peuvent être employées utilement dans des domaines nombreux et variés. Cela montre l'importance d'élaborer des outils performants pour maintenir les ontologies elles-mêmes. Dans cette section nous allons analyser les produits, dit « *intelligents* » (*sweet tools*) liés aux technologies du Web Sémantique et à la gestion des connaissances, selon plusieurs critères, tels que la fonctionnalité, le langage d'implémentation, l'ouverture des sources.

La majeure partie des informations présentées ici vient du site d'AI3 [1]. Au total nous avons examiné les informations sur 1038 produits correspondant aux 58 catégories différentes définies par l'auteur du site. Quarante pour cent des outils utilisent directement les ontologies. Ce sont :

— Les Frameworks (environnements logiciels) permettant de construire les ontologies « de A à Z » ;
— Les modules d'extension pour des sous-tâches distinctes ;
— Les systèmes de traitement de la langue naturelle. Nous avons ajouté à cette liste les générateurs et les éditeurs de RDF et OWL parce que ces langages sont recommandés par W3C pour le codage des ontologies, ainsi que les éditeurs qui supportent SPARQL, un langage de requêtes permettant d'utiliser les ontologies réalisées dans le format OWL.

Une vue synthétique du nombre de produits utilisables pour la construction et la maintenance des ontologies est présentée dans le tableau 3.1. qui donne une vision détaillée des outils selon leur fonction.

Le tableau 3.2 présente la répartition des produits en fonction du langage de programmation et selon l'ouverture, ou non, de la source. On voit que les langages les plus répandus sont Java et Python. Le fait que la majorité des produits est librement disponible contribue au développement progressif du domaine : l'accès au code permet à chacun de modifier et de perfectionner les outils.

1. http ://www.mkbergman.com/sweet-tools/

Catégorie de produit	Quantité	Exemple de produit
Framework	20	text2Onto
Environnent	162	NeON
Traitement de langue naturelle	37	MALLET
Plugin	20	SKOSEd
SPARQL	117	ARQ
RDF/OWL (éditeur)	79	WebProtégé

Table 3.1: Catégories des outils liés au développement, maintenance, alignement et évaluation des ontologies (selon leur fonction).

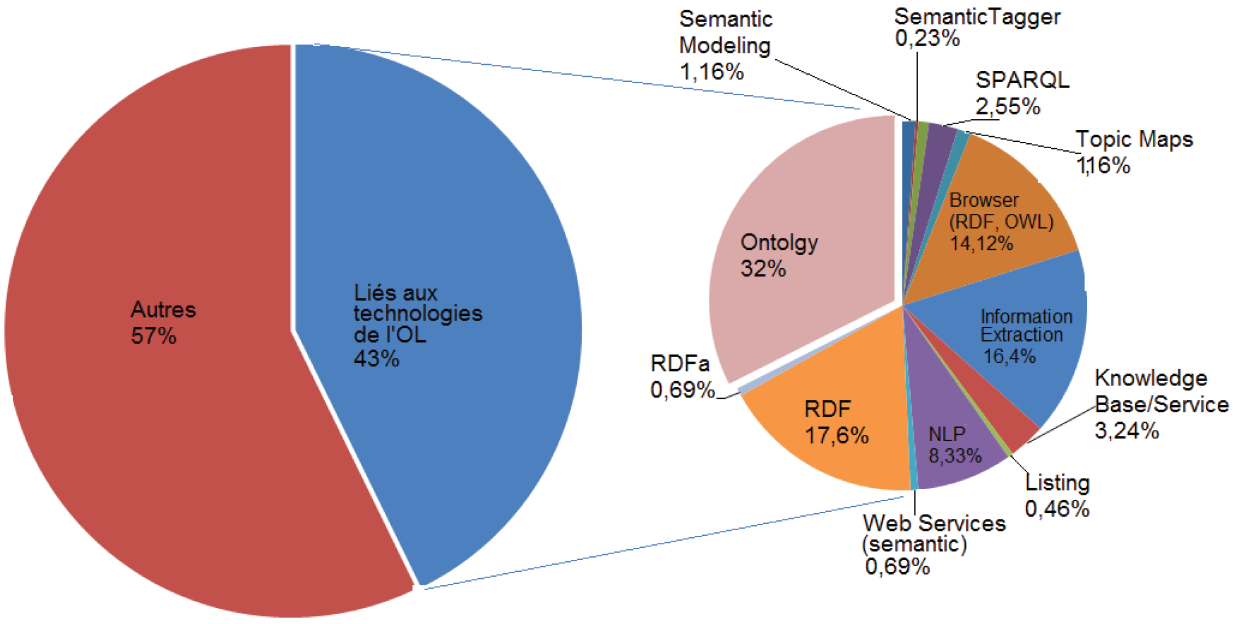

Figure 3.1: Distribution des outils employés dans l'ingénierie des ontologies.

3.3 Classification des ontologies

Il n'y a pas d'approche unifiée pour la classification des ontologies. Nous proposons deux critères qui nous semblent pertinents :

- Les objectifs de l'ontologie ;
- Le degré de la formalisation et des contraintes imposées sur les éléments de l'ontologie.

Langage de programmation	Open Source (OS)	N'est pas OS	En ligne	On ne sait pas s'il est OS	Total
Java/Java Script	501	50	28	57	636
C/C++,C#	51	7	2	1	61
Python	57	7	2	1	67
PHP	53		1	2	56
Ruby	31		3		34
Prolog	19	1		5	25
Perl	12		1		13
Autre	27	1	2	2	32

Table 3.2: Distribution des outils selon les langages de programmation et accessibilité au code.

3.3.1 Classification des ontologies selon leurs objectifs

Certains auteurs, comme *Lassila et McGuinness* (2001), *Dobrov et al.* (2008), distinguent les ontologies selon leur contenu ou le sujet abordé, mais il nous semble que ce sont plutôt les objectifs qui dictent les étapes ultérieures (l'analyse de faisabilité, les ressources, les méthodes, etc., cf. la figure 3.6). On distingue quatre catégories d'ontologies selon leurs objectifs, *Declerck et al.* (2012) :

1. Les ontologies de représentation ou meta-ontologies ;
2. Les ontologies de haut niveau ou, autre dénomination, les ontologies génériques ;
3. Les ontologies de domaine ;
4. Les ontologies d'application.

Les ontologies de différentes catégories forment une hiérarchie : les ontologies du bas de la figure 3.2 sont des spécifications des ontologies de niveaux supérieurs.

Les ontologies de représentation décrivent le domaine de représentation des connaissances et créent un langage pour la spécification d'autres ontologies de niveaux plus bas. Exemple : la description des catégories du langage OWL [2] par des moyens RDF/RDFS (cf. la figure 3.3).

Les ontologies de haut niveau doivent pouvoir être réutilisables dans de nombreuses applications, d'où cette dénomination.

2. http ://www.w3.org/2001/sw/RDFCore/Schema/20010618/#s2.1.1

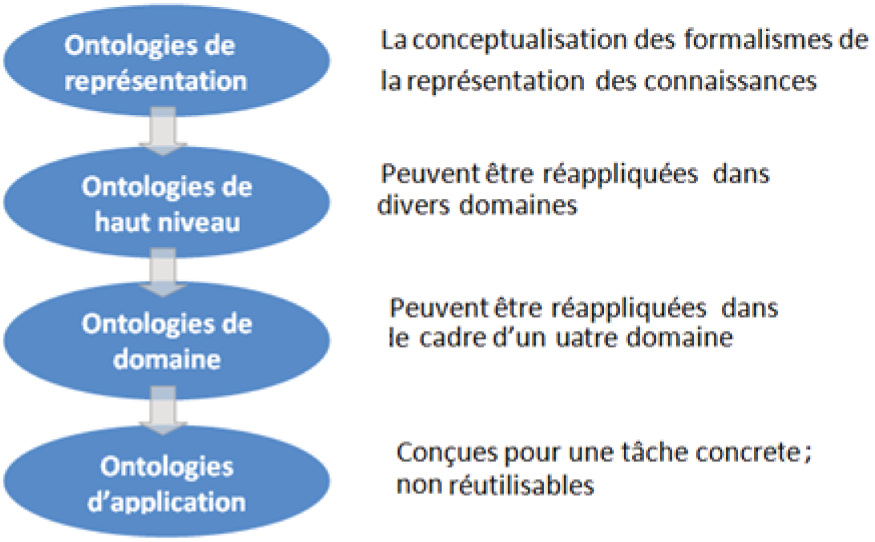

Figure 3.2: Classification des ontologies selon leurs objectifs.

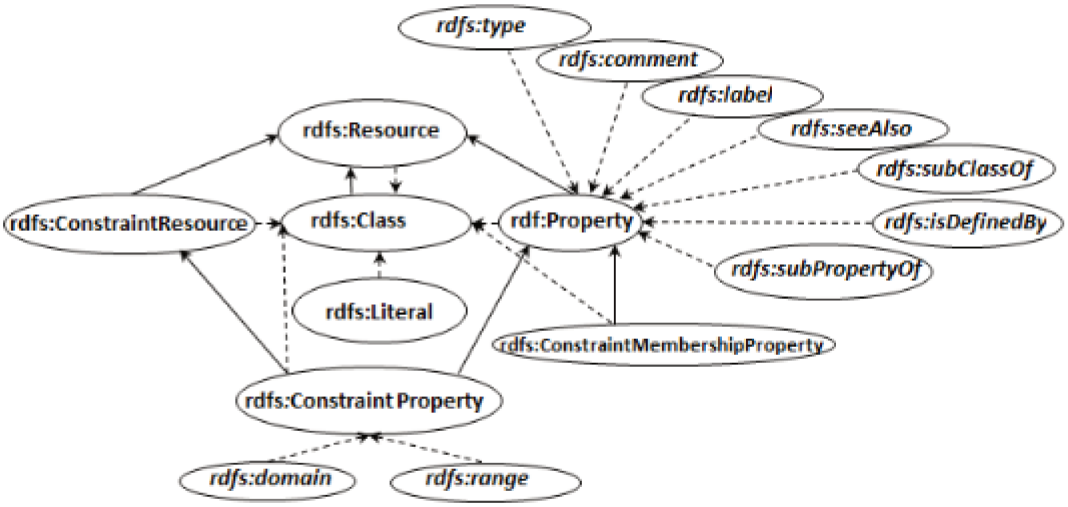

Figure 3.3: L'ontologie de représentation d'OWL.

Historiquement, les premiers projets ont tenté de réaliser des ontologies de haut niveau, conçues dans une optique philosophique : leur l'objectif était de réduire le lexique à un petit nombre de notions génériques, telles que la *Substance*, le *Phénomène*, le *Procès*, l'*Objet*, le*Rôle* etc.

Des exemples d'ontologies de haut niveau sont SOWA [3] (cf. figure 3.4 tiré du site de John F .Sowa http ://www.jfsowa.com/ontology/toplevel.htm.), OpenCyc [4] (qui est la ver-

3. http ://www.jfsowa.com/ontology/
4. http ://www.cyc.com/

sion ouverte du projet commercial Cyc lancé en 1984 par Doug Lenat), SUMO [5] (Suggested Upper Merged Ontology), (*Poli et al.* (2010), *Niles et Pease* (2003)). L'ontologie SUMO est alignée avec WordNet ; OpenCyc est alignée avec dbPedia, BabelNet, Wikipedia etc.

Un autre objectif est l'élaboration et le perfectionnement de langages formels pour présenter et manipuler des connaissances. À chaque projet correspond son propre langage : KIF pour SOWA, DALM pour SUMO, RDF pour OpenCyc, OWL (cf. le chapitre 6.2).

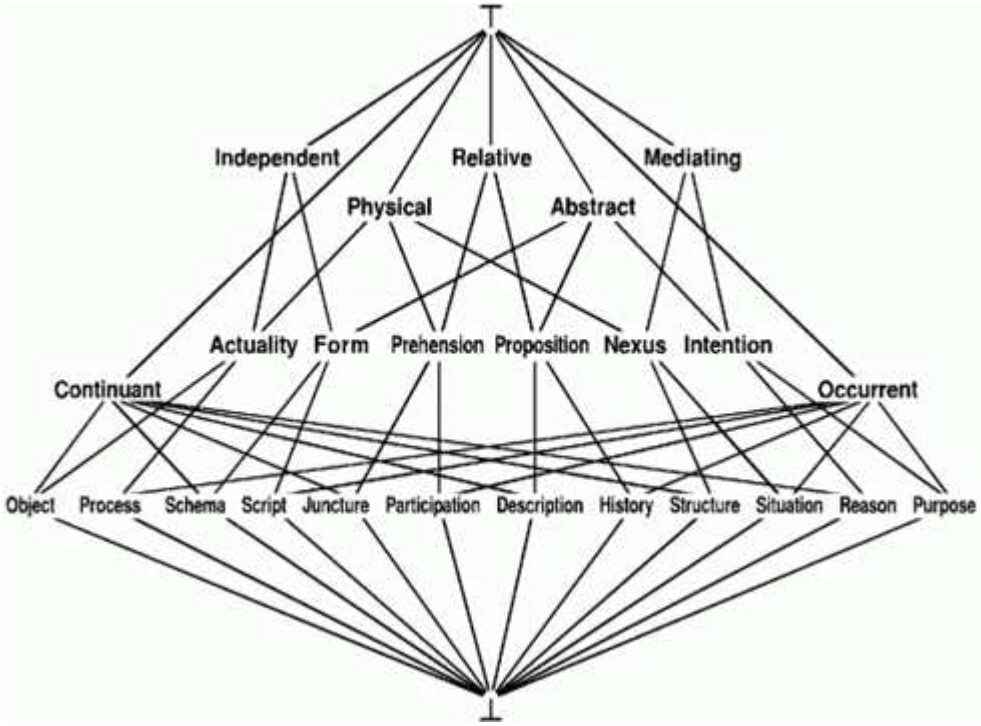

Figure 3.4: « Diamant » de SOWA.

Les ontologies de domaine ont pour but de modéliser les connaissances d'un domaine particulier. Cette catégorie est actuellement la plus demandée, *Simperl et al.* (2010). Pour donner un exemple, citons l'ontologie en médecine MENELAS, *Charlet et al.* (2009).

Les ontologies d'application sont conçues spécifiquement pour une tâche particulière ; elles ne sont en général pas réutilisables en dehors de ce cadre. Dans l'étude dont les résultats sont présentés dans *Simperl et al.* (2010), les auteurs montrent que la plupart des ontologies employées dans des entreprises se rapportent à cette catégorie (deuxième groupe forment les ontologies de domaine). La taille moyenne d'une ontologie est comprise entre 400 et 500 entités. Citons plusieurs exemples d'ontologies d'application : ***EFO*** (Experimental Factor Ontology) est utilisée afin de présenter les variables des données expérimentales d'expression génétique ; ***NIFSTD*** – ses modules couvrent les sous-domaines de la neuroscience, (*Malone et Parkinson* (2010)). ***TOVE*** [6]. (Toronto Virtual

5. http ://www.daml.org/ontologies/172
6. http ://www.eil.utoronto.ca/theory/enterprise-modelling/tove/

Enterprise Project) est un projet de création d'ontologies d'application modélisant des entreprises. Son objectif principal est l'élaboration d'ontologies donnant des réponses aux questions des utilisateurs sur la réingénierie des processus d'affaires.

3.3.2 Classification des ontologies par degré de formalisation

Dans *Uschold* (2008b), les auteurs proposent de distinguer les ontologies selon la spécification des constituants et des contraintes qui leurs sont imposées. Pour visualiser cette classification, on utilise le schéma connu sous le nom de « Le spectre des ontologies » reproduit figure 3.5. Cette classification a été réalisée pour la première fois dans le cadre d'AAAI-99, voir *McGuinness* (2003) ; la dernière mise à jour a été faite par McGuinness en 2007 [7].

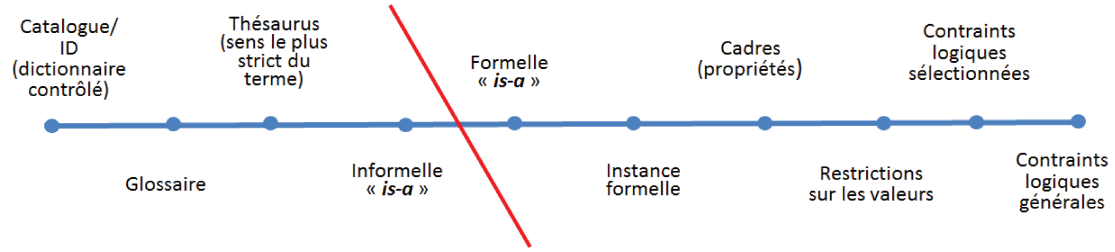

Figure 3.5: Spectre des ontologies.

Le premier point dans le spectre correspond au *dictionnaire contrôlé*, dont l'exemple le plus simple est le catalogue des identificateurs. Ici chaque terme est identifié par son ID et n'est associé à aucun contexte spécifique. Les catalogues donnent une interprétation unique (i.e. non polysémique) des termes. Par exemple, chaque fois qu'on fera référence au terme « la source » nous utiliserons la même signification correspondant à son ID dans le dictionnaire, et ceci quel que soit le contexte : qu'il s'agisse de « source radioactive », « source d'énergie » ou « des sources du Nil ».

Les *glossaires* présentent une autre spécification des ontologies : les termes y sont associés à des définitions écrites en langue naturelle. Mais les glossaires seuls ne sont pas suffisants pour que les agents logiciels puissent installer la hiérarchie des termes et les interpréter de façon non ambiguë.

Par rapport aux glossaires, les *thésaurus* apportent une sémantique supplémentaire pour les agents logiciels car ils définissent des relations entre les termes. En règle générale trois types de relations sont définis dans les glossaires : l'équivalence (la synonymie), la relation hiérarchique et l'association qui regroupe toutes les autres relations. Mais, dans les ressources de ce type, la propriété de transitivité n'est pas imposée strictement. Ce qui peut provoquer des erreurs lors du traitement automatique.

7. http ://ontolog.cim3.net

Le point suivant sur l'axe est la « *Taxonomie formelle* ». On y définit la relation hiérarchique du type « ***kind-of*** » ou « ***classe – sous-classe*** ». Cette relation est transitive, elle se formule comme suit : *si B est une sous-classe de la classe A, alors chaque* ***sous-classe*** *de la classe B est aussi une sous-classe de la classe A*.

Encore plus à droite sur l'axe de la figure 3.5, on trouve la catégorie des « *instances formelle*s ». Ici le niveau le plus bas de l'ontologie correspond aux instances de la classe pour lesquelles est définie la relation transitive ***is Instance Of*** : *si B est une sous-classe de la classe A, alors chaque* ***instance*** *de la classe B est aussi une sous-classe de la classe A*.

Les ontologies peuvent être également réalisées sous forme de cadres sémantiques où les propriétés des concepts sont prédéfinies par le biais des slots de chaque cadre. Une telle affectation des propriétés est particulièrement utile sur les niveaux supérieurs de la hiérarchie lorsque ces propriétés sont hérités par les sous-classes.

Les ontologies dont les propriétés de classes sont restreintes par le domaine de validité sont encore plus puissantes. Les valeurs des propriétés sont limitées - soit par l'ensemble prédéfini, par exemple les nombres naturels ou les symboles d'un alphabet, - soit par un sous-ensemble des autres concepts de l'ontologie. Il est possible d'introduire des restrictions supplémentaires sur les valeurs de toute propriété.

Généralement, plus riche est l'information qu'il faut transmettre à l'agent logiciel (par exemple, la déclaration de deux ou plus de deux classes disjointes), plus la structure de l'ontologie se complique.

Et enfin, l'expressivité de certains langages permet de construire des assertions logiques arbitraires, dites axiomes, sur les concepts.

Dans le cadre de cette thèse, nous nous focalisons sur les stratégies et les méthodes de construction d'une ontologie de domaine, c'est à dire une ontologie d'application, construite à partir d'un corpus spécialisé et de ressources lexicales complémentaires telles que les dictionnaires.

3.4 Méthodologies générales de construction des ontologies

La notion de méthodologie peut avoir plusieurs significations :

1. Ensemble cohérent de grandes lignes combinant des activités (ou procédures) avec des méthodes (ou techniques). Ou, comme il est mentionné dans IEEE Std.730.1 : « *a comprehensive, integrated series of techniques or methods creating a general systems theory of how a class of thought-intensive work ought to be performed* »[8].
2. Ensemble des techniques utilisées dans le cadre d'un projet déterminé. Dans cette optique, chaque projet a sa propre méthodologie.

8. Une série, exhaustive et intégrée, de techniques ou de méthodes permettant de créer une théorie générale des systèmes qui illustre la manière dont une classe de travail intensif intellectuel peut être réalisé.

3. Étude des méthodes d'une discipline donnée.

Bien sur, les méthodologies générales et les plateformes sont liées. Dans **?** l'auteur sépare rigoureusement ces deux catégories.

Notons que les notions de méthode et de méthodologie sont souvent utilisées dans la littérature de manière interchangeable. Néanmoins le terme « méthodologie » a un sens plus général ; une « méthode » permet la résolution d'une tâche plus limitée.

Dans la première section 3.4.1, nous allons décrire les actions qui assurent le cycle de vie complet d'une ontologie en lien avec le sens donné par l'IEE au mot méthodologie.

Ensuite, dans la section 3.5, nous allons présenter certaines plateformes faisant actuellement référence dans le domaine de l'apprentissage des ontologies.

Les méthodes adoptées pour les différentes étapes de la construction d'une ontologie seront présentées dans le chapitre 4.

Les premiers grands projets ontologiques ont été réalisés au milieu des années quatre-vingt-dix. La dernière décade du vingtième siècle s'est avérée très fructueuse et a vu se développer : CYC (*Lenat* (1995), *Panton et al.* (2006)) ; METHONTOLOGY (*López et al.* (1999)) ; une approche basée sur *le modèle de l'entreprise* (*Uschold et Gruninger* (1996), *Grüninger et Fox* (1995)) ; Unified Process (*DeNicola et al.* (2005)) ; IDEF5 (*Benjamin et al.* (1994)), ONIONS (*Gangemi et al.* (1996)), etc.

Le mérite principal de ces travaux est qu'ils ont permis de répertorier les étapes essentielles du cycle de vie d'une ontologie (cf. Fig. 3.6) et de développer des langages formels de présentation des connaissances.

Ici nous nous limitons à la présentation de trois méthodologies qui agrègent bien les démarches organisationnelles et techniques utilisées, d'une manière ou d'une autre, dans d'autres projets.

3.4.1 Methontology

Nous présentons ici

Methontology est l'une des premières stratégies largement acceptées pour l'élaboration des ontologies. Nous la présentons ici, bien qu'elle soit relativement ancienne, car elle décrit toutes les étapes du cycle de vie d'une ontologie. Proposée par M. Fernandez-Lopez et ses collègues en 1997 *López et al.* (1999), *López et al.* (2000), elle permet de planifier le cycle de vie de l'ontologie en clarifiant les étapes par lesquelles l'ontologie évolue pendant toute sa durée de vie et les tâches nécessaires à chaque étape ; elle permet un développement progressif de l'ontologie. Le processus se déroule depuis la spécification des résultats à atteindre jusqu'à la maintenance, à travers la conceptualisation, la formalisation et la mise en œuvre.

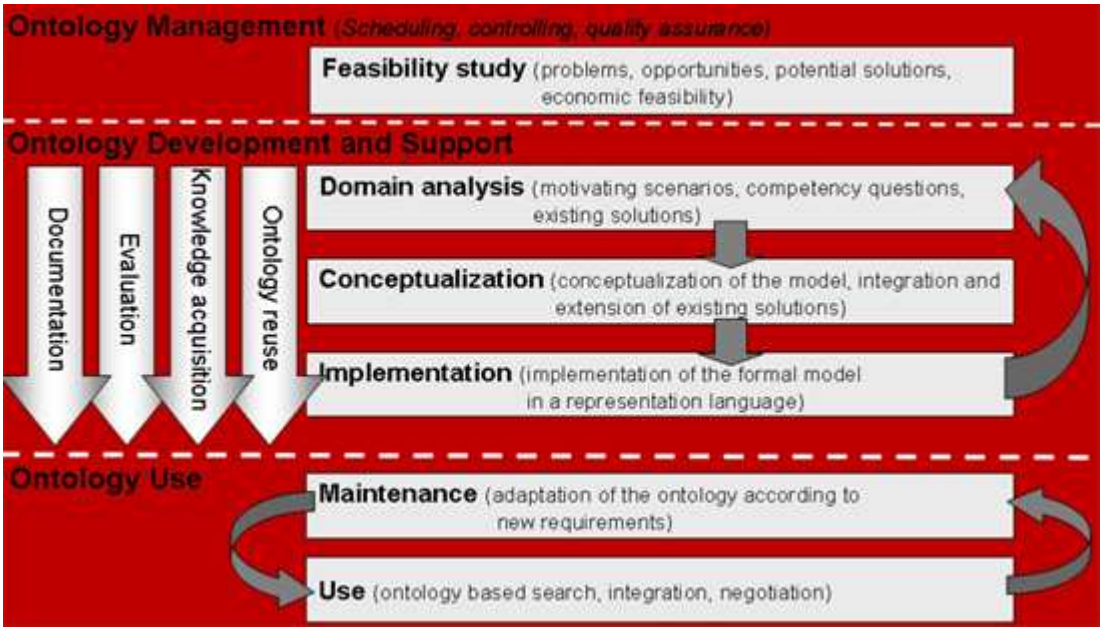

Figure 3.6: Méthodologie généraliste selon *Simperl et al.* (2010).

3.4.2 L'ingénierie des ontologies basée sur l'apprentissage (learning-driven ontology engineering process)

Cette méthodologie s'est construite en analysant plusieurs centaines de projets réels (≈ 500) développés dans différents secteurs économiques *Simperl et al.* (2008), *Simperl et al.* (2010). Les auteurs ont enquêté auprès des concepteurs, des réalisateurs et des utilisateurs d'ontologies afin de clarifier l'importance de chaque étape du développement d'une ontologie et son impact sur les résultats finaux. La méthodologie est proche de celle proposée par N. Noy, *Noy et al.* (2006), mais elle prend en compte plus de détails. Les auteurs ont montré l'importance de la cohérence dans la succession des étapes du cycle de vie d'une ontologie ainsi que les rôles (les responsabilités) de chacun des participants du projet. Finalement, ils ont proposé un modèle détaillé du processus d'apprentissage d'ontologie ; ce processus vise à implémenter le plus facilement possible l'acquisition des connaissances au cours de développement de l'ontologie.

La méthodologie proposée distingue les étapes suivantes :

1. L'étude générale de faisabilité dont l'objectif est d'évaluer les risques et les problèmes qui pourraient perturber l'acquisition des connaissances ; à l'issue de cette phase, la spécification des exigences (ontology requirements specification document, ORSD) peut commencer.
2. La mise au point de l'ORSD peut être considérée comme une instanciation de l'ORSD aux particularités de l'apprentissage ; certains points de la spécification peuvent être révisés et adaptés au nouveau contexte.

3. Les sources d'information à utiliser pour l'acquisition de connaissances ontologiques doivent être sélectionnées et configurées à cette étape. On identifie précisément les méthodes et outils qui vont être utilisés pour extraire l'ontologie visée, en fonction du corpus d'apprentissage sélectionné.
4. Cette étape détaille le processus d'apprentissage de l'ontologie. Après avoir identifié les sources d'information et les outils appropriés, on doit faire appel à des experts pour configurer l'infrastructure technique de la procédure d'acquisition de connaissances. La distinction entre la préparation et l'exécution de l'apprentissage est due à la complexité de cette dernière. Les spécialistes de l'apprentissage, avec la participation des experts, doivent organiser la procédure et dresser le parcours d'intégration des résultats intermédiaires.
5. A cette étape, s'effectue l'acquisition réelle de la connaissance ontologique sur la base de la configuration spécifiée. En cas de détection de problèmes graves dans la configuration des outils, une nouvelle itération du processus précédent s'impose.
6. Évaluation de l'ontologie conformément aux critères d'apprentissage formulés dans la spécification ORSD.
7. Les résultats doivent être intégrés dans l'ontologie finale. Cette étape, non spécifique aux processus d'apprentissage de l'ontologie, peut être effectuée en conformité avec des méthodes d'intégration et des outils existants dans l'ingénierie d'ontologie.

3.4.3 Approche du Modèle d'Entreprise (Enterprise Model Approach)

Une méthodologie standard pour la construction de l'ontologie a été décrite dans *Uschold et Gruninger* (1996). Les auteurs préconisent de définir, dès le tout début, le champ de couverture de l'ontologie et ce à quoi elle va servir. Cette étape fournit les cibles bien définies visées au cours de la construction d'ontologie. Tout le processus suit les trois étapes suivantes :

1. Modélisation préalable : on définit la liste des concepts et des relations clés, on élabore et on met en accord les définitions univoques textuelles de ces concepts et de ces relations ; on identifie les entités à référencer, les concepts et les relations et, enfin, on valide et on adopte les résultats.
2. Codage, avec un langage formel, de la conceptualisation capturée à l'étape précédente.
3. Évaluation formelle ; les critères utilisés peuvent être généraux ou spécifiques à un domaine particulier. Cette étape peut entraîner une révision des sorties des phases 2 et 3.

Dans l'approche du *modèle d'entreprise*, les phases formelles et informelles de la construction de l'ontologie sont bien distinguées. La phase informelle consiste à iden-

tifier les concepts et les relations clés et à donner des définitions explicites textuelles. Néanmoins, dans cette approche on ne donne aucune consigne spécifique sur l'identification des concepts ontologiques en dehors de recommandations générales sur l'acquisition des connaissances.

3.5 Projets et plateformes ontologiques

Dans cette partie nous allons aborder brièvement trois projets dont les idées de base résument, à notre avis, les approches de construction d'ontologies utilisant les méthodes d'apprentissage. Ces approches seront présentés dans le chapitre4.

D'autres plateformes intéressantes (ASIUM, DL-Learner, DODDLE, GATE, medSynDiKATE, NeOn, OntoGain, OntoGen, OntoLT, SVETLAN', TermExtractor) sont présentées dans l'Annexe A (6.2).

Le site `http://www.mkbergman.com/` recense et diffuse régulièrement les outils, plateformes et méthodes d'Intelligence Artificielle et leur actualisation.

3.5.1 TERMINAE

La plateforme TERMINAE matérialise la méthode proposée par *Aussenac-Gilles et al.* (2008). Elle s'appuie largement sur les ressources textuelles, option qui se développe en France depuis les années quatre-vingt-dix sous l'impulsion du groupe de recherches du LIPN. L'idée de base est qu'on peut construire le modèle conceptuel d'un domaine à partir des formes linguistiques réelles présentes dans les textes. La méthode utilise les techniques de Traitement Automatique de la Langue Naturelle (TALN), ainsi que des ressources terminologiques en ligne, ou d'autres ontologies pré-existantes, qui condensent des connaissances accumulées au préalable par les experts.

TERMINAE combine des outils d'acquisition des connaissances fondées sur la linguistique avec des techniques de modélisation, ce qui permet de maintenir les liens entre les modèles et les textes.

L'apprentissage de l'ontologie est effectué en trois niveaux : le niveau linguistique , le niveau de la normalisation et le niveau formel (cf. la figure. 3.7).

Les termes et les relations lexicales extraits des textes par des méthodes d'analyse linguistique. Ensuite, le niveau de normalisation comprend un modèle conceptuel exprimé par un réseau sémantique composé des concepts et des relations qui les lient. Les concepts ont deux dimensions : la dimension linguistique, montrant la proximité entre un concept et les syntagmes terminologiques dans le corpus, et la dimension pragmatique qui donne les raisons d'intégration du concept dans le modèle formel. Ce modèle conceptuel est peu formalisé mais il peut être facilement compris par le concepteur de l'ontologie.

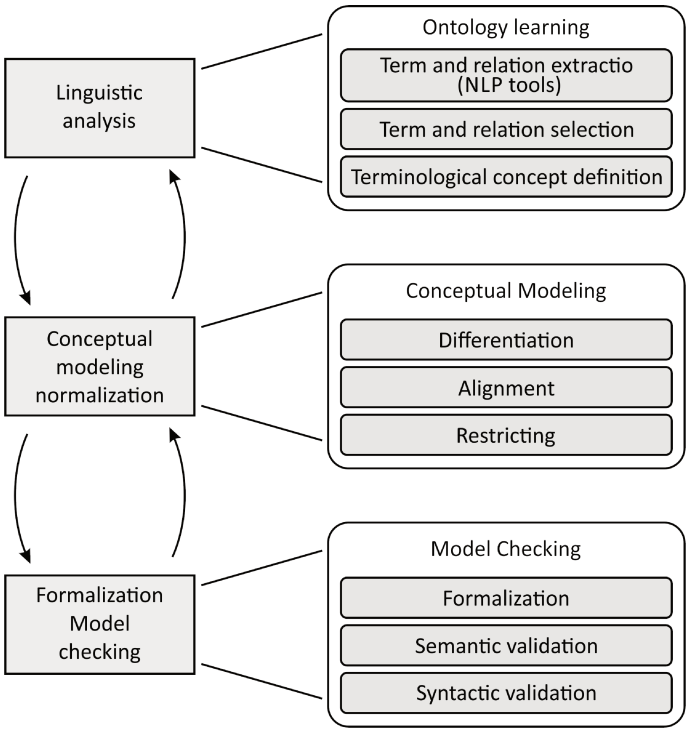

Figure 3.7: Fonctionnalité de la plateforme TERMINAE.

À la fin, les concepts et les relations conceptuelles sont mis en forme grâce à un langage de description formelle.

3.5.2 text2Onto

Text2Onto (*Cimiano et Völker* (2005)) est une plateforme conçue pour la construction d'ontologies à partir de ressources textuelles. C'est une version nouvelle, redéfinie et perfectionnée, de TextToOnto dont les principes de base sont eux-même hérités du système GATE (*Cunningham et al.* (2011)). L'architecture de Text2Onto est présentée dans la figure 3.8.

Son module central est le Modèle Probabiliste d'Ontologie (Probabilistic Ontology Model, POM) qui stocke les résultats d'application de différents algorithmes ; ceux-ci sont initialisés par un contrôleur qui a trois fonctions :

1. Amorcer le processus du prétraitement linguistique des données textuelles ;
2. Exécuter des algorithmes d'apprentissage dans un ordre prédéfini ;
3. Modifier les paramètres du modèle POM.

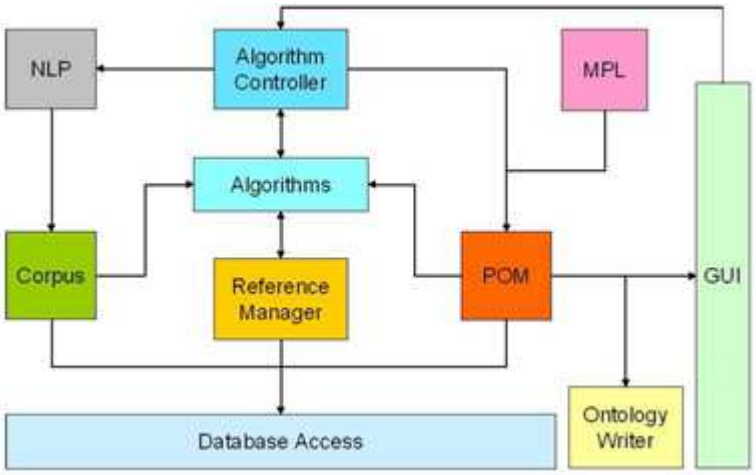

Figure 3.8: Schema modulaire de text2Onto.

Grâce à cette répartition en trois fonctions, aucun des algorithmes n'a la permission d'affecter directement le POM.

La nouveauté du POM, réalisé dans Text2Onto, est que la pertinence (la probabilité) des candidats-termes est ajoutée aux résultats présentés dans l'interface graphique, ce qui facilite la prise de décision par l'utilisateur.

Le deuxième nouveau paradigme de cette plateforme est la révélation explicite des changements dans les modèles probabilistes provoqués par de changement des données (data-driven change discovery). L'avantage de cette technique est qu'elle peut être appliquée à un sous-ensemble de textes ; on n'a donc pas besoin d'interroger tout le corpus pour corriger les scores des candidats-termes.

Le POM fait référence directe aux ensembles instanciés des primitives de modélisation. Text2Onto comprend une bibliothèque de modélisations des primitives (Modeling Primitive Library), chacune correspondant à la description complète d'un des éléments de l'ontologie tels que : concept (CLASS), concept hérité (SUBCLASSE-OF), concept d'instanciation (INSTANCE-OF), propriété/relation (RELATION), domaine et limitation (DOMAN/RANGE), relations méréologiques et relation d'équivalence.

Les primitives de modélisation ne sont pas implémentées dans des langages concrets de représentation des connaissances. Ce choix des concepteurs de Text2Onto offre plus de flexibilité pour la transformation finale des résultats en ontologie formelle.

Le support de l'instanciation des primitives de modélisation est réalisé par des algorithmes conçus pour l'extraction de candidats-termes, correspondant chacun à certaines « primitives de modélisation » dans le corpus. Plusieurs algorithmes différents peuvent être librement choisis par l'utilisateur pour l'extraction d'un même type de primitives.

La stratégie d'apprentissage d'ontologie de la plateforme Text2Onto est liée au paradigme dit « Layer Cake » qui sera présenté à la section 4.1 du chapitre 4.

3.5.3 Lemon (Lexicon Model for Ontologies)

Les modèles du type LEMON (Lexicon modèle for ontologies) poursuivent les buts différents de ceux de *l'apprentissage des ontologies*, mais l'idée d'élaborer des modèles linguistiques fondés sur les principes ontologiques est prometteuse : premièrement pour rendre plus lisibles les connaissances structurées (avec RDF(S)ou OWL) et déjà collectées dans le Web sémantique, et deuxièmement pour enrichir (voir construire) l'ontologie à partir des textes. C'est pourquoi nous abordons brièvement le modèle LEMON dans cette section.

Le modèle LEMON a été développé dans le cadre du projet « Monnet » pour créer un format standard permettant d'enrichir les objets ontologiques avec des informations linguistiques qui sont déjà codées dans un des formats du Web sémantique et, par là, de garantir la liaison entre la sémantique d'un concept et sa réalisation linguistique.

Le modèle est basé sur le principe de « la sémantique par la référence » *McCrae et al.* (2011), *Buitelaar et al.* (2011). Ce principe signifie que toute entrée lexicale peut être liée, de manière explicite, à un objet d'une ontologie.

LEMON est écrit avec le langage *Turtle* dont la syntaxe textuelle permet de réaliser le graph RDF sous forme de patrons facilement compréhensibles et adaptés à un large usage partagé ; c'est ce qui a permis de caractériser ce modèle comme « ontologique » (ontology-based).

Notons que, depuis 2014, ce modèle est recommandé par le consortium W3 pour la présentation des données dans les ressources lexicales.

Le noyau du module LEMON est présenté sur la figure 3.9. L'objet principal dans ce modèle est *Lexicon* pour l'ensemble des entrées lexicales. Par la suite, chaque entrée lexicale est envisagée comme un objet distinct muni des informations morphosyntaxiques, *LexicalEntry* et *LexicalForm* respectivement.

LexicalSense rajoute les informations et les contraintes sur l'usage de l'entrée lexicale dans le contexte du domaine d'application.

Dans le modèle LEMON, tout élément peut être décrit à l'aide d'une propriété générique *LexicalProperty* d'où on peut dériver d'autres propriétés ; ainsi toutes les propriétés lexicales peuvent être regroupées. Actuellement, le français peut être modelisé avec LEMON, comme plusieurs autres langues.

Le module qui nous intéresse ici est celui de « *syntaxe et mappage* » *syntax and mapping module*, qui réalise la description syntaxique des entrées lexicales et leurs références à l'ontologie (cf. la figure 3.10).

L'élément *Frame* est le cadre syntaxique dans lequel l'entrée peut apparaître. C'est la description explicite du prédicat avec ses arguments (i.e. son comportement syntaxique).

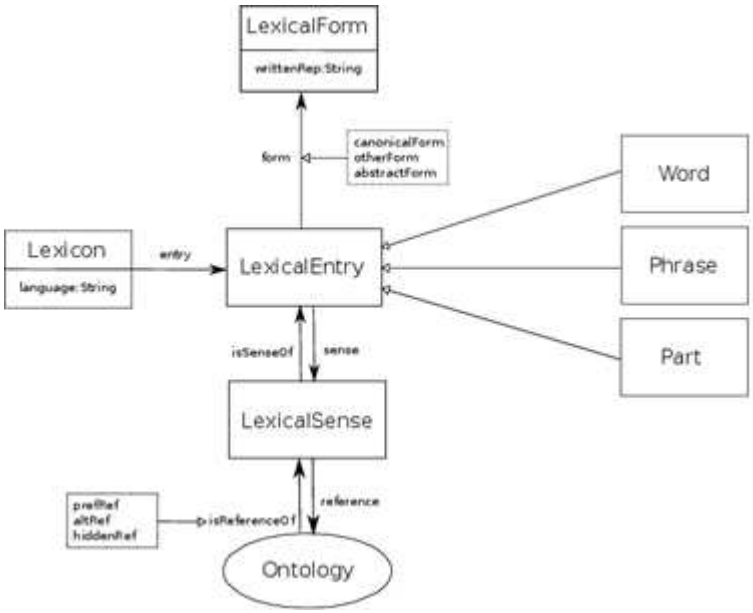

Figure 3.9: Schéma de Lemon.

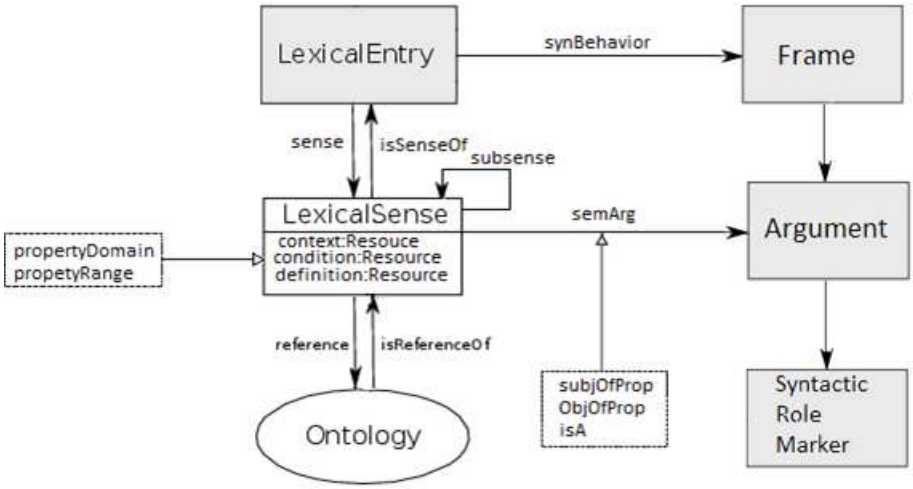

Figure 3.10: Module syntaxique de Lemon.

3.6 Méthodes pour l'apprentissage d'ontologie

3.6.1 L'Analyse de Concepts Formels, FCA

Rappel de la méhode FCA

L'Analyse de Concepts Formels (FCA, *Formal Concept Analysis*), est une méthode privilégiée de construction d'ontologie ; elle est basée sur la théorie des ensembles ordonnés

(ou treillis) et permet de constituer une hiérarchie de concepts à partir de l'analyse des attributs qu'ils partagent.

Il existe plusieurs extensions de la méthode FCA, également applicables à la construction d'une ontologie de domaine, à savoir l'Analyse Relationnelle de Concepts et l'Analyse Logique de Concepts, mais nous ne les présentons pas ici.

Selon les propres dires de Ph. Cimiano, le « goulet d'étranglement » dans la construction d'une ontologie est la modélisation du domaine, où le rôle important est la définition des relations entre les concepts, *Cimiano et al.* (2005).

La méthode FCA permet de visualiser les dépendances entre les objets à l'aide d'un treillis de Galois *Ganter et Wille* (1997). Elle peut être utilisée lors de l'analyse des données afin de détecter les relations entre les éléments d'un système et, dans le cas d'une ontologie, entre les étiquettes linguistiques des concepts présents dans un corpus.

Les relations se manifestent à travers les attributs décrivant les propriétés. Á leur tour, les attributs doivent être proches des catégories de la raison humaine de façon à être compréhensibles. Ainsi la FCA peut être envisagée comme un algorithme de regroupement conceptuel (conceptual clustering) ; son cadre général est présenté dans la figure 3.11 tirée de *Cimiano et al.* (2005).

Figure 3.11: L'ensemble des étapes de la construction d'une ontologie avec FCA.

Deux notions principales sont à la base de la FCA : celle de *contexte formel* et de *concept formel* dont les définitions sont formulées ci-dessous. La hiérarchie résultante, regroupant les objets partageant les mêmes propriétés, est appelée *treillis des concepts*.

Définition. Un *contexte formel* est le triplet $K = (G; M; I)$ où G et M sont des ensembles finis et disjoints dont les éléments, g et m, sont liés par la relation binaire I : $I \subseteq G \times M$ ou gIm.

Les éléments de l'ensemble G sont dénommés *objets*, les éléments de l'ensemble M sont dénommés *attributs* et les élément de l'ensemble I représentent l'incidence du contexte, c'est à dire le fait que les éléments de l'ensemble G possèdent des attributs de l'ensemble M.

On définit l'ensemble *A'* pour $A \subseteq G$ comme l'ensemble des attributs communs pour tous les objets de *A* :

$$A' := \{m \subseteq M \mid (\forall g \in A : (g, m) \in I)\}$$

et l'ensemble *B'* pour $B \subseteq M$ comme l'ensemble des objets qui possèdent tous les attributs *A'* :

$$B' := \{g \subseteq G \mid (\forall m \in B : (g, m) \in I)\}$$

Définition. La paire *(A,B)* de deux ensembles disjoints des objets *A* et des attributs *B*, respectivement, est le *concept formel* du contexte formel $K = (G; M; I)$, si l'affirmation suivante est correcte : chaque objet $g \in A$ possède tous les attributs de l'ensemble d'attributs *B* et, vice versa, chaque attribut $m \in B$ est possédé par tous les objets de l'ensemble des objet *A*.

Étant donné cette définition, la paire (*A,B*) est un concept formel du (*G; M; I*) si et seulement si $A \subseteq G$, $B \subseteq M$, $A' = B$, $B' = A$.

Les concepts formels obtenus dans le contexte formel sont ordonnés par la relation hiérarchique :

$$(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 (\Leftrightarrow B_2 \subseteq B_1)$$

Traditionnellement, pour présenter le contexte formel on utilise un tableau dont les en-têtes de lignes correspondent aux objets et les en-têtes des colonnes correspondent à leurs attributs. La table 3.3 illustre le fragment du contexte formel pour la distinction entre les organes ou tissus du corps humain et les pathologies cancéreuses. Les résultats présentés dans le tableau ont été obtenus au cours de nos expérimentations sur les textes du domaine de la radiologie et de l'utilisation des isotopes radioactifs pour le traitement des cancers *Orobinska et Sharonova* (2011). Ici *G* correspond à l'ensemble des organes humains et leurs pathologies carcinomateuses ; *M* correspond à l'ensemble des propriétés. Le treillis résultant est illustré par la figure 3.12. L'emplacement des nœuds du treillis indique l'héritage des propriétés par les objets. La propriété est affichée au-dessus d'un nœud tandis que l'élément qui possède cette propriété est affiché au-dessous du nœud. De la même façon, tous les objets qui se placent au-dessous de l'objet *g*, sous le nœud marqué par la propriété **m**, sont plus spécifiques que *g*.

Les connaissances qui correspondent au contexte formel présenté dans la table 3.3 se formulent comme suit : des organes du corps humain (poumons, estomac, foie, thyroïde) ainsi que des tumeurs malignes (tumeur de l'estomac, cancer du poumon, cancer de la thyroïde, cancer du foie) peuvent être détruits par l'exposition au rayonnement ionisant, mais seules les maladies peuvent être diagnostiquées et guéries alors que les organes risquent, eux, d'être abimés par le rayonnement.

Cette table 3.3 est représentée comme hiérarchie de concepts dans la figure 3.13.

Le passage du treillis à l'ontologie se réalise à l'aide de la fonction α telle que $\alpha : B(G, M, I) \longrightarrow TBox \cup ABox$. *B(G,M,I)* est le treillis des concepts obtenu par FCA. *TBox*

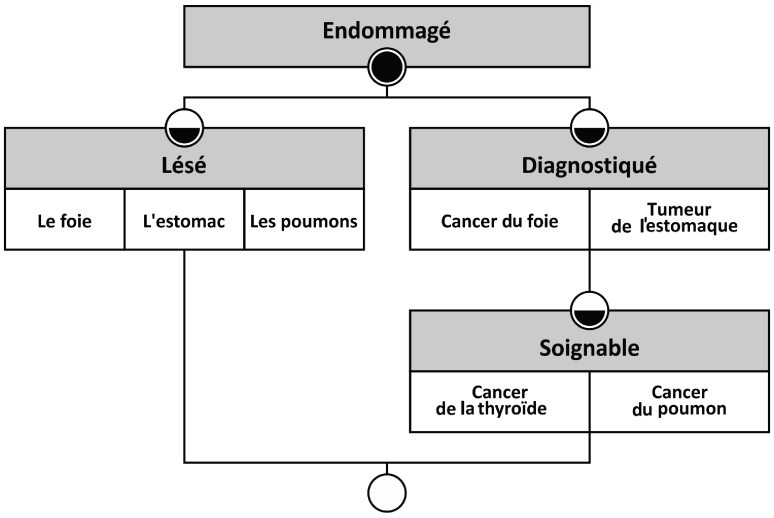

Figure 3.12: Le treillis correspondant au contexte formel donné.

correspond à la conceptualisation de domaine, i.e. à l'ensemble des concepts et leurs relations ; *ABox* est l'ensemble des affirmations sur chaque concept, soit la description de ses propriétés, i.e. son intention. La conjonction de *TBox* et *ABox* forme la base de connaissances, *Baader et al.* (2010).

Dans les termes de la logique descriptive, *TBox* est le noyau d'ontologie (ou l'ontologie de base) réalisé par le triplet $O := (C, \subseteq c, A)$, où C est l'ensemble des concepts, $\subseteq c$ est la relation hiérarchique de subsomption (*is-a*) et *A* est l'ensemble des attributs des concepts (ou ses propriétés).

La base de connaissances de l'ontologie *O* est représentée par la structure $KB := (I, i_C, i_A)$ où $i_C := C \rightarrow 2^I$ est la fonction d'assignation des entités, et $i_A := C \rightarrow 2^A$ est la fonction d'assignation des attributs.

La fonction de transformation α qui formalise le passage peut être représentée à l'aide de la table 3.4

L'ontologie obtenue est présentée à l'expert qui doit vérifier la correspondance de l'ensemble des objets au concept en partant des propriétés partagées par des éléments de l'ensemble. Par exemple, les objets qui possèdent les propriétés (*diagnostiqué, abimé*) peuvent être étiquetés par la notion d'*organe*, tandis que le groupe avec les propriétés (*diagnostiqué, détruit*) peut être désigné comme *cancer*.

L'outil mathématique de la logique descriptive (LD) est adapté à la présentation formelle d'une ontologie. Les concepts sont définis par l'opération de conjonction de ses attributs, chacun déclaré par le quantificateur existentiel. Pour notre exemple, le résultat de la transformation du treillis en ontologie est présenté figure 3.13. La définition de chaque concept de ce fragment d'ontologie est présentée dans la table 3.5.

	Peut être endommagé	**Peut être diagnostiqué**	**Peut être lésé**	**Peut être guérissable**
Tumeur de l'estomac	×	×		
Le cancer du poumon	×	×		×
Les poumons	×		×	
Cancer de la thyroïde	×	×		×
L'estomac	×		×	
Le foie	×		×	
La thyroïde	×		×	
Le cancer du foie	×	×		

Table 3.3: Exemple du contexte formel représentant les propriétés reliant les organes et les pathologies cancéreuses.

Treillis des concepts	**Ontologie en LD**
Contexte formel $K(G, M, I)$	Concept atomique $c \equiv \alpha(K)$
Objet $g \in G$	Entité $\alpha(g)$ dans $ABox\bot(g)$
Propriété $m \in M$	Rôle atomique $\alpha(m) \equiv \exists m.\top$ dans *TBox*
Élément $(g, m) \in I$	Proposition
Concept $c = (A, B) \in C$	Concept défini dans *TBox* $\alpha(c) \equiv \cap m \in B\alpha(m)$
$\forall(c, \overline{c}) \in C \times C$ tel que $c \prec \overline{c}$	Inclusion des axiomes $\alpha(m) \subseteq \alpha(\overline{m})$ dans *TBox*
Ensemble des attributs des concepts $\Lambda_1^n c_i$	Conjonction des concepts $\alpha(c_1) \cap \ldots \cap alpha(c_n)$

Table 3.4: Formalisation du passage du treillis à l'ontologie.

L'Analyse Relationnelle de Concepts, ARC

L'Analyse Relationnelle de Concepts peut être envisagée comme une extension de l'Analyse de Concepts Formels (FCA) ; son objectif est de prendre en considération, non seulement les attributs des concepts, mais aussi les relations entre les concepts. Initialement la méthode a été présentée dans *Rouane et al.* (2007). La notion centrale de l'ARC est celle de l'*ensemble des contextes relationnels* (*K,R*) où *K* est l'ensemble des contextes $K_i = (G_i, M_i, I_i)$, étant précisé qu'à chaque ensemble d'objets *G* correspond un seul contexte ; *R* est l'ensemble des relations $r_i \subseteq G_i \times G_j$ entre deux ensembles d'objets.

La méthode basée sur l'ARC exploite l'ensemble des treillis de concepts qui, ayant été superposés, donnent le treillis final dans lequel les concepts sont liés entre eux par des relations non seulement hiérarchiques mais aussi associatives, horizontales.

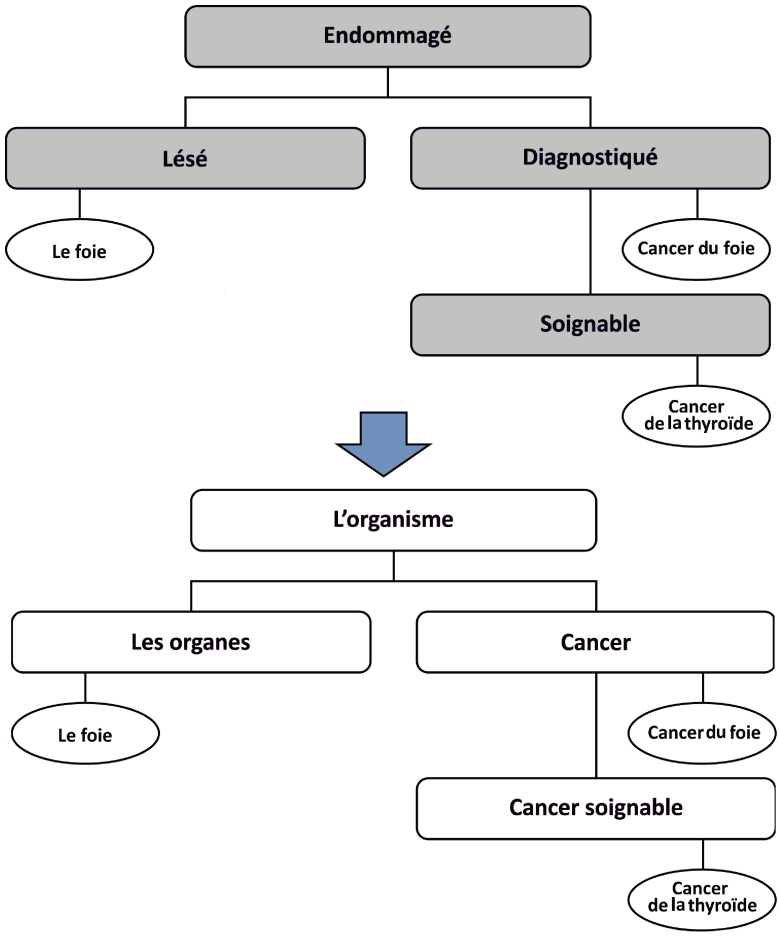

Figure 3.13: Transformation du treillis en l'ontologie et définition des concepts.

Reprenons l'exemple du domaine de la radiologie. Soit deux contextes construits à la base du FCA présentés dans les tables 3.3 et 3.6 : $K_1 = (G_1, M_1, I_1)$ qui contient l'énumération des organes et leur pathologies et $K_2 = (G_2, M_2, I_2)$ l'ensemble des mesures utilisées en radiologie.

L'intégration des relations et des treillis est réalisée par le processus de pondération décrit en détails dans *Rouane et al.* (2007). Le treillis résultant donne, par exemple, les relations : *diagnostiqué à l'aide* entre les objets « cancer de la thyroïde » et « iode-131 » et *guérissable par* entre les objets « cancer du poumon » et « cobalt-60 » (cf. la table 3.7).

Grâce aux relations établies entre les concepts, on peut enrichir le noyau d'ontologie et le transformer en une ontologie plus complète (cf. la figure 3.14).

Concept
$Objet$:= ∃*endommagé*
$Organe$:= ∃*endommagé* ∩ ∃*lésé*
$Cancer$:= ∃*diagnostiqué* ∩ ∃*endommagé*
$CancerGuerissable$:= ∃*diagnostiqué* ∩ ∃*endommagé* ∩ ∃*guérissable*

Table 3.5: Définition des concepts de l'ontologie en Logique Descriptive.

	Période radioactive <2h	Période radioactive entre 24h et 10 jours	Période radioactive >100 jours
iode-131		×	
cobalt-60			×
carbone-11	×		

Table 3.6: Le fragment du contexte formel pour les radionucléides.

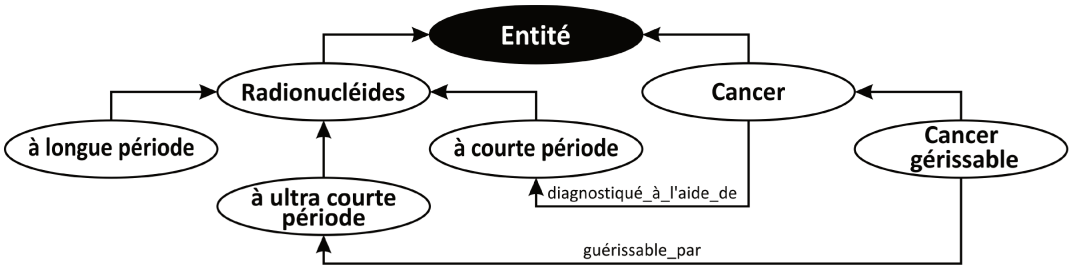

Figure 3.14: L'ontologie complétée par les relations associatives.

3.6.2 Méthode de l'identification comparative

La méthode de l'identification comparative s'organise autour du modèle donné par le prédicat $E(x_1, x_2)$ du type $E(x_1, x_2) = D(F[x_1], [x_2])$, où x_1 et x_2 sont les éléments de l'ensemble des signaux d'entrée ; $y_1 = F[x_1]$ et $y_2 = F[x_2]$ sont les éléments de l'ensemble *B* des signaux de sortie ; *D* est le prédicat standard d'égalité qui est défini sur le carré cartésien de l'ensemble B. La méthode est décrite plus en détail par la suite.

Historique

L'idée de l'identification comparative a été proposée pour la première fois par I. Newton pour la théorie sur la vision des couleurs par l'homme. Les signaux d'entrée du système visuel de l'homme, (les rayons lumineux) sont accessibles à la mesure objective

Diagnostiqué à l'aide de			
	Iode-131	Cobalt-60	Carbone-11
Cancer de la thyroïde	×		
Cancer du foie			×
Guérissable par			
Cancer du poumon		×	
Cancer de la thyroïde		×	

Table 3.7: Treillis des relations entre les concepts.

mais il est impossible de mesurer les couleurs qui sont les signaux de sortie de l'organe de la vision, parce que la perception de la couleur est subjective et inaccessible à l'observation directe physique.

L'idée de Newton était que le candidat est capable d'établir l'égalité des couleurs de deux rayonnements lumineux qu'on lui présente : même si le processus de comparaison est subjectif, son résultat peut être enregistré physiquement par la réaction binaire du candidat. Par exemple, dans chaque épreuve on enregistre une des deux valeurs : 1, si les couleurs coïncident, et 0, si non.

En généralisant, le problème peut être formulé comme suit : pour deux signaux connus, d'entrée *x* et de sortie *y*, il faut définir la loi de transformation $y = F(x)$.

l'Algèbre des Prédicats Finaux (APF)

La méthode de l'identification comparative se fonde sur l'Algèbre des Prédicats Finaux (APF) qui est une généralisation de la logique propositionnelle et de l'algèbre de Boole ; la méthode a été développée par Ju. P. Shabanov-Kushnarenko et ses successeurs *Shabanov-Kushnarenko* (1984), *Sharonova* (2010).

L'APF permet d'élaborer un modèle de représentation des connaissances d'un domaine en réduisant la redondance d'une langue naturelle ; elle utilise un ensemble d'équations, chacune étant formée par des combinaisons de propositions simples (ou prédicats) à l'aide d'opérateurs logiques. On résout le système d'équations de prédicat à l'aide des algorithmes traditionnels, avec un langage de programmation orienté objet (tel Java, par exemple).

Comme l'APF peut manier des variables « texte », c'est un outil adapté à la modélisation des relations d'une langue naturelle, tout en respectant les exigences du formalisme d'une langue artificielle, notamment la complétude et la non-contradiction (consistance). Il est prouvé que l'APF est « complète », c'est à dire qu'elle permet la description de toutes les relations finies.

L'avantage principal de l'APF est la représentation de la langue naturelle sous forme d'un système d'équations logiques ; cela permet d'effectuer les tâches de traitement de

tous les niveaux de la langue (morphologiques, syntaxiques, sémantiques) sans changement de support linguistique.

Plusieurs définitions

Définition 1. Une proposition est une expression de toute nature dont on peut affirmer qu'elle est vraie ou fausse.

Définition 2. Un prédicat est une fonction de domaine de validité $0, 1$ (1 pour *Vrai* et 0 pour *Faux*), définie sur le produit cartésien de *n* ensembles d'objets de nature quelconque.

Définition 3. La relation binaire R sur les ensembles X et Y est un sous-ensemble du produit cartésien $X \times Y$ soit l'ensemble des paires ordonnées $(x, y) \in X \times Y$ ou xRy. Par analogie, on peut généraliser la définition d'une relation comme le produit cartésien des *n* ensembles A_n non vides (qui ne sont pas obligatoirement différents). Dans ce cas, on dit que la relation R est à *n* dimensions (ou est d'*arité n*). Sur ces relations, on définit les opérations de disjonction, conjonction et soustraction logique.

Définition 4. La relation binaire R définie sur les ensembles X et Y est fonctionnelle si xRy_1 et xRy_2 impliquent $y_1 = y_2$; soit

$$\forall x \in X, \forall y, \in Y, (x, y_1) \in R, (x, y_2) \in R \Rightarrow y_1 = y_2$$

Toute relation R d'arité *n* définie sur les ensembles $X_1, X_2, \cdots X_n - 1, Y$ est fonctionnelle si $(x_1, x_2, \cdots x_n - 1, y_1 \in R)$ et $(x_1, x_2, \cdots x_n - 1, y_2 \in R)$ implique $y_1 = y_2$. Cela signifie qu'à chaque valeur $x \in X$ correspond une seule valeur $y \in Y$.

Les relations qui ne sont pas fonctionnelles sont qualifiées de polyfonctionnelles : à chaque ensemble de variables $(x_1 \in X_1, x_2 \in X_2, \cdots x_n - 1 \in X_n - 1)$ peut correspondre plusieurs valeurs $(y_1, y_2 \cdots y_i \in Y)$ avec $y_1 \neq y_2 \neq \cdots \neq y_i$.

Avec les définitions 1 à 3, on peut définir un prédicat comme une fonction sur l'ensemble de *n* éléments, chacun présentant une certaine relation réalisée sous forme d'un prédicat de reconnaissance (recognition predicate).

À tout jeu de variables $(x_1, x_2, \cdots x_n)$ on peut mettre en correspondance l'équation canonique, i.e. l'équation du type $f_R(x_1, x_2, \cdots x_n) = 1$. Le prédicat de f_R se défini comme suit :

$$f_R = \begin{cases} 1 \; si \; (x_1, x_2, \ldots, x_n) \in R \\ 0 \; si \; (x_1, x_2, \ldots, x_n) \notin R \end{cases}$$

Inversement, chaque élément x_i peut être représenté par le prédicat de reconnaissance :

$$x_i^a = \begin{cases} 1\ si\ x_i = a \\ 0\ si\ x_i \neq a \end{cases}$$

L'Algèbre des Prédicats Finaux (APF) convient pour la description des connaissances représentées par les textes écrits en langue naturelle (LN). Dans une langue naturelle, on distingue trois niveaux des relations :

1. Les relations définies sur l'ensemble des lettres de la langue donnée. À ce niveau correspondent les prédicats de mots (et leurs paradigmes syntaxiques) ;
2. Les relations définis sur l'ensemble des mots de la LN et déterminées par les règles syntaxiques de la langue. Les relations de ce niveau correspondent aux opérations sur les prédicats de mots ;
3. Les relations définies sur l'ensemble des concepts présentés à travers le texte qui comprennent les deux premiers groupes de prédicats et correspondent aux opérations sémantiques sur les concepts.

Exemple d'application de la méthode de l'identification comparative

Nos expérimentations ont été réalisées sur des corpus en langue française et en langue russe.

En l'utilisant comme modèle de représentation des connaissances sur une ontologie de domaine, l'APF permet de passer de la description algorithmique des relations entre concepts à leur description sous forme d'équations.

Dans sa définition formelle, une ontologie comprend des relations de hiérarchie du type *is-a*. Dans toute langue naturelle, la relation hiérarchique est la relation binaire R, définie par l'ensemble des formes lexicales *D*, telle que pour chaque $d \in D$, existe toujours $d'Rd \in D$. De plus, R n'est pas réflexive pour tout $d \in D$, mais elle est transitive et antisymétrique. Le nombre des variables dans le prédicat $f_R(x_1, x_2, \ldots x_n) = 1$ correspond au nombre *n* des niveaux de la hiérarchie.

Chaque niveau de la hiérarchie peut être présenté comme suit :

$$\begin{array}{l} x_1^1 \equiv (x_2^{a_1} \vee x_2^{a_2} \vee \ldots \vee x_2^{a_i}) = 1 \text{ où } (a_1, a_2, \ldots, a_i) \in A \\ x_2^{a_1} \equiv (x_3^{b_1} \vee x_3^{b_2} \vee \ldots \vee x_3^{b_k}) = 1 \text{ où } (b_1, b_2, \ldots, b_m) \in B, m \geq k, \\ \ldots\ldots\ldots\ldots\ldots \\ x_{n-1}^1 \equiv (x_n^{z_1} \vee x_n^{z_2} \vee \ldots \vee x_n^{z_j}) = 1 \text{ où } (z_1, z_2, \ldots, z_l) \in Z, l \geq j \end{array}$$

où $A, B, \ldots Z$ sont des sous-ensembles d'un dictionnaire qui représentent les niveaux différents de la hiérarchie. Selon le théorème de la décomposition, on peut écrire le prédicat $f_R(x_1, x_2, \ldots x_n)$ sous forme (sous la condition que $a = 1$, $b = 1$ et $l = 1$) :

$$(x_1^1 \equiv \vee x_2^{a_1})(x_2^{a_i} \equiv \vee x_3^{b_m})...(x_{n-1}^{c_s} \equiv \vee x_n^{t_l}) = 1$$

Un exemple de fragment de la hiérarchie de l'ontologie sur la radioprotection est présenté sous forme d'arbre sur le 3.15.

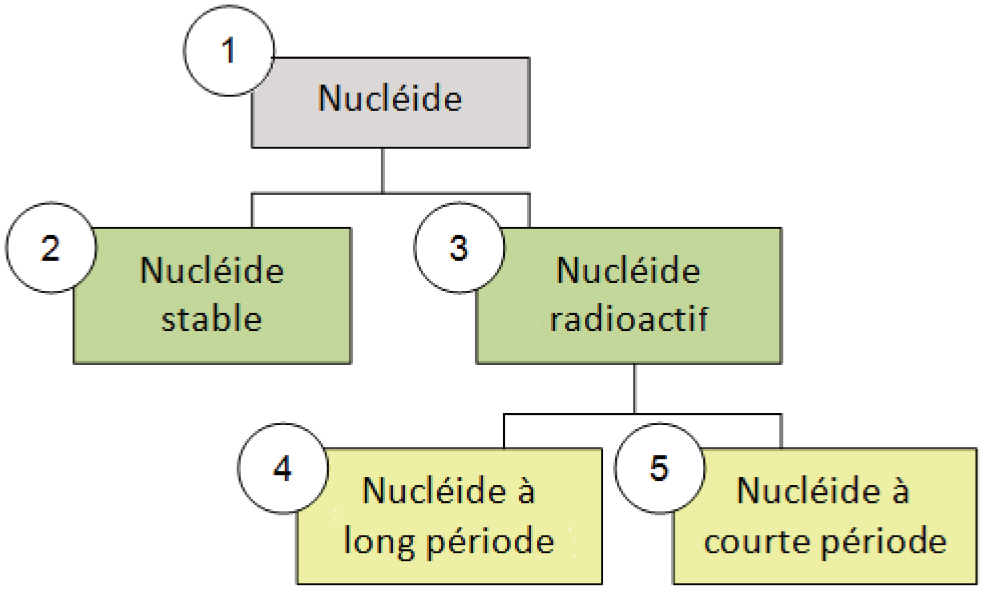

Figure 3.15: Hiérarchie des notions.

Ici on a les variables $x_1 = 1$, $x_2 = 2, 3$, $x_3 = 4, 5$.

L'équation qui décrit cet arbre est :

$$(x_1^1 \equiv x_2^2 \vee x_2^3)(x_2^3 \equiv x_3^4 \vee x_3^5) = 1$$

Sa solution correspond aux feuilles de l'arbre.

3.7 Conclusion du chapitre 3

Les approches générales de l'apprentissage d'ontologie visent à affiner les méthodes automatiques pour réduire le travail manuel. Le succès de ces méthodes dépend beaucoup de la qualité et de l'exhaustivité des ressources linguistiques utilisées comme « sources primaires » pour le développement d'une ontologie.

Les experts ont deux tâches : premièrement former un corpus représentatif de l'ensemble du domaine et choisir les ressources spécialisées telles que les dictionnaires ou thésaurus ; deuxièmement valider les résultats des étapes intermédiaires du processus de développement de l'ontologie (apprentissage).

Les problèmes d'organisation (faisabilité, cadrage, ordonnancement) sont cruciaux pour toutes les méthodologies proposées.

Chapitre 4

L'apprentissage des ontologies : méthodes et techniques

4.1 Spécificité de la tâche

Comme mentionné dans le chapitre 3, la construction d'une ontologie comprend plusieurs étapes ; on utilise tout un ensemble de techniques pour acquérir les éléments constituants l'ontologie.

Une des tâches est l'intégration des termes isolés dans une structure hiérarchisée de concepts, munie de règles d'inférence et d'axiomes ; traditionnellement, on utilise une stratégie connue sous le nom de « *Layer Cake* » (*Buitelaar et Magnini* (2005), *Wong et al.* (2012)) ; plusieurs étapes se succèdent, et à chaque stade on utilise comme entrées les résultats de l'étape précédente (Fig. 4.1, Fig. 4.2).

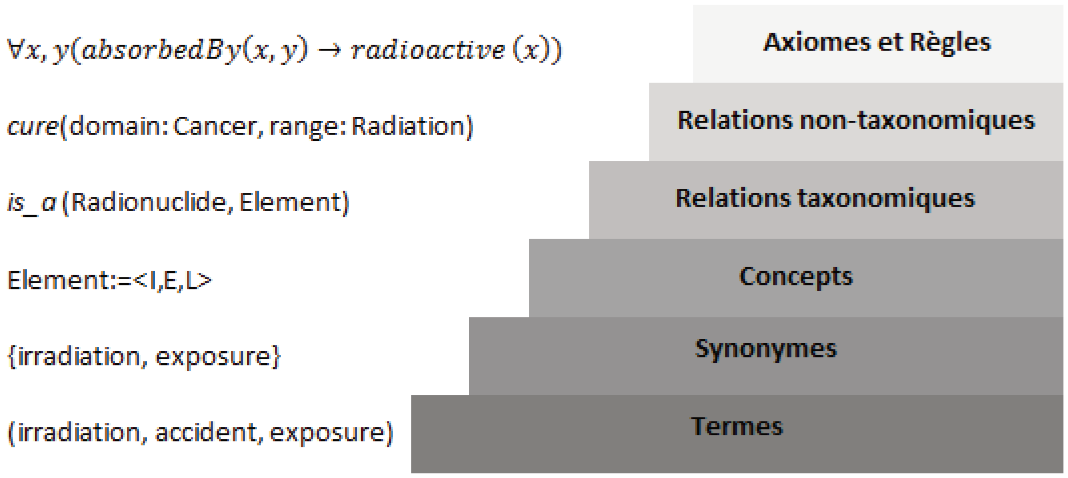

Figure 4.1: Ontology learning "layer cake".

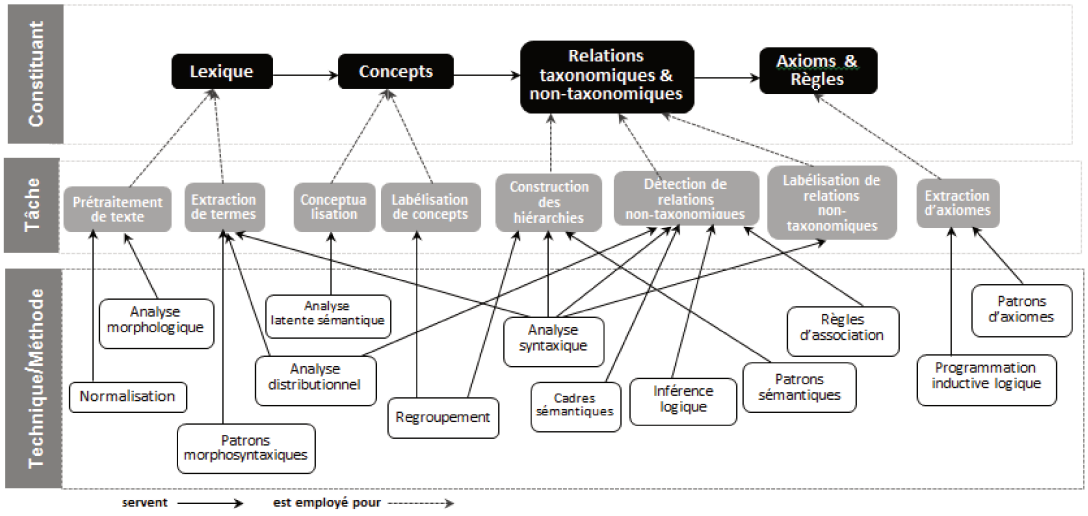

Figure 4.2: Enchaînement des tâches, indication des techniques adoptées pour chacune et éléments d'ontologie produits.

Cette stratégie *Layer Cake* a été mise en œuvre, de façon plus complète, dans la plateforme ***Text2Onto*** (cf. la section 3.5.2 du chapitre 3).

L'approche *Layer Cake* a été critiquée par *Mondary et al.* (2008). Les auteurs soulignent notamment que les étapes de construction de l'ontologie se déroulant l'une après l'autre dans un ordre prédéfini, la conceptualisation devient complètement dépendante du résultat précédent. Néanmoins, selon *L'Homme* (2004), *Layer Cake* permet de découvrir les notions importantes du domaine, et leurs dénominations restent raisonnables.

Mais, pour nous, un travail préliminaire avec des experts du domaine est nécessaire afin d'amorcer le travail à partir des notions les plus importantes ; la fouille de textes et l'analyse linguistique sont appelées à faciliter ce travail.

4.2 La terminologie, une sous-langue spécialisée

L'ingénierie des ontologies s'intéresse particulièrement à la *terminologie* puisque l'ontologie de domaine est construite à partir des termes et, plus largement, du lexique de domaine.

La *terminologie* vise à fournir une norme commune pour le travail terminologique. C'est une discipline utilisant plusieurs théories et plusieurs méthodes. Ses recommandations sont prises en compte par l'Organisation Internationale de Normalisation (ISO), et en particulier par le Comité Technique 37 d'ISO (qui prépare les standards et autres documents concernant la méthodologie et les principes de travail sur la terminologie et

les ressources linguistiques). Les origines de ces normes proviennent du travail d'Eugene Wüster et de l'école terminologique de Vienne qu'il a animée, *Foo* (2012).

Dans *Bourigault et al.* (2001) les auteurs distinguent trois catégories de tâches qui sont propres au domaine de la *terminologie computationnelle* :

1. identification et filtrage automatique des candidats terminologiques ;
2. regroupement des variations et synonymes ;
3. dépistage des relations qui les lient.

Ces problèmes sont donc proches de ceux qu'on vise à résoudre pour l'apprentissage des ontologies.

La terminologie d'un domaine est considérée comme un sous-langage inscrit dans la langue naturelle et caractérisé par deux [1] aspects, *L'Homme* (2004), *Neveu* (2008), *Foo* (2012) :

1. L'aspect cognitif qui relie les formes linguistiques (les termes eux-mêmes) et leur contenu conceptuel, i.e. les référents dans le monde réel ;
2. L'aspect linguistique qui donne les règles syntaxiques auxquelles les termes sont assujettis.

Les méthodes que nous allons discuter portent sur ces deux dimensions.

Quelles que soient leurs cibles (termes, concepts, relations), les méthodes d'apprentissage peuvent être réparties en trois catégories : statistiques, linguistiques et hybrides, *Biemann* (2005). Mais toutes les méthodes d'apprentissage utilisent des modèles probabilistes basées sur des mesures statistiques. De ce fait, lorsqu'on dit « les méthodes linguistiques » on veut souligner que des informations linguistiques additionnelles y sont prises en compte. Par exemple, tous les mots sont étiquetés par des balises indiquant leur partie du discours, c'est à dire leur fonction syntaxique dans la phrase. On va qualifier de « méthodes statistiques » celles qui s'appuient uniquement sur les textes bruts pour y découvrir les entités recherchées.

4.3 Extraction des termes

Dans cette section, nous présentons les méthodes d'extraction des termes du corpus textuel, qu'elles utilisent, ou non, des ressources complémentaires.

Les termes peuvent être constitués de mots isolés, ou de suites des mots. Dans les domaines techniques la plupart des termes sont des termes composés (ou multi-mots).

1. Ici nous ne prendrons pas en considération l'aspect communicatif qui examine l'utilisation des termes par des gens, leur évolution et transformation.

Par exemple, dans *Orobinska et al.* (2013) les auteurs ont montré que les termes de la Sécurité Radiologique comprennent en moyenne 3 unités lexicales.

En linguistique, les termes sont des combinaisons syntaxiques hiérarchisées de mots, dénommés *syntagmes terminologiques*, ou *synapsies* (*Cabré et al.* (1998)). Cela veut dire que, pour l'acquisition de termes, les méthodes doivent être plus élaborées que celles qu'on utilise pour d'autres tâches de Data Mining comme l'extraction des thématiques (Topic Extraction) ou tous les traits lexicaux, mots, lemmes ou stemmes, peuvent être considérés comme composants d'une thématique *Gaussier et Yvon* (2011).

Notons une autre différence entre les méthodes d'extraction des termes du corpus textuel et celles de DataMining : si on veut pouvoir extraire des termes composés, il ne faut pas utiliser certaines étapes traditionnelles de pré-traitement des textes comme le coupage des terminaisons (stemming) ou l'élimination des « mots vides ».

4.3.1 Méthodes basées sur la fréquence

Dans ces méthodes, on fait l'hypothèse que les fréquences d'occurrence des termes du domaine sont particulièrement élevées.

Nous mentionnons ici huit caractéristiques utiles pour la sélection des termes candidats. La table 4.1 récapitule leurs formules de calcul.

Pour les décrire nous allons utiliser les notations suivantes :

- $TF(w)$ est la fréquence d'un mot ou d'un syntagme dans le corpus ;
- $DF(w)$ est le nombre de documents qui contiennent le mot ou le syntagme *w* au moins une fois ;
- $TF(w|d)$ est la fréquence du mot ou du syntagme *w* dans le document *d* ;
- $|W|$ est le nombre total des mots ou des syntagmes dans le corpus ;
- $|D|$ est le nombre total des documents dans le corpus ;
- $|W_d|$ est nombre des mots ou des syntagmes dans le document *d*.

Les deux premières caractéristiques sont la fréquence de terme (TF) et la présence/absence de terme dans un document ou dans tous les documents du corpus (DF). Ces mesures sont utilisées pour le calcul de caractéristiques plus complexes.

Term Frequency – Inversed Document Frequency (TF-IDF) favorise les mots ou les syntagmes qu'on rencontre souvent dans un sous-ensemble limité de documents du corpus.

Term Frequency – Residual Inverse Document Frequency (TF-RIDF) est une modification de la mesure précédente qui utilise la loi de Poisson en faisant l'hypothèse que la distribution de termes recherchés, à la différence des mots communs, ne correspond pas à celle de Poisson, *Church et Gale* (1995).

Domain Consensus (DC) est basée sur le calcul de l'entropie ; cette caractéristique favorise les mots ou les syntagmes plus fréquents dans certains documents du corpus, *Sclano et Velardi* (2007).

Term Contribution (TC) pénalise au contraire les mots fréquents qui sont distribués uniformément dans les documents du corpus *Liu et al.* (2005).

Term Variance Quality (TVQ) pénalise les mots et syntagmes qui ne se présentent qu'une seule fois dans la majorité des documents du corpus ; TVQ correspond à la variance de la fréquence du terme parmi les documents où le mot (ou le syntagme) est rencontré au moins une fois, *Dhillon et al.* (2003). Les mots ou syntagmes qui sont présents dans peu de documents de corpus, ou qui sont distribués uniformément parmi ces documents, ont des valeurs basses pour ***Term Variance***, *Liu et al.* (2005).

	Formule
TF-IDF	$TF(w) \times log\frac{\|D\|}{DF(w)}$
TF-RIDF	$TF(w) \times \left(log\frac{\|D\|}{DF(w)} - \left(-log\left(1 - e^{-\frac{TF(w)}{\|D\|}}\right)\right)\right)$
Domain Consensus	$-\sum_{d \in D}\left(\frac{TF(w\|d)}{\|W_d\|} \times log\frac{TF(w\|d)}{\|W_d\|}\right)$
Term Contribution	$\sum_{d_i,d_j \in D : d_i \neq d_j}\left(TF(w\|d_i) \times log\frac{\|D\|}{DF(w)}\right) \times \left(TF(w\|d_j) \times log\frac{\|D\|}{DF(w)}\right)$
Term Variance	$\sum_{d \in D}\left(TF(w\|d) - \frac{TF(w)}{\|D\|}\right)$
Term Variance Quality	$\sum_{d \in D} TF(w\|d)^2 - \frac{1}{\|D\|}\left(\sum_{d \in D} TF(w\|d)\right)^2$

Table 4.1: Récapitulatif des caractéristiques fréquentielles pour l'extraction des termes.

4.3.2 Méthodes basées sur les corpus contrastés

Ces méthodes sont basées sur la comparaison de la distribution de fréquences des mots dans le corpus spécialisé (ou corpus ciblé) et le corpus de référence (ou corpus contrasté) qui contient les textes plus généraux : on suppose que le comportement des termes dans les deux collections est sensiblement différents. Pour que les résultats obtenus soient corrects, il faut que les corpus aient des tailles comparables.

Pour mesurer ces caractéristiques, nous complétons la liste précédente :

- $TF_r(w)$ est la fréquence du mot ou du syntagme *w* dans le corpus de référence ;
- $DF_r(w)$ est le nombre de documents du ou des corpus de référence qui contiennent le mot ou le syntagme *w* ;
- $|D_r|$ est le nombre total des documents dans le (ou les) corpus de référence ;
- $|W_r|$ est le nombre total des mots ou des syntagmes dans le (ou les corpus) de référence ;
- $|C_w|$ est le nombre de corpus qui contiennent le mots ou le syntagme *w* ;
- $|C|$ est le nombre de corpus à comparer y compris les corpus contrastés.

La caractéristique de base dans ce groupes des mesures est ***Weirdness*** (l'anomalie), (cf. l'équation 4.1) qui est le ratio entre les fréquences relatives du mot *w* dans le corpus ciblé et le corpus contrasté, *Ahmad et al.* (1999) :

$$Weirdness(w) = \frac{TF(w)}{|W|} / \frac{TF_r(w)}{|W|_r} \tag{4.1}$$

L'indice ***Relevance*** (cf. l'équation 4.2) exploite le même principe : les mots et les locutions peu fréquents ont des valeurs faibles pour cette caractéristique. A l'inverse, *Relevance* a une valeur significativement élevée si l'expression est fréquente, mais pas trop, dans le corpus ciblé ni trop rare dans les textes du corpus contrasté, *Anselmo et al.* (2001).

$$Relevance(w) = 1 - \frac{1}{log_2\left(2 + \frac{TF(w) \times DF(W)}{TF_r(w)}\right)} \tag{4.2}$$

La mesure ***TF-IDF*** (cf. l'équation 4.3) fait partie des méthodes de ce groupe mais, ici, on calcule *IDF* sur le corpus contrasté. Sa nouvelle expression est ***KF-IDF*** qui favorise les mots et les locutions plus fréquentes dans le corpus ciblé que dans le corpus contrasté *Kurz et Xu* (2002) :

$$KF - IDF(w) = DF(w) \times log\left(\frac{|C|}{|C_w|} + 1\right) \tag{4.3}$$

Deux autres modifications de *TF-IDF* sont ***Contrastive Weight*** (cf. l'équation 4.4), *Basili et al.* (2001) et ***Discriminative Weight*** (cf. l'équation 4.5), *Wong et al.* (2007). La

première porte sur l'hypothèse que la distribution des mots communs doit être la même dans les deux corpus.

$$CW(w) = logTF(w) \times log\left(\frac{|W| + |W_r|}{TF(w) + TF_r}\right) \tag{4.4}$$

Discriminative Weight pénalise les mots fréquents du corpus ciblé s'ils sont plus spécifiques dans le corpus contrasté.

$$DW(w) = DP(w) \times DT(w) \tag{4.5}$$

où

$$DP(w) = log_{10}(TF(w) + 10) \times log_{10}\left(\frac{|W| + |W_r|}{TF(w) + TF_r(w)}\right) \tag{4.6}$$

et

$$DT(w) = log_2\left(\frac{TF(w) + 1}{TF_r(w) + 1} + 1\right) \tag{4.7}$$

Le nombre ***Loglikelihood*** (cf. l'équation 4.8) favorise les mots isolés, et les groupes de mots, dont la fréquence relative est plus élevée dans le corpus ciblé que dans le corpus contrasté, *Gelbukh et al.* (2010).

$$Loglikelihood(w) = 2 \times \left(TF(w) \times log\frac{TF(w)}{TF^{exp}(w)} + TF_r(w) \times log\frac{TF_r(w)}{TF^{exp}(w)}\right) \tag{4.8}$$

où

$$TF^{exp}(w) = |W| \times \frac{TF(w) + TF_r(w)}{|W| + |W_r|} \tag{4.9}$$

et

$$TF_r^{exp}(w) = |W_r| \times \frac{TF(w) + TF_r(w)}{|W| + |W_r|} \tag{4.10}$$

4.3.3 Méthodes basées sur la mesure d'association entre les mots

Ces méthodes traditionnelles utilisent les mesures d'association entre deux mots, ou entre deux fragments de texte, pour estimer la corrélation mutuelle de deux candidats

en termes. Elles sont destinées à l'extraction des termes composés de plus de deux mots ; elles ne sont pas applicables pour l'extraction de termes constitués de mots isolés. Notons que les mesures d'association sont utilisables par ailleurs pour la conceptualisation, ainsi que pour l'extraction des relations entre les concepts (cf. la section 4.4).

Dans les formules de cette section, nous allons utiliser les notations suivantes :

- $\bar{x}$ est tout mot différent de mot *x* ;
- $r(w)$ et $l(w)$ est le nombre total de mots différents qui se trouvent avant (à gauche) et après (à droite) du mot *w*.

L'***Information Mutuelle*** (cf. l'équation 4.11), proposée par S. Kullback en 1959 (*Kullback* (1997)), a été utilisée pour la première fois afin d'estimer la probabilité de la co-occurrence des mots dans *Church et Hanks* (1990). Elle continue à être très utilisée pour l'extraction de termes car elle est simple et claire.

$$MI(x,y) = |W| \times log\frac{TF(x,y)}{TF(x) \times TF(y)} \tag{4.11}$$

Si les mots *x* et *y* sont rencontrés dans le corpus indépendamment l'un de l'autre, leur information mutuelle est égale à zéro. Au contraire, plus la valeur de *MI* est élevée, plus ces mots sont liés sémantiquement. La formule ci-dessus n'est pas adaptée au cas où un des mots est absent. Plusieurs travaux ont proposé des modifications de l'équation 4.11 dans le but de résoudre ce problème. Par exemple, dans *Zhang et al.* (2009) les auteurs proposent la mesure dite « renforcée », ***Enhanced Mutual Information*** (cf. l'équation 4.12), définie comme le rapport entre la probabilité d'occurrence de la paire de mots et le produit des probabilités des occurrences individuelles de chaque mot *en dehors de la paire*.

$$EMI(x,y) = log\frac{TF(x,y)}{(TF(x) - TF(x,y))(TF(y) - TF(x,y))} \tag{4.12}$$

La mesure de l'information mutuelle normalisée ***Normalized MI*** a été proposée dans *Bouma* (2009) ; elle permet d'améliorer les résultats dans le cas de faibles fréquences d'apparition des mots (cf. l'équation 4.13) :

$$NormalizedMI(x,y) = \frac{log\frac{|W| \times TF(x,y)}{TF(x) \times TF(y)}}{-log\frac{TF(x,y)}{|W|}} \tag{4.13}$$

Cubic MI (cf. l'équation 4.14) est une autre modification de *MI* ; elle a été proposée par B. Daille, dans le même but, celui d'adapter la formule au cas des co-occurrences peu fréquentes *Daille* (1994).

$$CubicMI(x,y) = log\frac{TF(x,y)^3}{|W| \times TF(x) \times TF(y)} \tag{4.14}$$

B. Daille a aussi proposé la mesure ***True MI*** (cf. l'équation 4.15) :

$$TrueMI(x,y) = TF(x,y) \times log\frac{TF(x,y)}{TF(x) \times TF(y)} \tag{4.15}$$

Dans *Silva et al.* (1999), la ***Symmetrical Conditional Probability*** permet de vérifier la cohésion entre les mots *x* ety (cf. l'équation 4.16) :

$$SCP(x,y) = \frac{TF(x,y)^2}{TF(x) \times TF(y)} \tag{4.16}$$

Dans *Smadja et al.* (1996), les auteurs proposent d'utiliser le coefficient de Dice ***Coefficient de Dice***, initialement emprunté de la théorie de l'information pour la traduction automatique (cf. l'équation 4.17).

$$DC(x,y) = \frac{2 \times TF(x,y)}{TF(x) + TF(y)} \tag{4.17}$$

Dans *Kitamura et Matsumoto* (1996), la modification proposée par des auteurs a permis d'améliorer la qualité d'extraction de bigrams fréquents, et de les classer (cf. l'équation 4.18).

$$ModifiedDC(x,y) = log(TF(x,y)) \times \frac{2 \times TF(x,y)}{TF(x) + TF(y)} \tag{4.18}$$

Une nouvelle modification et généralisation de l'*Information Mutuelle* a été proposée dans *Park et al.* (2002). Les auteurs proposent une mesure pour calculer la cohésion de termes composés de plusieurs mots ($n > 2$) et assigner les valeurs les plus élevées aux termes dont la fréquence de co-occurrence est plus élevée (cf. l'équation 4.19).

$$GeneralizedDC(w,y) = \frac{r_w \times logTF(w) \times TF(w)}{TF(y)} \tag{4.19}$$

Dans *Daille* (1994), B. Daille a également adapté plusieurs mesures d'association, empruntées de la théorie de l'information, pour l'extraction des termes composés de deux mots ; parmi elles le ***Simple Matching Coefficient*** (SMC), le ***Coefficient de Kulczinsky*** et le ***Coefficient de Yule***.

Le ***SMC*** (cf. l'équation 4.20) additionne les nombres de co-occurrences, et d'occurrences disjointes, de deux mots (indépendamment l'une de l'autre).

$$SMC(x,y) = \frac{TF(x,y) + TF(\bar{x},\bar{y})}{|W|} \tag{4.20}$$

Le ***Coefficient de Kulczinsky*** (cf. l'équation 4.21) varie de 0 à 1 et, dans le cas où le mot *x* n'apparaît qu'avec le mot *x*, il est supérieur de 0.5 :

$$Kulczinsky\,Coefficient(x,y) = \frac{TF(x,y)}{2}\left(\frac{1}{TF(x)} + \frac{1}{TF(y)}\right) \tag{4.21}$$

Le ***Coefficient de Yule*** (cf. l'équation 4.22) varie de -1 à +1. Si les mots sont indépendants dans le corpus, il est égal à zéro. Pour les mots qui sont toujours ensemble sa valeur est égale à +1 et pour les mots qui ne sont jamais ensemble, il est égal à -1.

$$YUC(x,y) = \frac{TF(xy) \times TF(\bar{x}\bar{y}) - TF(x\bar{y}) \times TF(\bar{x}y)}{TF(xy) \times TF(\bar{x}\bar{y}) - TF(x\bar{y}) \times TF(\bar{x}y)} \tag{4.22}$$

Pour extraire des termes multi-mots, on peut aussi utiliser le ***Coefficient de Jaccard*** (cf. l'équation 4.23) qui est le ratio du nombre de co-occurrences de deux mots et de la somme des occurrences de chacun d'eux sans l'autre.

$$JaccardCoefficient(x,y) = \frac{TF(x,y)}{TF(x,\bar{y}) + TF(\bar{x},y)} \tag{4.23}$$

Les mesures de ***Chi-Square*** (cf. l'équation 4.24) et ***T-Score*** (cf. l'équation 4.25) sont utilisées pour mesurer le degré de dépendance de deux mots composant une locution (4.24).

$$\textit{Chi-Square}(x,y) = \frac{\left(TF(x,y) - \frac{TF(x)\times TF(y)}{|W|}\right)^2}{TF(x) \times TF(y)} \tag{4.24}$$

$$\textit{T-Score}(x,y) = \frac{TF(x,y) - \frac{TF(x)\times TF(y)}{|W|}}{\sqrt{TF(x,y)}} \tag{4.25}$$

Le ***Gravity Count*** (cf. l'équation 4.26), proposé dans *Daudaravicius et al.* (2004), est une mesure d'associativité qui permet d'estimer la fréquence à laquelle le deuxième mot dans la paire suit (i.e. apparaît à droite) le premier mot, ou l'inverse.

$$GravityCoun(x,y) = log\frac{(TF(x,y)r(x)}{TF(x)} + log\frac{(TF(x,y)l(y)}{TF(y)} \tag{4.26}$$

Le dernier outil de test paramétrique que nous allons citer pour la sélection des termes est ***LogLikelihood Ratio*** (cf. l'équation 4.27). Il compare les valeurs maximales de deux fonctions de vraisemblance selon les hypothèses que les deux mots forment, ou non, un syntagme, *Dunning* (1993).

$$LLR(xy) = 2 \times \Big(TF(xy) \times log \frac{|W| \times TF(xy)}{TF(x) \times TF(y)} + TF(x\bar{y}) \times log \frac{|W| \times TF(x\bar{y})}{TF(x) \times TF(\bar{y})} + \\ + TF(\bar{x}y) \times log \frac{|W| \times TF(\bar{x}y)}{TF(\bar{x}) \times TF(y)} + TF(\bar{x}\bar{y}) \times log \frac{|W| \times TF(\bar{x}\bar{y})}{TF(\bar{x}) \times TF(\bar{y})} \Big) \tag{4.27}$$

4.3.4 Méthodes basées sur le contexte

Ces méthodes sont basées sur le contexte permettant de repérer les termes. Ici, le contexte, gauche ou droit, est défini par des mots isolés ou des fragments composés de plusieurs mots qui servent de bornes entre lesquelles les termes se trouvent régulièrement. Avec cette approche on parle de « termes imbriqués » (*nested terms*). En pratique, les valeurs des bornes sont heuristiques ; elles peuvent être spécifiques au domaine ou bien plus générales. Ci-dessous, nous allons lister les mesures permettant de choisir le seuil de sélection des candidats en termes repérés dans le texte à l'aide du contexte.

Les notations utilisées dans les formules sont :

- P_w est l'ensemble de toutes les phrases qui contiennent le mot ou la locution *w* ;
- C_w est l'ensemble de tous les contextes du mot ou de la locution *w* ;
- $|W_c|$ est le nombre de mots contextuels du mot ou de la locution *w* ;
- $TF_w(c)$ est la fréquence du mot *c* en tant que mot contextuel du mot ou de la locution *w* ;
- $F_{max(w)}$ est la valeur maximale de la fréquence d'un n-gramme qui contient le mot ou la locution *w* (i.e. un contexte plus le mot ou la locution *w*) ;
- P_w^N est l'ensemble *N* des les n-grammes les plus fréquents qui contiennent le mot ou la locution *w* ;
- $l_{token}(w)$ et $r_{token}(w)$ sont les sommes des fréquences des mots qui, dans les textes, se trouvent à gauche, respectivement à droite, du mot ou de la locution *w* ;
- $l_{type}(w)$ et $r_{type}(w)$ est le nombre des mots contextuels uniques qui se trouvent juste à gauche, respectivement à droite, du mot ou de la locution *w* ;
- $|w|$ est le nombre des mots dans la locution *w*.

Un modèle exploitant le contexte a été présenté dans *Ananiadou* (1994). La caractéristique utilisée dans les méthodes de ce groupe est ***C-Value*** (cf. l'équation 4.28). Initialement, elle a été utilisée pour trouver les termes composés de plusieurs mots. Ainsi ***C-Value*** favorise les candidats-termes plus longs, et pénalise les locutions qui rentrent fréquemment à l'intérieur des groupes nominaux.

$$C-Value(w) = \begin{cases} log_2|w| \times \left(TF(w) - \frac{\sum_{p \in P_w} TF(p)}{P_w} \right), & \text{si la phrase enveloppe } w \\ log_2|w| \times TF(w), & \text{sinon} \end{cases} \quad (4.28)$$

Dans *Frantzi et Ananiadou* (1997) on trouve une généralisation de cette mesure pour les termes à un seul mot (cf. l'équation 4.29) :

$$\textit{C-Value(w)} = TF(w) - \frac{\sum_{p \in P_w} TF(p)}{|P_w|} \quad (4.29)$$

La modification la plus connue de ***C-Value*** est ***NC-Value*** (cf. l'équation 4.30) qui rajoute à ***(C-Value)*** l'information contextuelle ; l'équation 4.30 permet d'estimer le taux d'indépendance de chaque mot dans le texte ou, autrement dit, de vérifier si le mot *m* est nécessairement associé aux autres mots, *Frantzi et Ananiadou* (1997).

$$\text{NC-Value}(w) = \frac{1}{|W|} \times MC - Value(w) \times cweight(w), \text{ où} \quad (4.30)$$

$$cweight(w) = \sum_{e \in C_w} weight(c) + 1 \text{ et } weight(c) = \frac{1}{2} \left(\frac{|W_c|}{|W|} + \frac{\sum_{e \in W_C} TF(e)}{TF(c)} \right) \quad (4.31)$$

Dans *Frantzi et al.* (2000) les auteurs proposent une autre variation de ***NC-Value*** (4.28).

$$\textit{NC-Value(w)} = 0.8 \times \textit{C-Value(w)} + 0.2 \times \sum_{c \in C_w} TF(c) \quad (4.32)$$

Le ***Insideness*** (cf. l'équation 4.33) et le ***SumN*** (cf. l'équation 4.34), *Nokel et al.* (2012), sont des mesures envisageant les mots et les locution dans le contexte des phrases (sentences) qui les entourent.

D'autres mesures considèrent les mots et les locutions dans le contexte des phrases qui les entourent, par exemple ***Insideness*** (4.33) et ***SumN*** (4.34), *Nokel et al.* (2012).

$$Insideness(w) = \frac{TF_{max}(W)}{TF(w)} \quad (4.33)$$

$$SumN(w) = \frac{\sum_{p \in P_w^N} TF(p)}{N \times TF(w)} \quad (4.34)$$

Le ***Insideness*** permet de trouver les parties (fragments) de vrais termes, tandis que ***SumN*** permet de vérifier si un mot ou une locution est utile pour la construction du lexique de domaine.

Voici encore d'autres mesures construites sur l'hypothèse que certains mots sont utilisés plus souvent dans des unités terminologiques. On mesure l'accroissement de la probabilité que les fragments contenant ces unités soient elles-mêmes des termes. *Nakagawa et Mori* (2003) proposent les deux formules : ***Token-LR*** (cf. l'équation 4.35) et ***Type-LR*** (cf. l'équation 4.36).

$$\textit{Token-LR(w)} = \sqrt{l_{token}(w) \times r_{token}(w)} \tag{4.35}$$

$$\textit{Type-LR(w)} = \sqrt{l_{type} \times r_{type}} \tag{4.36}$$

Mais ces deux mesures ne considèrent que les mots contextuels, sans prendre en compte les candidats-termes eux-mêmes. Dans le même ouvrage, afin de pallier cet inconvénient, a été proposé la mesure FLR, qui a également deux variantes : ***Token-FLR*** (cf. l'équation 4.37) et ***Type-FLR*** (cf. l'équation 4.38).

$$\textit{Token-FLR(w)} = TF(w) \times \textit{Token-LR(w)} \tag{4.37}$$

$$\textit{Type-FLR(w)} = TF(w) \times \textit{Type-LR(w)} \tag{4.38}$$

4.3.5 Méthodes basées sur les thématiques (topic-based)

Ces méthodes utilisent les mesures basées sur ce qu'on appelle des thématiques (sujets, topics) ; l'idée sous-jacente est que chaque texte peut être décrit par un ensemble de mots caractérisant son sujet ; à l'origine c'était une méthode de classification non-supervisée de textes. Actuellement l'extraction des thématiques est un domaine à part entière d'ingénierie des connaissances et elle dispose de ses propres méthodes. Chaque thématique est présentée comme une liste ordonnée de mots fréquemment cooccurrents.

Pour la construction de thématiques, on distingue deux types de modèle : probabiliste et non probabiliste. Les *modèles probabilistes* représentent les textes comme un mélange de thématiques, chaque thématique étant considérée comme une distribution de probabilité sur les mots : LDA (Latent Dirichlet Analysis), PLSI (Probabilistic Latent Semantic Analysis), *Blei et Lafferty* (2006).

Les *modèles non probabilistes* sont basés sur les méthodes d'agrégation (clustering) tels que K-Means, agrégation hiérarchique (hierarchical agglomerative clustering) etc.,

Rizoiu et Velcin (2011). Ces méthodes nous suggèrent que les termes du domaine pourraient correspondre à certaines sous-thématiques. Dans *Bolshakova et al.* (2013) les auteurs ont démontré que l'algorithme NMF (Non-Negative Matrix Factorization) avec la minimisation de la divergence de Kullback–Leibler (KL Divergence Minimization) est le meilleur modèle pour l'extraction de termes. En résumé, étant donné une matrice non-négative *V* termes – documents, l'algorithme essaie de la décomposer en produit de deux matrices non-négatives (une matrice "termes – document" *W*, et une matrice "thématique – document" *H*) , telles que $V = W \times H$.

La diversité et l'abondance des méthodes d'extraction de termes témoignent qu'il n'existe pas « une solution meilleure » pour tous les cas. La prise de décision dépend de nombreux facteurs : le domaine, les textes, le type d'information qu'on cherche etc. Dans notre cas nous avons choisi l'information mutuelle et l'indice de Jaccard.

4.4 Extraction des relations

4.4.1 Extraction des relations – une tâche à plusieurs niveaux

Selon *Bachimont* (2000), la définition des relations entre les éléments de l'ontologie peut être envisagée sous plusieurs angles, linguistique, sémantique et ontologique, chacun contribuant à interprétation des constituants de l'ontologie et de leurs relations.

La dernière étape de l'apprentissage ontologique est l'implémentation des résultats dans un langage formel.

Au niveau sémantique, les relations entre les concepts ainsi que l'assignation des propriétés aux concepts se présentent par le biais des prédicats. Ce fait implique la nécessité, au préalable, de les répertorier, classifier et structurer (cf. les sections 2.4 et 2.5 du chapitre 2).

On rappelle qu'en linguistique (*Huddleston et Pullum* (2005)) :

> *…un prédicat décrit généralement une propriété de la personne ou de la chose visée par le sujet, ou bien décrit une situation dans laquelle cette personne ou chose joue un certain rôle dans les clauses élémentaires décrivant une action ; le sujet indique normalement l'acteur, la personne ou la chose qui exécute l'action, alors que le prédicat décrit l'action*.

Il faut noter que chaque langue naturelle a plusieurs moyens, syntaxiques et lexicaux, pour exprimer la même relation ; par exemple, les phrases « *les effets nocifs de la radio-exposition* » et « *la radio-exposition se caractérise par des effets nocifs* » expriment la même relation d'agrégation du type « *Objet* $\leftrightarrow$ *Propriété(s)* » (cf. Table 4.2). Pourtant, dans le premier cas, elle est mise en œuvre par une préposition et un article défini (« *de la* ») ; dans le second cas, elle est mise en œuvre par le verbe pronominal « *se caractériser* » au présent.

Finalement, il est difficile de définir les frontières entre les différentes approches d'extraction des relations car elles sont complémentaires et se renforcent réciproquement.

4.4.2 Classification des méthodes d'extraction de relations

On utilise deux groupes de méthodes d'apprentissage pour aboutir à des résultats satisfaisants dans l'extraction des relations : celles qui analysent la distribution des co-occurrences des mots (cf. section 4.3.3) et celles utilisant des patrons lexico-syntaxiques (ex. « TELS QUE + {Liste des groupes nominaux} ») ; ces dernières sont liées aux méthodes basées sur le contexte, cf. la section 4.3.4.

Les méthodes exploitant la distribution des co-occurrences des unités lexicales (des mots et des locutions) ont deux branches :

- Considérer uniquement la co-occurrence des mots dans le même énoncé, sans considérer d'autres propriétés lexicales ou sémantiques ; c'est la méthode de « *window-based approache* » (méthode de la fenêtre glissante) ;
- Utiliser aussi les structures syntaxiques et les rôles thématiques des arguments dans ces structures, *GuoDong et al.* (2005).

Les méthodes basées sur l'analyse distributionnelle ont l'avantage de repérer la présence d'un lien statistiquement significatif entre les termes, mais il reste à distinguer les relations syntagmatiques (comme les collocations où les relations nom-verbe) et les relations paradigmatiques (comme la synonymie, hyperonymie), *Fabre et D.* (2006), *Morlane-Hondère et Fabre* (2012).

Les méthodes basées sur les patrons lexico-syntaxiques ont été introduites par M. Hearst *Hearst* (1992). Elles ont permis d'améliorer la précision des résultats mais elle ne permettent de construire qu'un seul type de relations, *Perinet et Hamon* (2013), et chacune d'elles sont relativement rares dans les corpus. Néanmoins ces méthodes, basées sur les patrons pour l'extraction des relations non-taxonomiques, continuent à se développer, *Barriere* (2008), *Budanitsky et Hirst* (2001).

Le perfectionnement de ces méthodes est stimulé par les progrès des différents outils d'analyse linguistique : l'analyse morphologique permettant de définir la partie du discours (*part of speach*) de chaque mot (*Falk et al.* (2014)) et l'analyse des dépendances syntaxiques dans les phrases (*de Marneffe et al.* (2006)).

Les démarches de type linguistique qui permettent d'améliorer les résultats sont l'analyse syntaxique et morphologique, la lemmatisations et la racinisation (stemming) des mots. Elles facilitent la récupération, par des moyens différents, des fragments des phrases dans lesquelles les candidats-termes sont explicitement liés. Ainsi, on a vu récemment apparaitre des travaux utilisant l'apprentissage approfondi (deep learning), basés sur l'analyse syntaxique des énoncés et les calculs de dépendances entre les mots ; citons *Daojian et al.* (2014). Notons que ces méthodes peuvent aussi être classées dans la catégorie de l'analyse distributionnelle.

Jusqu'à tout récemment, le seul moyen d'apprentissage pour l'identification et la classification des relations restait l'utilisation d'heuristiques très coûteuses en temps et nécessitant la collaboration de linguistes spécialisés. Mais à partir des années 2000, les ressources lexiques et lexicographiques sont de plus en plus utilisées pour l'identification des relations et leur caractérisation *Velardi et al.* (2013), *Specia et Motta* (2006), *Schutz et Buitelaar* (2005), *Ciaramita et al.* (2005).

Les méthodes qui ont influencé nos propres recherches sont décrites dans la section 4.4.4.

4.4.3 Classification des relations

Pour l'apprentissage des ontologies, les relations peuvent être classées selon deux aspects : sémantique et ontologique.

Selon la définition formelle (*Cimiano et al.* (2005), *Serra et al.* (2013)), l'ontologie peut être décrite comme suit :

$$O = (C, H, I, R, P, A)$$

Dans ce 6–tuplet sont listés les éléments suivants : $C = C_C \cup C_I$ est l'ensemble des classes, soit des concepts (C_C) dont chacun correspond à l'ensemble des entités (C_i) qui peuplent ce concept.

H (cf. l'équation 4.39), I (cf. l'équation 4.40) et R (l'équation 4.41) sont les types de relations ; P (4.42) est l'ensemble de propriétés des concepts et A (4.43) est l'ensemble des axiomes et règles qui contrôlent la cohérence de l'ontologie et permettent l'inférence de connaissances nouvelles.

$$H = \{kind\_of(c_1, c_2) \mid c_1 \in C_C,\ c_2 \in C_C\} \tag{4.39}$$

$$I = \{is\_a(c_1, c_2) \mid c_1 \in C_C \wedge c_2 \in C_C\} \tag{4.40}$$

$$R = \{rel\_k\ (c_1, c_2,\ \ldots\ c_n) \mid \forall i, c_i \in C\} \tag{4.41}$$

$$P = \{prop\_\mathsf{C}(c_k, datatype) \mid c_k \in C_C\} \wedge \{prop\_I(c_k, value) \mid c_k \in C_I\} \tag{4.42}$$

$$A = \{condition_x \Rightarrow conclution_y\ (c_1, c_2,\ \ldots\ c_n) \mid \forall j,\ c_j \in C_C\} \tag{4.43}$$

On distingue les deux types des relations entre les concepts : les relations hiérarchisées de subordination (ou subsomption) et les relations associatives (non-taxonomiques, non-subordonnées).

Les relations hiérarchisées, elles-même, se divisent en deux sous-catégories : *Kind-of* ou *Classe–Sous-Classe* (4.39), et *Is-a* ou *Classe–Entité* (4.40).

Lors de l'implémentation dans un langage formel, les relations qu'on a mises en évidence se voient assigner certains propriétés, telles que la transitivité, la symétrie, l'ordre, l'équivalence ; en même temps, on leur impose des restrictions.

Cet ensemble de types de relations ontologiques n'est pas suffisant pour modéliser de façon satisfaisante un domaine spécialisé, ni pour fournir des règles de détection des traits linguistiques des relations dans les textes.

Néanmoins, à notre connaissances, il n'existe pas pour le moment de classification universelle des relations prédicatives, bien que certaines relations soient unanimement acceptées : synonymie, antonymie, causalité, relations hiérarchiques telles que hyperonymie, méronymie, etc.

Une de premières systématisations de la terminologie en tant que sous-langue formelle a été réalisée par E. Wüster qui a proposé de distinguer les relations logiques et ontologiques. Dans chaque groupe de relations, il existe une hiérarchie. Les relations logique sont définies comme relation de ressemblance. Les relations ontologiques sont comprises comme des relations de contiguïté dans l'espace et le temps, *Wüster* (1985).

Beth Levin a accompli un travail important sur la classification lexico-sémantique des verbes anglais, basée sur l'analyse de leur structure argumentale (*Levin* (1993)), *Kipper et al.* (2007)). Gaston Gross a produit une classification équivalente pour le frainçais (*Gross* (1975)). La méthode propose de regrouper les objets en fonction de leurs rôles dans les schémas de prédicats (*Gross* (2008)).

Dans *Rosario et Hearst* (2001) les auteurs proposent 13 classes qui décrivent les relations sémantiques possibles dans un groupe nominal entre le nom central et ses dérivés. Dans *Girju et al.* (2005) les auteurs proposent le système de 35 classes pour les relations sémantiques dans les groupes nominaux.

Dans notre travail, nous nous intéressons principalement aux relations récurrentes entre les verbes et les noms, car ce sont les verbes qui jouent le plus souvent le rôle de prédicats liant les concepts dans les structures argumentales.

Dans *Nayhanova* (2008) l'auteur propose une typologie des relations liant les termes, cf. Table 4.2, ainsi qu'une correspondance entre les relations ontologiques et les prédicats verbaux cf. Table 4.3.

Catégorie de relations	Groupe de relations	Relation	Type de notion
Qualitatifs	Hiérarchie	Kind-of (Is-a) Attribut ↔ Valeur d'attribut Invariant ↔ Variant	Abstrait et concret
	Agrégation	Intégralité ↔ Part (Part-of) Objet ↔ Propriété(s) Objet ↔ Localisation Niveau ↔ unité de niveau	Appartenance
	Fonctionnelle	Objet d'action ↔ Action ↔ Sujet d'action Cause ↔ Conséquence Condition ↔ Action Événement ↔ Action Aspect (État) ↔ Action Événement ↔ Aspect (État) Terme ↔ Synonyme Données ↔ Action Données ↔ Grandeurs	Processus
	Sémiotique	Terme ↔ Mode d'usage Terme ↔ Mode de représentation Terme ↔ Signe de terme	Le fond et la forme
Quantitatif	Équivalence	Terme ↔ Synonyme de terme	Équivalence et l'opposition
	Corrélation	Terme ↔ Corrélatif de terme	

Table 4.2: Types de relations entre les termes.

4.4.4 Les méthodes d'extraction des relations

Dans cette section, nous allons décrire plus en détail des méthodes dont le point commun est l'utilisation des verbes en tant que marqueurs des relations entre les concepts. Ces méthodes différent selon les pré-requis pour le traitement des données initiales, la profondeur de l'analyse linguistique, et les critères de sélection des relations-candidates.

Un des premiers algorithmes d'extraction automatique des relations causatives est décrit dans *Girju et Moldovan* (2002). Ces auteurs ont repris une méthode décrite en 1992 par C. Fillmore afin d'apprendre la structure du cadre de RISQUE (cf. la section 2.6.1.1, p. 27), *Fillmore et Atkins* (1992).

L'objectif principal des expériences de *Villaverde et al.* (2009) est de trouver le seuil de confiance pour sélectionner les verbes. Les ressources initiales sont : un corpus de textes

Groupe de relations	**Relation**	**Prédicat**	**Ensemble des valeurs de l'argument premier**
Hiérarchie	Kind-of (Is-a) Attribut ↔ Valeur d'attribut Invariant ↔ Variant	$PHier(a, x, y)$	{Classe, Type} {Catégorie,Valeur} {Invariant,Variant}
Agrégation	Intégralité ↔ Part (Part-of) Objet ↔ Propriété(s) Objet ↔ Localisation Niveau ↔ unité de niveau	$PAggr(a, x, y)$	{Classe,Kind}
Fonctionnelle	Objet d'action ↔ Action ↔ Sujet d'action Cause ↔ Conséquence Condition ↔ Action Événement ↔ Action Aspect (État) ↔ Action Événement ↔ Aspect (État) Terme ↔ Synonyme Données ↔ Action Données ↔ Grandeurs	$PFun(a, x, y, z)$	{Fonction, Cause, Condition, Événement, État, ÉtatObj, Outil, Date, Quantité }
Sémiotique	Terme ↔ Mode d'usage Terme ↔ Mode de représentation Terme ↔ Signe de terme	$PForm(a, x, y, z)$	{Expression, Représentation, Symbole}
Équivalence	Terme ↔ Synonyme de terme	$PEquiv(a, x, y)$	{Synonyme}
Corrélation	Terme ↔ Corrélat de terme	$PCor(a, x, y)$	{Corrélat, Opposé}

Table 4.3: Correspondance entre les relations ontologiques et les prédicats.

spécifiques au domaine, une liste de candidats concepts ou une hiérarchie de concepts. Le traitement du corpus élimine les *mots vides*. A la fin du prétraitement le texte se présente sous la forme d'un *sac de mots* « *bag-of-words* ». Les auteurs utilisent les synonymes des concepts initiaux trouvés dans WordNet et cherchent le verbe qui les lie. Les règles d'association sont basées sur l'algorithme d'Agrawal (*Agrawal et al.* (1993)). Le problème principal est la sélection des candidats verbes appropriés pour relier les concepts.

Stevenson (2004) utilise l'analyse syntaxique pour identifier dans un corpus les patrons du type SVO (Sujet-Prédicat-Objet). Le mécanisme d'apprentissage permet de trouver de nouveaux partons, similaires aux patrons dont on a appris la pertinence. L'auteur utilise une mesure de similarité basée sur la valeur de contenu d'information (Information Content). Pour commencer le processus d'apprentissage de nouveaux patrons, il faut avoir des patrons déjà acceptés avec des scores égaux à 1. Une contrainte importante est

que les entités nommées doivent avoir été annotées préalablement, ce qui est impossible à faire de façon automatique. De plus, la méthode permet bien de former des *topics*, mais pas de distinguer les concepts au sens d'une ontologie.

Dans *Kavalec et al.* (2004), les auteurs proposent une méthode d'apprentissage des règles d'association basée sur les paramètres probabilistes d'occurrence des verbes en association avec les deux concepts donnés. Ils supposent que l'association existe dans des limites d'une distance spatiale entre les concepts. Ils sélectionnent les verbes dont la probabilité conditionnelle d'association est supérieure à un seuil empirique. Mais les verbes fréquents sont souvent très génériques, ce qui peut engendrer une redondance des résultats quand on utilise ces verbes comme marqueurs pour chercher de nouvelles entités. La méthode a été étendue par le rajout de la direction de la relation entre les concepts (*Punuru et Chen* (2012)).

Dans *Serra et Girardi* (2011) et *Serra et al.* (2013) les auteurs proposent un algorithme d'apprentissage supervisé de relations non taxonomiques à partir d'un corpus (LNTRO – Learning Non-Taxonomic Relationships of Ontologies). Les auteurs soulignent les deux aspects fondamentaux en *Ontology Learning* : - la qualité et la disponibilité des connaissances préalables sur le domaine ; - le format des données d'où les connaissances doivent être extraites. A la différence des méthodes de cette catégorie, les auteurs proposent d'utiliser la distance maximale entre les concepts pour assurer leur filiation importante par biais d'un verbe. Mais les paramètres de la méthode doivent chaque fois être réglés par un spécialiste.

Makki et al. (2008) proposent une méthode semi-automatique de peuplement d'une ontologie, basée sur l'application de règles de reconnaissance des instances ; les relations sémantiques entre les concepts y sont prédéfinies à l'aide des experts qui doivent proposer une ontologie générique. Les relations entre les concepts sont exprimées par les verbes. Les règles de reconnaissance se présentent sous forme de patrons du type (Instance-a ,Vab ; Instance-b) où les « Instances » sont les groupes nominaux à droite et à gauche de verbe. La relation inclut tous les verbes-synonymes proposés par WordNet, ce qui fait diminuer la précision des résultats car, parmi les synonymes, peuvent se trouver des verbes qui ont des sens spécifiques, différents de ceux utilisés dans le domaine qu'on étudie.

4.5 Conceptualisation

Dans le domaine de l'apprentissage des ontologies, la conceptualisation peut être interprétée de deux façons. D'une part, il s'agit du processus d'acquisition, dans le corpus, des candidats-termes et de leurs liens lexicaux. D'autre part, la conceptualisation se comprend comme le résultat de la transformation des unités recueillies en concepts et relations ontologiques. Ces démarches sont difficiles à formaliser. Pour cette raison, *Mondary et al.* (2008) soulignent la nécessité de travailler étroitement avec des experts du domaine dans toutes les étapes de la construction de l'ontologie. La même idée est

partagée par *Héon et al.* (2009) et d'autres chercheurs qui adaptent la méthodologie de modélisation par objets typés, MOT (*Paquette* (1996b)), pour faciliter la production de la spécification formelle (en OWL, par exemple) en partant du modèle semi-formel obtenu lors de l'élicitation des préférences de l'expert (c'est à dire quand on aide l'expert à formaliser ses connaissances). Ce modèle semi-formel est plus facilement lisible par les spécialistes qui ne sont pas informaticiens.

Finalement, la conceptualisation correspond au modèle du domaine où les concepts [2] sont compris comme des étiquettes linguistiques pour évoquer certaines notions. Dans le cas d'un domaine spécialisé, la terminologie du domaine restreint ces moyens, en faisant de chaque terme une entité conceptuelle associée à une certaine notion invariante *Roche* (2005).

Donc, afin de faciliter la conceptualisation, il faut se concentrer sur les méthodes de regroupement des unités lexicales sémantiquement proches, et sur la désambiguïsation du sens des mots.

Une tendance récente est d'utiliser les arbres de dépendances tirés des théories grammaticales. On appelle les méthodes de ce groupe « *deep learning methods* ».

On distingue deux théories qui exploitent les dépendances syntaxiques : celle de TAG (*Tree Adjointing Grammar*), dont l'hypothèse est que toute dépendance syntaxique est exprimée localement par un arbre élémentaire unique *Frank* (2004), et DRT (*Discourse Representation Structures*) *Cimiano et al.* (2014).

Dans le premier cas, on part du sujet de la phrase comme tête de l'arbre pour déployer la structure arborescente. Dans le deuxième cas, on part du verbe.

Le problème de l'ambiguïté se pose lors de l'étiquetage des concepts. Les méthodes qui proposent une solution à l'ambiguïté utilisent des ressources lexicales extérieures (cf. 2.6). Par exemple, dans *Flati et Navigli* (2013), *Velardi et al.* (2013), *Moro et al.* (2014) les auteurs proposent d'utiliser des ressources lexicales multilingues, notamment BabelNet, pour les différents aspects de traitement linguistique. En particulier, ils proposent une approche unifiée multilingue, basée sur un graphe ; ils obtiennent un liage des entités et une désambiguïsation du sens de mots basée sur une identification floue de significations combinée à l'heuristique la plus dense d'un sous-graphe ; cette approche permet de sélectionner des interprétations sémantiques très cohérentes.

Dans *Fabre et D.* (2006) les auteurs présentent des méthodes utilisant WordNet pour repérer la synonymie entre les mots. Mais, pour nous, WOLF qui est l'analogue français de WordNet, n'est pas suffisamment complet pour les domaines scientifiques et techniques. Nous avons utilisé une autre ressource en ligne : CRISCO, le Dictionnaire Électronique des Synonymes (cf. la section 5.4.3 du chapitre 5).

2. cf. la définition d'un concept dans les contextes différents dans Table 2.1 de la section 2.3.1

4.6 Conclusion du chapitre 4

Dans ce chapitre, nous avons présenté un état de l'art sur les méthodes destinées à l'extraction des termes, des relations et de la conceptualisation.

Les méthodes de sélection de termes sont nombreuses et variées et il est difficile de choisir « la meilleure méthode » car les résultats dépendent de plusieurs facteurs tels que le domaine étudié et la définition des termes (sont-ils toujours des mots isolés, ou peuvent-ils être des multi-mots ?). L'analyse syntaxique permet de rajouter des informations additionnables telles que les parties du discours, par exemple, et d'enrichir l'exploitation du corpus afin d'améliorer les résultats. Mais cela complique les calculs.

Un problème majeur auquel on est confronté lors de l'extraction des relations est la nécessité de les classifier correctement. En règle générale, il faut faire le choix entre les méthodes qui permettent l'extraction d'un seul type de relation, et les méthodes qui cherchent des chaînes des mots supposés être unis par certaines relations à définir. La solution de ce problème se trouve dans l'utilisation de modèles conceptuels, sur lesquels on s'est préalablement accordé avec des experts du domaine.

Chapitre 5

Nos expérimentations

5.1 Introduction

Dans ce chapitre nous décrivons les résultats obtenus au cours de nos recherches. Ils ont été présentés dans des revues et des conférences, notamment IMS-2012, TIA-2013, TOTH-2014, Eastern-European Journal of Eenterprise Technologies, Bionica Intellecta (Бионика интеллекта), Herald of the NTU « KhPI » (Вестник НТУ « ХПИ »).

Au départ, l'objectif principal de nos recherches était de :

- Réaliser un état de l'art des différentes approches et méthodes de construction d'une ontologie de domaine ;
- Explorer les tendances récentes pour l'apprentissage d'une ontologie de domaine ;
- Formuler des recommandations sur les étapes et méthodes d'apprentissage d'ontologie, et les vérifier expérimentalement.

Au final, nous avons proposé une approche complète permettant de créer une ontologie de domaine « de A à Z » en utilisant largement les ressources linguistiques en ligne telles que les dictionnaires des synonymes. La création se réalise en deux temps :
– la première étape installe le noyau d'ontologie ;
– la deuxième étape effectue l'extension et l'instanciation de l'ontologie.

Pour la construction de notre ontologie, nous adoptons l'approche terminologique ; le résultat recherché est l'organisation du vocabulaire spécialisé, sous forme d'une hiérarchie de termes groupés autour des concepts qu'ils représentent, ainsi que leurs relations sémantiques. Les concepts, sont décrits via des listes de termes les plus génériques possible. De plus, nous devons formuler les règles de reconnaissance permettant l'extension du vocabulaire.

Avant de décrire en détail notre approche et ses méthodes, nous rappelons deux principes de départ formulés dans le chapitre 3.

Premièrement, nous sommes convaincus qu'il faut organiser un processus guidé avec, à l'issue de chaque étape, une validation des résultats intermédiaires par des experts. Ceci est proche de la stratégie de la plateforme TERMINAE (cf. la section 3.5.1, p. 45).

Deuxièmement, les concepts supérieurs n'étant généralement pas très nombreux, entre dix et vingt en moyenne, *Velardi et al.* (2013), il vaut mieux utiliser au début un modèle construit avec l'expert du domaine. Autrement dit, la conceptualisation anticipée devient une étape principal dans tout le processus et les efforts des développeurs doivent se concentrer sur l'élaboration de méthodes (semi-)automatiques de formalisation des connaissances. Ceci permet d'éviter l'accumulation progressive de faux résultats intermédiaires ; elles sont à la base de nos méthodes.

Le chapitre est organisé comme suit : la section 5.2 présente le domaine d'application choisi : la sécurité radiologique. La section 5.3 décrit les principes généraux de notre approche ; la section 5.4 discute du choix d'un dictionnaire en ligne, puis présente les démarches entreprises pour l'installation de nos deux corpus et leur pré-traitement syntaxique. La section 5.5 récapitule l'enchainement de nos méthodes et présente les résultats d'expérimentations. La section 5.6 conclut le chapitre.

5.2 Domaine d'application : la radioprotection

5.2.1 Le choix du domaine

Nous avons choisi le domaine de la sécurité radiologique (ou radioprotection, RP) comme domaine d'application pour tester les méthodes proposées. Le choix s'explique par plusieurs raisons :

- La radioprotection est un des domaines dont dépend la santé des personnes et l'état de l'environnement.
- Le domaine de la radioprotection, étant à la fois scientifique et technique, il utilise une terminologie riche et spécifique.
- De très nombreux textes officiels, éditée par l'AIEA (Agence Internationale pour l'Énergie Atomique) et les commissions nationales sur la Radio-Protection, sont disponibles en plusieurs langues. Cela nous a permis de créer les corpus parallèles et de comparer la qualité des résultats des expérimentations sur les langues française et russe qui ont des structures grammaticales différentes.
- La possibilité de consulter des experts du domaine qui travaillent au Centre National Scientifique « Institut de la Métrologie » de Kharkov en Ukraine.

L'objectif principal de la radioprotection (ou sécurité radiologique) est la minimisation des risques liés à l'exposition aux rayonnements ionisants des personnes et de l'environnement. La gestion des risques demande des actions coordonnées de spécialistes de nombreux domaines tels que la médecine, la physique, l'ingénierie, le droit, la technologie, etc. (cf. la figure 5.1 tirée du *Wilson et al.* (1997)). La réglementation les risques radiologiques est publiée dans les séries d'ISO et dans les Normes de sûreté de l'AIEA.

5.2.2 Les principaux aspects de la radioprotection

La prise des décisions dépendant d'un ensemble de facteurs de nature très variée, il est primordial de construire et de maintenir une base de connaissances liée à l'ontologie du domaine (ou à plusieurs ontologies de sous-domaines spécifiques).

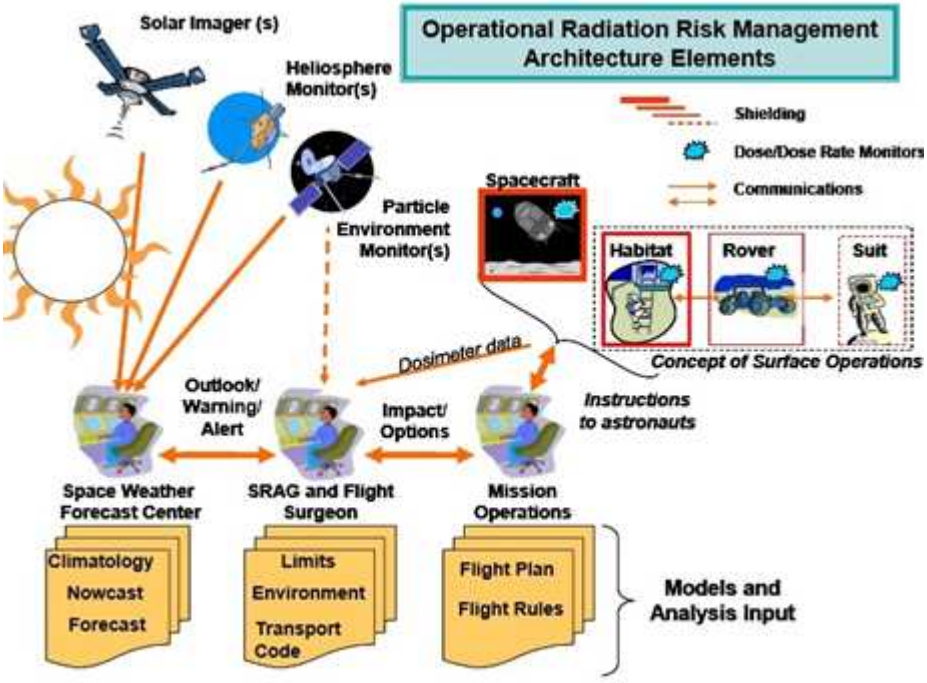

Figure 5.1: Éléments du système de gestion des risques radiologiques.

On distingue trois catégories « d'objets » visés par les mesures de sécurité :

1. Le personnel, c'est à dire les gens qui sont en contact avec des matériaux radioactifs dans leur activité professionnelle.
2. La population qui peut être exposée aux rayonnements.
3. Les lieux et locaux dans lesquels travaillent ces personnes.

Pour le personnel, les outils de sécurisation sont :

- Les restrictions du travail près des sources radioactives selon l'âge, le sexe, l'état de santé, les doses précédentes d'exposition aux rayonnements, etc. ;
- Le respect du code de sécurité et des niveaux d'exposition maxima ;
- L'organisation de conditions de travail respectant les normes et règles de la sécurité radiologique ;
- L'utilisation d'équipements de protection individuelle ;
- L'organisation d'un système d'information pour le contrôle des niveaux de rayonnement dans l'environnement de travail et de vie ;
- L'adoption de mesures additionnelles de protection du personnel en cas d'augmentation de la radiation ou de menace d'accident.

Pour l'environnement (i.e. les lieux et les locaux) :

- La qualité du plan d'organisation des sites de radiation, de leur entretien et des conditions de mise hors service ;
- Le choix de l'emplacement de chaque site de radiation ;
- La protection physique des sources de radiation ;
- La délimitation de périmètres différenciés autour et à l'intérieur des sites les plus dangereux ;
- L'octroi de licences pour des activités liées au travail avec les sources radiatives et l'évaluation sanitaire et épidémiologique des produits et des procédés d'utilisation ;
- L'organisation d'un système de mesure et de contrôle de radiation ;
- La formation continue du personnel et les instruction à la population.

Pour la population :

- Le respect des normes de sécurité radiologique ;
- La fixation des quotas d'exposition en fonction de la source de radiation ;
- L'organisation d'un système d'information et de contrôle du niveau de radiation ;
- La planification des mesures de sécurité dans les conditions normales, et en cas de menace d'accident ;
- L'approvisionnement systématique et gratuit de la population en équipements de protection individuelle tels que les masques respiratoires, les masques à gaz, les doses d'iode, etc.

5.2.3 L'organisation de la collaboration avec un spécialiste du domaine

Les méthodes que nous proposons supposent une intervention humaine pour valider les résultats. Le spécialiste avec lequel j'ai été en contact tout au long de ma thèse est le

directeur du « Laboratoire Scientifique des Rayonnements Ionisants et de la Dosimétrie des Radiations » qui appartient à l'« Institut de Métrologie » du Centre National Scientifique d'Ukraine. Ce laboratoire est responsable des contrôle réguliers de l'équipement de mesure radiologique pour les centrales nucléaires, les mines, les hôpitaux, etc. Le laboratoire est habilité à la certification des nouveaux outils et il participe aux sessions de l'AIEA.

La première intervention de l'expert a porté sur les consignes pour constituer le corpus qui devait inclure des documents sur toutes les activités du domaine de la Radioprotection (cf. la section 5.4.1). Les autres consultations ont été données à chaque étape, en fonction de la disponibilité de l'expert.

5.3 Notre approche : cadre général

Dans la communauté des chercheurs et des informaticiens du domaine de l'ingénierie d'ontologie, il est maintenant reconnu que la construction d'une ontologie suppose plusieurs étapes. En règle générale, on commence par l'installation d'un noyau d'ontologie qui comprend soit la simple énumération des dénominations des concepts, soit, en plus, une hiérarchie de ces concepts. Le noyau d'ontologie sert à l'extraction des nouveaux candidats-termes. Après cette première étape, on définit les relations sémantiques entre les concepts. C'est, entre autre, la stratégie « Layer cake » (cf. la section 4.1, p. 61).

Partant du point de vue que la construction d'une ontologie se divise en deux grandes parties - la construction du noyau d'ontologie, puis son enrichissement par des termes nouveaux - nous proposons d'inclure l'extraction des relations sémantiques dès la première étape. Autrement dit nous proposons de donner la même importance aux concepts et aux relations qui les lient.

Ainsi, nous commençons par la conceptualisation anticipée du domaine de modélisation et, en même temps, la modélisation linguistique anticipée de corpus d'entrée (cf. la section 5.5.1, p. 93) et nous introduisons la notion de *cadre prédicatif* (défini ci-dessous). La définition, dès la première étape, des relations régissant la quêtes de nouveaux candidats-termes améliore les résultats de la deuxième étape.

Nous proposons le mode opératoire suivant :

1. En consultation avec les experts, définir une liste limitée de termes généraux et des catégories de relations sémantiques entre ces termes ;
2. En prenant chacun des termes précédents comme référence d'un concept, grouper autour d'eux ses synonymes pour constituer les classes sémantiques représentant les concepts à travers leurs extensions sémantiques (cf. la définition dans la section 5.5.1, p. 93) ;

3. Former le cadre prédicatif sous la forme de l'ensemble de classes lexico-sémantiques des verbes ;
4. Appliquer le cadre prédicatif pour l'extraction de nouveaux candidats-termes pour peupler l'ontologie.

Notons que l'ordre des items 2 et 3 est libre.

Éclaircissons la notion de *cadre prédicatif*. En partant de l'idée que toute structure se définit par les liens fonctionnels entre ses constituants, nous étendons la notion classique de cadre sémantique (cf. section 2.4.2.1) à l'ensemble des prédicats qui caractérisent la structure du modèle ; dans la projection linguistique, ce sont majoritairement les verbes.

Définition. Le *cadre prédicatif* est l'ensemble des classes sémantiques des verbes qui représentent les relations entre les concepts de l'ontologie au niveau linguistique.

Les figures 5.2 et 5.3 visualisent notre approche. La figure 5.2 illustre la connexion étroite entre l'extraction des concepts et la définition des relations sémantiques, ainsi que leur conditionnement réciproque. La figure 5.3 décrit les sous-processus, impliqués dans la construction de l'ontologie ; ils y sont regroupés en fonction de l'objectif à atteindre : l'extraction de nouveaux composants, l'enrichissement de l'ontologie, et son évaluation.

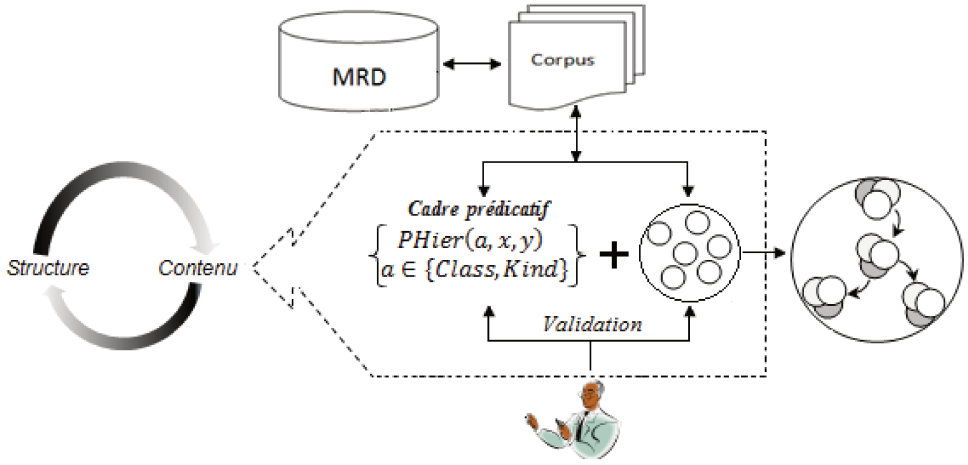

Figure 5.2: La construction du noyau de l'ontologie.

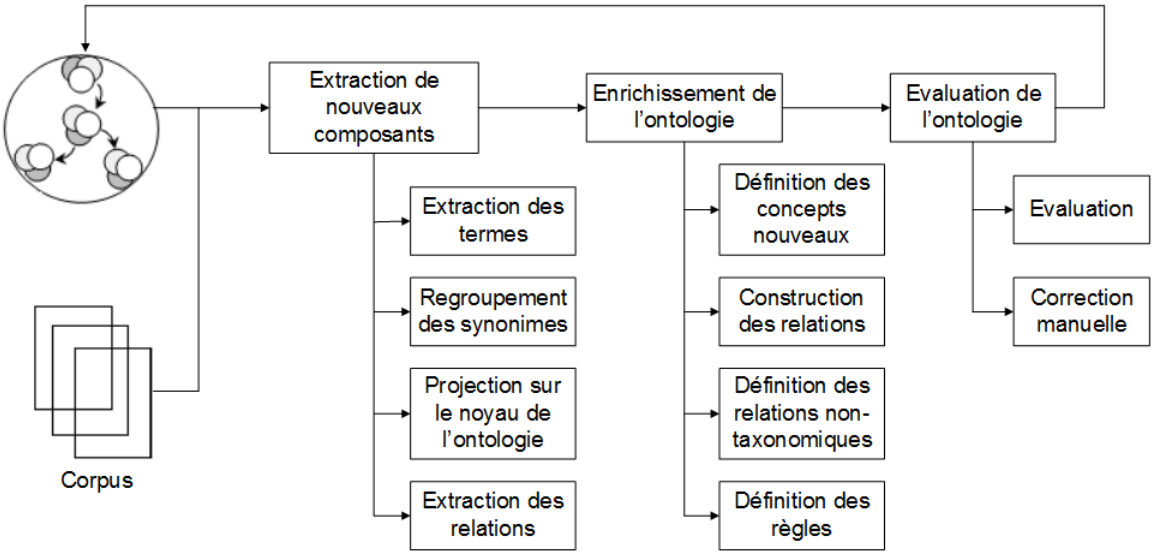

Figure 5.3: Flot de travail pour la construction de l'ontologie de domaine.

5.4 Acquisition des ressources

5.4.1 Installation des deux corpus

Le qualité du résultat final dépend beaucoup des ressources utilisées. Ces ressources sont variées, *Gómez-Pérez et Manzano-Macho* (2003), *Zhou* (2007), *Drumond et Girardi* (2008) : glossaires, dictionnaires ou thésaurus (spécialisés ou non), bases de données, textes. Parmi les ressources mentionnées, les textes sont considérés comme les plus importants parce qu'ils sont nombreux et facilement accessibles, et qu'ils contiennent, à l'état latent, les connaissances conceptuelles modélisant le domaine, les entités illustrant des phénomènes, les états, les processus, etc.

Les principales difficultés dans l'analyse et le traitement des textes sont : leur redondance par rapport à l'information pertinente qu'on y cherche, la diversité des formes morphologiques pour le même lemme et la diversité des formes lexicales pour la même notion ou concept. Ceci nécessite la mise en œuvre de méthodes d'analyse à la fois statistiques et linguistiques. C'est dans cette direction que se développe l'apprentissage d'ontologie depuis 10 ans *Cimiano et al.* (2014), *Mondary et al.* (2008). Pour cette raison, la première tâche pour la construction d'une ontologie de domaine est l'installation d'un corpus couvrant au maximum le lexique du domaine. Pour constituer nos corpus, nous avons suivi les préconisations pour la constitution de corpus généraux et spécialisés présentées dans *Wynne* (2005).

Nous avons installé deux corpus, français et russe. La plus grande partie des deux corpus rassemble des documents identiques par leurs structures et leurs contenus. Ce sont notamment des documents officiels accessibles en ligne, tels que les Normes de Sûreté

Radiologique et les rapports publiés par l'IAEA et les Commissions nationales et internationales de protection radiologique. Le corpus français contient, en plus, des articles de revues spécialisées publiés depuis 1999. Ce corpus français contient environ 1 500 000 mots dans 63 documents ; le corpus russe contient 600 000 mots dans 48 documents.

La plupart de ces textes étant initialement en format PDF, nous avons dû les convertir dans un format .TXT utilisable pour le traitement, puis les nettoyer de tous les caractères parasites.

5.4.2 Étiquetage des textes

L'étape suivante du traitement du corpus est l'étiquetage des textes avec les balises morpho-syntaxiques.

Actuellement, l'éventail des outils librement disponibles pour l'analyse syntaxique et morphologique offre un vaste choix. Pour le corpus français, le choix en faveur de TreeTagger a été fait à partir des résultats présentés dans *Falk et al.* (2014) sur les 7 parseurs du français et à partir de notre expérience ; nous avons testé plusieurs outils, notamment les analyseurs de Stanford et Brill. La qualité de TreeTagger s'est avérée particulièrement bonne pour les substantifs, les adjectifs et les verbes ; or ces parties du discours nous intéressent en priorité car ils caractérisent particulièrement le lexique du domaine. TreeTagger est l'outil-pionnier d'étiquetage automatique basé sur un modèle probabiliste d'arbre de décision *Schmid* (2000). Le parseur distingue 33 formes morphologiques et grammaticales pour le français et il renvoie le lemmes de chaque mots d'entrée. Pour nos objectifs, on n'a pas besoin d'une analyse morphologique aussi détailléee ; par exemple la distinction des temps des verbes, ou du genre des substantifs et adjectifs, ne nous est pas utile. Pour ces raisons, au cours de l'étiquetage du corpus français, nous avons réduit le nombre des formes distinguées de 33 à 15 (Table 5.1).

Par exemple, la sortie de TreeTagger pour le fragment de texte « *…détriment causé au personnel médical est pris en compte de façon subsidiaire* » est la séquence suivante :

« *détriment/NOM/détriment causé/PPP/causer au/PRP/au personnel/NOM/personnel médical/ADJ/médical est/VER/être pris/PPP/prendre en/PRP/en compte/NOM/compte de/PRP/de façon/NOM/façon subsidiaire /ADJ/subsidiaire ./SENT/.* »

Pour le corpus russe nous avons testé deux outils : la version de TreeTagger pour le russe, et le module morphologique proposé dans le cadre du projet AOT [1], basé sur un automate fini et un dictionnaire morphologique de 161 000 lemmes. Ce dernier est construit à partir du dictionnaire grammatical russe de Zalizniak (А. Зализняк), *Zaliznyak* (1977), *Sokirko* (2004). Bien que les résultats des tests obtenus avec ce dernier outil aient été un

1. www.aot.ru

Balise	Descripteur grammatical
ABR	Abréviation
ADJ	Adjectif
ADV	Adverbe
ART	Articles
DET	Pronom possessif
KON	Conjonction
NAM	Nom propre
NUM	Nombre ou tout chiffre
PPP	Participe passé
PRO	Tous les autres pronoms
PRP	Toutes les prépositions (y compris celles liées aux articles)
PUN	Toute ponctuation autre que le point
SENT	Point
SYM	Symbole
VER	Toutes les formes des verbes sauf PPP

Table 5.1: Liste réduite de nos 17 balises morpho-syntactiques pour le français, regroupant celles de TreeTagger.

peu meilleurs que ceux obtenus à l'aide de TreeTagger, nous avons décidé d'utiliser le dernier pour garder l'homogénéité du cycle de traitement des deux corpus.

TreeTagger pour le russe distingue plus de 750 formes morphologiques. Dans ce cas, nous avons réduit le nombre des formes à 8. Leur description est présentée dans la Table 5.2.

TreeTagger	Descripteur grammatical	Tag simplifié
Afc*, Afp*	Toutes les formes des adjectifs	ADJ
Mc*	Toutes les formes des noms et adjectifs numéraux	NUM
Nc*	Toutes les formes substantives	Noun
P-*	Tous les pronoms	PRN
R*	Adverbes	ADV
Vm*	Tous les verbes	Verb
Vmps*, Vmpp*	Tous les participes	Part
Aposition	Prépositions	Adp

Table 5.2: Liste réduite de nos 8 balises morpho-syntactiques pour le russe, regroupant celles de TreeTagger.

Voici, par exemple, la sortie de TreeTagger pour le fragment de texte : «*Уровни общего радиактивного загрязнения кожи определены с учетом проникновения части загрязнения через неповрежденную кожу...* » (Les niveaux de la contamination radioactive totale de la peau sont définis en tenant compte la pénétration de la partie de contamination dans la peau intacte).

Уровни/Noun/уровень общего/ADJ/общий/ радиактивного/ADJ/радиактивный загрязнения/Noun/загрязнение кожи/Noun/кожа определены/Part/определить с/Adp/с учетом/Noun/учет/ проникновения/Noun/проникновение части/Noun/часть загрязнения/Noun/загрязнение через/Adp/через неповрежденную/ADJ/неповрежденный кожу/Noun/кожа.

Lors de l'étiquetage des deux corpus nous n'avons pas effectué d'apprentissage de Treetagger sur le lexique spécifique au domaine car les résultats se sont avérés fiables, sans réglage additionnel. Sur notre grand corpus français, nous n'avons observé que deux types d'erreurs de TreeTagger : celles dues aux fautes d'orthographe causées par la conversion des fichiers PDF en format texte, et quelques erreurs d'étiquetage causées par l'insuffisance du modèle de langue ; par exemple le nom « personnel », au sens du groupe des gens qui travaillent, a été interprété comme un adjectif (plus de détails sur cet aspect sont donnés dans la section 5.5.2).

5.4.3 Sélection des ressources lexicales

Nous avons testé plusieurs dictionnaires de synonymes pour la langue française et pour la langue russe, notamment WOLF, BabelNet, ALEXANDIA, CRISCO, RusNet, PyTez, DCS [2] (cf. la section 2.6). A la suite de ces tests, nous avons choisi CRISCO et DCS. Le critère principal de choix d'une ressource lexicale est sa cardinalité (sa complétude).

Notons que les listes de synonymes d'une même entrée lexicale sont différentes selon les dictionnaires, cf. la table 5.3 pour les synonymes du terme *dommage*.

Dictionnaire	WOLF	BabelNet	ALEXANDIA	CRISCO
Nb des synonymes	6	11	12	38

Table 5.3: Nombre de synonymes du terme *dommage* dans les différents dictionnaires.

Mais l'utilisation d'un dictionnaire de synonymes n'est pas suffisante : la plupart des mots étant polysémiques, leurs synonymes sont différents d'un contexte à l'autre ; autrement dit ils peuvent rentrer dans plusieurs synsets [3] et ceci ne respecte pas le principe fondamental d'une ontologie qui est de garantir l'interprétation univoque des données.

Il est donc nécessaire de définir des critères fiables permettant à chaque fois de choisir la signification de l'un ou l'autre des candidats-termes. Comme cela a été annoncé, nous utilisons pour cela une modélisation linguistique et sémantique du corpus et du domaine.

2. Dictionnaire complet des synonymes (Полный словарь синонимов русского языка)
3. Dans la terminologie de WordNet un sysnet signifie une liste des synonymes.

5.5 Méthodes et Résultats

5.5.1 Modélisation sémantique et linguistique du noyau d'ontologie

La terminologie spécifique d'un certain domaine est univoque. Ainsi les dénominations des phénomènes concernés, des grandeurs physiques, des unités de mesure sont strictement définies et listées dans les glossaires spécialisés. Leur dénomination ne varie guère dans les textes techniques et scientifiques, sauf par le changement des terminaisons grammaticales (ex. centrale nucléaire – centrales nucléaires). De nombreuses méthodes permettent de traiter ces variations grammaticales (cf. la section 4.3).

Par contre, par sa définition même, un concept présente toujours toute une classe d'objets possédant des propriétés similaires. Il y a deux moyens de définir un concept : soit par son intention, i.e. la définition explicite restreignant ses propriétés, soit par son extension, i.e. par l'énumération des objets possédant sa propriété caractéristique (cf. la section 2.3.2 du chapitre 2).

Un exemple de l'intention d'un concept est la définition du *risque* (*radiologique*) donnée dans le glossaire de sûreté de l'AIEA [4] :

> *Probabilité qu'un effet sanitaire déterminé survienne chez une personne ou dans un groupe à la suite d'une exposition à des rayonnements.*

Tandis que l'extension de ce concept *dommage* est présenté dans ce glossaire par des termes tels que *perte*, *décès*, *dégradation*, etc.

Les langages formels de description d'ontologie tel qu'OWL peuvent traiter les définitions en langue naturelle. L'apprentissage des définitions peut donc être considérée comme une des tâches lors de la construction d'une ontologie. Mais nous jugerons qu'au stade initial de la construction d'ontologie, il est plus pratique et plus facile de chercher des mots isolés, pertinents du domaine, que des fragments longs. Nous avons donc choisi la représentation des concepts par leurs extensions d'où la définition suivante du noyau d'ontologie.

Définition. Le noyau d'ontologie est la combinaison de la liste des classes sémantiques des noms, chaque classe correspondant à l'extension d'un concept, et du cadre prédicatif modélisant leurs relations sémantiques (cf. la section 5.3 et la figure 5.2).

Donc, l'installation du noyau d'ontologie comprend les étapes suivantes :

1. En partant de la liste des termes initiaux, installer les classes sémantiques de noms, composées des synonymes des termes initiaux.
2. Définir les catégories de relations conceptuelles et les prédicats qui correspondent à chacune d'entre eux.

4. Glossaire de sûreté de l'AIEA, Terminologie employée en sûreté nucléaire et radioprotection, Édition 2007.

3. Trouver dans le corpus les verbes réalisant chaque prédicat et installer les classes sémantiques de verbes, composées des synonymes des verbes sélectionnés comme bases prédicatives.

Dans les sections qui suivent, nous présentons les méthodes adoptées pour chaque étape, ainsi que les résultats des expérimentations.

Description de la méthode

Le déroulement de la méthode est schématisé dans la figure 5.4.

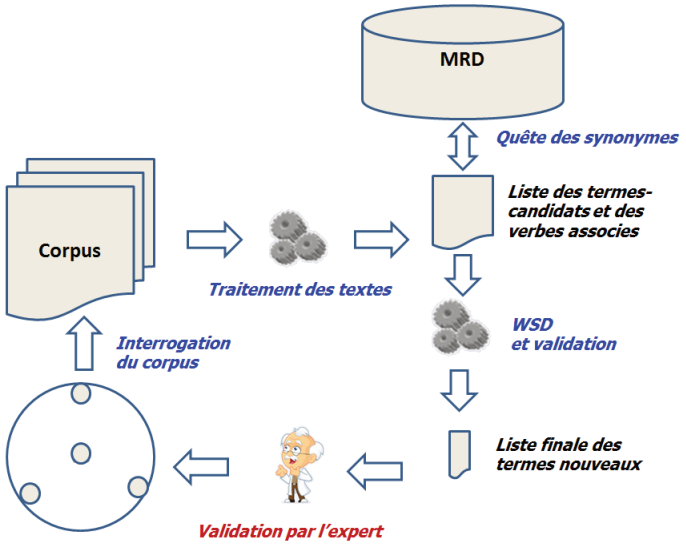

Figure 5.4: Schéma de la méthode d'installation du noyau d'ontologie.

On peut différencier trois étapes, chacune étant subdivisée en plusieurs sous-étapes :

I. Préparation des données ;
II. Analyse syntaxique des textes et formation du contexte (sélection des verbes caractéristiques) ;
III. Validation des résultats.

Préparation des données.

Cette étape correspond aux tâches suivantes :

- Mise en forme du corpus et première modélisation du domaine à l'aide des experts de ce domaine (cf. la section 5.4.1).

- Balisage de corpus avec l'analyseur syntaxique (cf. la section 5.4.2).

Choix de la liste initiale des concepts.

La sélection des concepts initiaux a été réalisé en 4 étapes :

1. Au départ, nous avons extrait des deux corpus, français et russe, les 100 « meilleurs » candidats-termes selon l'indice *TF-IDF* (cf. la table 4.1, p. 65) ; la liste a été organisée par ordre décroissant des scores.
2. Nous avons aligné les 100 canditas-termes précédents avec le glossaire de l'IAEA [5] sur la radioprotection, en éliminant de la liste les noms non spécifiques au domaine.
3. Pour constituer la liste des concepts généraux du domaine, nous nous sommes servis du cadre de RISQUE (cf. la section 2.6.1.1) qui synthétise la présentation des situations liées aux risques de toute nature.
4. Enfin, la validation définitive par l'expert nous a permis de retenir une liste de dix mots, ceux-ci devenant les dénominations initiales des concepts. Cette liste comprend les termes suivants (dans l'ordre alphabétique) : (*dommage, exposition, contrôle, personnel, population, protection, rayonnement, risque, sûreté, source*) ; en langue russe, la liste correspondante est : (*ущерб, облучение, контроль, персонал, население, защита, излучение, риск, безопасность, источник*).

L'analyse syntaxique de texte.

Nous réalisons une première analyse grammaticale pour récupérer dans le corpus les couples du type (*w,v*), où *w* est le nom et *v* est le verbe dans la même phrase [6]. Dans la liste complète de toutes les paires *nom–verbe*, nous conservons celles qui contiennent les termes prédéfinis ou leurs synonymes suggérés par le dictionnaire. Un module en Java a été écrit pour cette étape.

La sélection des verbes caractéristiques.

L'évaluation des synonymes des termes initiaux a été réalisée selon la méthode FCA (cf. la section 3.6.1, p. 49) : on estime que deux noms sont de *vrais synonymes* s'ils sont associés aux mêmes verbes caractéristiques.

Afin de sélectionner les verbes caractéristiques qui forment le contexte formel de chaque concept, nous avons proposé le coefficient $\boldsymbol{K}$ pour mesurer le degré d'association entre chaque terme général et chacun des verbes qui lui sont associés dans le corpus.

Le coefficient $\boldsymbol{K}$ (cf. la formule 5.1) est le produit de l'information mutuelle, MI (cf. la formule 4.11, p. 68) et du coefficient de Jaccard (cf.la formule 4.23, p. 70). Cette formule

5. The International Atomic Energy Agency
6. Les séparateur des phrases sont tous les signes de ponctuation sauf la virgule

composite a été choisie pour atténuer les valeurs trop grandes liées aux verbes apparaissent rarement dans le corpus. En reprenant les notations de la section 4.3.3 :

$$K = MI(c_i, v_j) \times JaccardCoefficient(c_i, v_j) \tag{5.1}$$

où

$$MIc_i, v_j) = |W| \times log \frac{TF(c_i, v_j)}{TF(c_i) \times TF(v_j)}$$

et

$$JaccardCoefficient(c_i, v_j) = \frac{TF(c_i, v_j)}{TF(c_i, \bar{v}_j) + TF(\bar{c}_i, v_j)}$$

Dans les formules sur *MI* et *Coefficient de Jaccard*, nous avons remplacé *x* et *y* par c_i et v_j respectivement, où c_i est un concept en tant que sujet [7] du verbe v_j.

Ceci peut être formellement résumé comme suit.

- $W = \{w\}$ est l'ensemble des noms [8] apparaissant au moins une fois dans le corpus ;
- $V = \{v\}$ est l'ensemble des verbes apparaissant au moins une fois dans le corpus ;
- $C = \{c_i | i = 1 \cdots n, n \in N\}$ est la liste des concepts représentant le modèle de domaine ;
- $CV = c_i, v_j$ est l'ensemble des paires « nom–verbe » où le nom correspond à un concept c_i qui est sujet du verbe v_j ;
- $DL_i = \{l_{ij}\}$ est l'ensemble des synonymes de concept c_i trouvés dans le corpus ; $DL_i = W \bigcup DL_i$ et $DL_i = \{w \subset W | \forall w : (w, c_i) \in I_{syn}\}$;
- $DL = \{DL_i | i = 1 \cdots n, n \in N\}$ est l'ensemble des listes de synonymes qui ont été trouvés pour les concepts à l'aide du dictionnaire CRISCO ;
- $I_{syn} : DL \times W$ est la relation (binaire et transitive) de synonymie.

Les graphiques Fig. 5.5 et Fig. 5.6 montrent les valeurs du coefficient d'association entre le concept *dommage* et les verbes qui lui sont associés, respectivement dans les corpus français et russe. Le seuil a été établi heuristiquement au niveau 1,3E-4. Les verbes pour lesquels la valeur de coefficient d'association est inférieure à cette grandeur ont été conservés en tant que verbes caractéristiques.

7. Nous supposons que le sujet est le substantif à gauche d'un verbe situé dans le même énoncé.
8. L'ensemble des lemmes pour être plus précis.

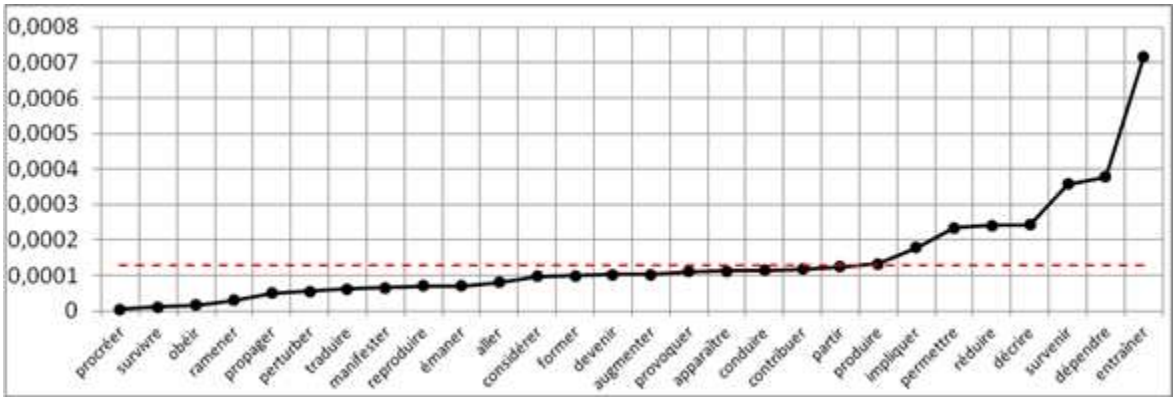

Figure 5.5: Corpus français : les verbes associés au concept *dommage*, rangés selon le coefficient d'association ***K*** = 1,3E-4 (la formule 5.1)

.

Les résultats

Les résultats des expérimentations sur la labellisation des concepts initiaux à partir des verbes caractéristiques sont rassemblés dans les tableaux 5.4, 5.5, 5.6, 5.7, 5.8. Dans les premières colonnes des tableaux 5.6 et 5.8, les mots sont les traductions russes de ceux de la première colonne des tableaux 5.5 et 5.7.

	Total des substantifs	**Total des verbes**	**Total des verbes associés avec les concepts**	**Total des prédicats**
français	2464	1094	493	211
russe	1774	981	547	324

Table 5.4: Informations agrégées sur les deux corpus.

La première colonne contient la liste des concepts. Les valeurs qui se trouvent dans les autres colonnes des tableaux 5.5 et 5.6 sont :

Verbes associés contient le nombre de verbes associés au concept(on compte ici tous les verbes du corpus liés au concept donné).

Prédicats correspond au nombre de verbes dont le coefficient d'association est au-dessous du seuil (ils forment le *contexte formel* selon la méthode AFC).

Candidats-synonymes contient le nombre de noms qui ont été trouvés dans le dictionnaire des synonymes CRISCO pour chaque concept initial.

Candidats-synonymes dans le corpus contient le nombre des synonymes présupposés présents dans le corpus.

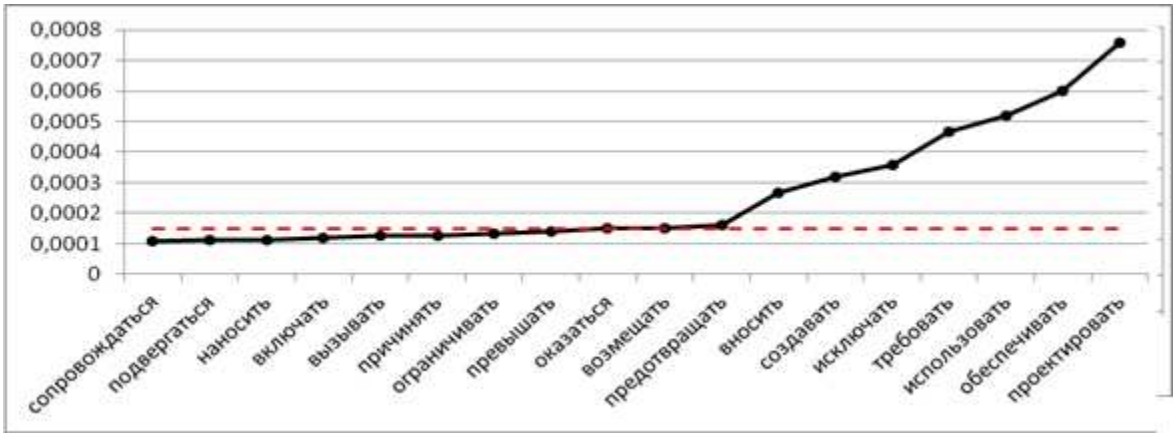

Figure 5.6: Corpus russe : les verbes associés au concept *ущерб* (*dommage*), rangés selon le coefficient d'association ***K*** = 1,3E-4 (la formule 5.1)

.

Concept	**Verbes associés**	**Prédicats**	**Candidats-synonymes**	**Candidats-synonymes dans le corpus**	**Candidats-termes respectant le seuil**	**Termes validés par l'expert**
dommage	28	20	38	11	7	6
exposition	92	69	34	22	5	1
contrôle	111	92	17	17	7	4
personnel	69	54	–	–	–	–
population	42	31	9	5	5	3
protection	148	127	77	32	25	17
rayonnement	144	103	26	12	10	5
risque	163	117	12	6	4	2
sûreté	86	51	31	12	9	4
source	159	117	35	22	9	4

Table 5.5: Résumé des résultats de labellisation de termes généraux en français.

Candidats-termes sélectionnés contient le nombre de synonymes associés aux prédicats.

Termes validés contient le nombre de noms qui ont été finalement acceptés par l'expert.

Exemple qualitatif en français : un des concepts du domaine de la Sécurité Radiologique correspond au mot *dommage* en français, ou à son équivalent en russe *ущерб*. Dans le dictionnaire des synonymes CRISCO, on trouve 38 synonymes potentiels. Dans

Concept	Verbes associés	Prédicats	Candidats-synonymes	Candidats-synonymes dans le corpus	Candidats-termes respectant le seuil	Termes validés par l'expert
ущерб	19	11	22	7	4	3
облучение	41	31	3	2	2	1
контроль	58	34	52	14	12	8
персонал	23	17	38	11	5	3
население	26	16	8	1	1	1
защита	51	41	48	19	7	4
излучение	48	35	10	3	1	1
риск	64		51	25	7	4
безопасность	57	46	31	5	3	2
источник	39	29	37	2	6	2

Table 5.6: Résumé des résultats de la labellisation des termes généraux en russe.

Concept	Labels linguistiques
dommage	*dégradation, destruction, détérioration, détriment, lésion, perte*
exposition	*exposé*
contrôle	*inspection, révision, sondage, surveillance*
personnel	–
population	*habitant, gens, démographie*
protection	*assistance, abri, secours, sauvegarde, préservation, défense, clôture, recommandation, soutien, appui, tutelle, couvercle, écran, verrouillage, conservation, parrainage, restauration*
rayonnement	*chaleur, lumière, irradiation, radiation, rayon*
risque	*danger, menace*
sûreté	*garantie, sécurité, précaution*
source	*point, mine, endroit, cause*

Table 5.7: Résumé des résultats de la labellisation des termes généraux en français.

le corpus seuls 11 mots de cette liste sont présents : (*atteinte, brèche, dégât, dégradation, destruction, détérioration, détriment, ennui, intérêt, lésion, perte*). Par ailleurs, dans le corpus, le mot *dommage* paraît en tant que sujet avec 28 verbes différents. Mais seuls 20 de ces verbes ont un coefficient d'association $\boldsymbol{K}$ inférieur au seuil choisi. Enfin, parmi les 11 termes-candidats, seulement 7 se trouvent avec les mêmes verbes dans le corpus : (*brèche, dégradation, destruction, détérioration, détriment, lésion, perte*). Enfin le mot *brèche* a été exclu par l'expert de domaine.

Concept	Labels linguistiques
ущерб	*вред, повреждение, авария*
облучение	*экспозиция*
контроль	*инспектирование, инспекция, наблюдение, надзор, обследование, осмотр, проверка, ревизия*
персонал	*кадры, люди, состав*
население	*жители*
защита	*ограждение, охрана, предохранение, сохранение, экранирование*
излучение	*радиация*
риск	*вероятность, возможность, опасность, угроза*
безопасность	*надежность, пожаробезопасность*
источник	*ресурс, очаг*

Table 5.8: Résumé des résultats de labellisation des termes généraux en russe.

Par ailleurs, CRISCO a interprété le mot *personnel* en tant qu'adjectif, en conséquence, il était impossible d'utiliser la liste des synonymes obtenus pour le traitement dans les étapes suivantes.

La table 5.9 montre la corrélation entre les résultats obtenus sur les deux corpus, français et russe. Dans la dernière colonne de ce tableau, le signe $\subset$ signifie que les ensembles des synonymes en deux langues sont corrélés mais qu'une des deux listes est plus courte, c'est à dire qu'elle est un sous-ensemble de l'autre.

Concept		
dommage $\leftrightarrow$ *ущерб*	9 : 4	{Dommage} $\supset$ {Ущерб}
exposition $\leftrightarrow$ *облучение*	2 : 2	{Exposition} $\equiv$ {Облучение}
contrôle $\leftrightarrow$ *контроль*	5 : 9	{Contrôle} $\subset$ {Контроль}
personnel $\leftrightarrow$ *персонал*	– : 4	–
population $\leftrightarrow$ *население*	4 : 2	{Population} $\supset$ {Население}
protection $\leftrightarrow$ *защита*	18 : 6	{Protection} $\supset$ {Защита}
rayonnement $\leftrightarrow$ *излучение*	6 : 2	{Rayonnement} $\supset$ {Излучение}
risque $\leftrightarrow$ *риск*	3 : 5	{Risque} $\subset$ {Риск}
sûreté $\leftrightarrow$ *безопасность*	4 : 3	{Sûreté} $\supset$ {Безопасность}
source $\leftrightarrow$ *источник*	5 : 3	{Source} $\supset$ {Источник}

Table 5.9: Correspondance entre les résultats de labellisation de termes généraux en français et en russe.

Exemple qualitatif en russe : dans le dictionnaire russe DCS il y a 21 synonymes du terme initial *ущерб* et dans le corpus on ne trouve que 7 de ces mots (*вред, потеря,*

повреждение, авария, осложнение, ухудшение) ; cinq d'entre eux sont associés à les prédicats sélectionnés, et trois sont validés par l'expert.

En français la plus faible précision est 20% (pour le terme *exposition*) et la meilleure précision est 87% (pour le terme *dommage*). En russe les résultats sont 33% pour le terme *источник* et 75% pour le terme *ущерб*.

Notons que les classes sémantiques de noms et de verbes correspondent au niveau linguistique de l'interprétation d'un modèle conceptuel ; à partir d'eux il est possible de passer au niveau ontologique de cette conceptualisation à travers des concepts et prédicats.

5.5.2 Cadre prédicatif

Description de la méthode

Comme il a été mentionné au début du chapitre, les relations contribuent à la construction d'une ontologie au même titre que les concepts.

Dans la section précédente, nous avons présenté la méthode exploitant les verbes caractéristiques en tant que contexte pour compléter les concepts du noyau d'ontologie par des lexèmes ayant la même signification que termes initiaux.

Dans cette section nous proposons la description de notre méthode qui aide à établir les relations associatives entre les concepts eux-mêmes moyennant le *cadre prédicatif*.

Définition. Par *cadre prédicatif* nous entendons l'ensemble des indices lexicaux qui explicitent les relations entre les concepts et permettent de les détecter dans le corpus.

Nous nous concentrons sur les verbes car ils sont les principaux agents prédicatifs : à chaque catégorie de relation sémantique correspond un certain prédicat et chaque prédicat peut être réalisé à l'aide de plusieurs verbes qui, dans ce cas, forment une classe sémantique.

Par cette méthode, nous poursuivons les objectifs suivants : - trouver les types de relations sémantiques entre les concepts de l'ontologie, - pour chaque type de relations chercher le prédicat correspondant, - former les classes de verbes susceptibles d'exprimer la même relation.

L'idée qui a inspiré notre méthode est brièvement évoquée dans la section 3.5.3 (p. 48). Elle concerne la définition d'une *ontologie formelle* comme théorie logique introduisant un vocabulaire et des règles garantissant l'interprétation univoque de ce vocabulaire, *Guarino et Welty* (2004). En d'autres mots, une ontologie correspond à un système $\boldsymbol{S}$ qui peut être complètement décrit par l'ensemble fini des variables et de leurs valeurs et dont l'interprétation est imposée par certaines contraintes. Plusieurs définitions, tirés de *Cimiano et al.* (2014), éclairent ces déclarations.

Définition d'un *univers*. Soit un système ***S*** caractérisé par un ensemble fini de variables ; l'univers W[9] représente la situation où, pour caractériser le système ***S***, un sous-ensemble des variables est sélectionné,chaque variable correspondant à une certaine propriété de l'univers observé.

Définition d'un *espace de domaine*. Soit un système ***S***, alors un *espace de domaine* est une paire (*D,W*) où *D* est un ensemble particulier d'éléments et *W* est un ensemble d'univers.

Définition d'une *relation conceptuelle*. Soit un système ***S*** ainsi q'un espace de domaine *(D,W)*, alors une *relation conceptuelle* est la fonction $\rho^n : W \rightarrow P(D^n)$ où $P(D^n)$ désigne la puissance de tous les *n*-uplets sur D.

Définition d'une *conceptualisation*. Une conceptualisation est un triplet (*D,W,R*) où *D* est un espace de domaine, *W* est l'ensemble des univers possibles et *R* est l'ensemble des relations conceptuelles sur *D,W*.

On peut en conclure que la conceptualisation signifie l'imposition de restrictions sur l'interprétation d'un univers par la limitation de ses constituants ; autrement dit, pour la construction d'une ontologie particulière, nous ne prenons en considération qu'un nombre fini d'éléments liés par un nombre fini de relations ; les relations conceptuelles s'interprètent alors comme la projection de l'univers *W* sur l'ensemble des prédicats d'arité *n*. Cela veut dire que les relations jouent le rôle de restrictions dans une ontologie.

Mais l'affirmation inverse est aussi vraie : à toute relation, on peut associer un prédicat et, en connaissant sa structure, il est possible de reconstituer les éléments qui forment l'espace du domaine. Cette réciprocité entre les éléments et leurs relations peut être utilisée pour la construction d'une ontologie.

Par exemple CAUSER est une relation conceptuelle binaire qui met en correspondance toutes les entités du concept `Rayonnement` et `Dommage`.

Il y a deux façons de mettre en correspondance la conceptualisation formelle d'un système d'information et un univers réel. La première façon est d'essayer de représenter les relations par extension en les énumérant pour chaque univers ; mais c'est impossible puisque le nombre des univers possibles est infini. La deuxième façon est de représenter les relations par intention, ici par une théorie logique : les différents modèles correspondant aussi étroitement que possible aux univers, conformément aux conceptualisations choisies.

L'implémentation et les résultats

La diversité des moyens grammaticaux et lexicaux d'un langue pour exprimer les relations entre les objets du monde réel complique leur mise en évidence dans les textes. Une

9. du mot anglais *world*

des façons les plus explicites de le faire est l'emploi des verbes ; ceci nécessite d'introduire une phase d'analyse syntaxique dans les méthodes d'extraction de relations. Dans notre méthode, nous utilisons une analyse superficielle des phrases pour extraire les triplets sujet–verbe–objet (SVO), sujet et objet étant représentés par des termes désignant les concepts. En règle générale, le sujet est exprimé par un groupe nominal à gauche du verbe, tandis que l'objet est un groupe nominal à droite du verbe. Dans le cas d'une construction passive, ces places sont inversés. L'exploitation des lemmes permet de réduire la sensibilité de la méthode à cette inversion car l'objectif de cette étape est de trouver les verbes qui lient toutes les paires formées de termes initiaux.

Les résultats quantitatifs de l'extraction des triplets SVO (où les sujets et les objets (compléments) sont les paires de concepts) sont présentés dans la table 5.10. L'expert a retenu six types de relations sémantiques importantes dans le domaine de la radioprotection (selon la classification proposée dans la section 4.4.3). Elles sont listées dans la table 5.11.

On lui a aussi présenté la liste des 57 verbes qui lient les concepts dans le corpus. Parmi eux, l'expert a retenu 21 verbes qu'il a jugés les plus pertinents pour constituer la base de prédicats. La liste complète de ces verbes est présentée dans le tableau de l'annexe **B** (p. 127) où les verbes retenus comme pertinents sont mis en gras. Notons que plusieurs verbes se sont avérés synonymes et que la liste définitive des groupes de verbes ayant pour le but former les classes sémantiques contient 15 éléments :

1. *assurer, garantir*
2. *entrainer, provoquer*
3. *conduire, induire*
4. *conserver*
5. *constituer*
6. *contrôler, superviser*
7. *éliminer*
8. *évaluer, mesurer*
9. *diagnostiquer*
10. *limiter, réduire*
11. *localiser*
12. *maintenir*
13. *produire*
14. *risquer*
15. *subir*

Ces verbes servent de « points de départ » pour constituer autour d'eux les classes sémantiques de verbes. Cette démarche est accomplie en deux étapes.

	dommage	*exposition*	*contrôle*	*personnel*	*population*	*protection*	*rayonnement*	*risque*	*sûreté*	*source*
dommage	2			1					3	
exposition		2			2	2		4	3	
contrôle		4	3		1	3	1	3		2
personnel	4	4		1			2		5	
population		2	1				2	2		
protection		4	1	4	1		2	7	3	1
rayonnement	4	4	2	3	1	1	4	4		4
risque	1	1		1	1	1	2			4
sûreté		1	2	1		3		2	10	4
source	1	3	4			3	1	1	2	8

Table 5.10: Récapitulatif sur l'extraction des triplets SVO à partir des concepts initiaux.

Type de relation	**Prédicat** (paramètre ***a***)	**Verbe** (paramètre ***z***)
cause ↔ conséquence	*PCaus(x,y,z)*	*provoquer*
objet ↔ propriété/attribut	*PProp(x,y,z)*	*caractériser*
objet d'action ↔ sujet d'action	*PAct(x,y,z)*	*protéger*
événement ↔ action	*PEven(x,y,z)*	*risquer*
état ↔ action	*PStat(x,y,z)*	*réduire*
objet ↔ état	*PStat(x,y,z)*	*contrôler*

Table 5.11: Les catégories de relations sémantiques entre les concepts du domaine de la Sécurité Radiologique.

En première étape, comme pour les noms dans la méthode précédente, nous récupérons dans le corpus les synonymes potentiels des verbes, sélectionnés à l'aide du Dictionnaire de Synonymes CRISCO qui est en ligne [10]. Les résultats de ces procédures sont présentés dans les tables 5.13 et 5.14.

Le premier tableau contient les informations agrégées sur le nombre des différents verbes et noms, récupérés dans le corpus après étiquetage morphologique, ainsi que le nombre de verbes et de noms pour lesquels des synonymes sont trouvés dans le Dictionnaire CRISCO.

10. http ://www.crisco.unicaen.fr/des/

Paire sujet–objet	Verbe
exposition–dommage	induire, provoquer, conduire, diagnostiquer
rayonnement–dommage	conduire
exposition–rayonnement	caractériser
personnel-exposition	réduire
risque-contrôle	garantir, contrôler
protection–contrôle	entraîner
source–protection	contrôler
risque–population	évaluer
sécurité – risque	garantir
population – rayonnement	risquer, subir
exposition – source	produire

Table 5.12: Exemples des verbes liant les concepts du noyau de l'ontologie.

Dans la table 5.14, les résultats sont groupés selon la fréquence des synonymes du verbe ou du nom. On voit dans ce tableau que les listes de synonymes sont très larges. Chaque liste contenant des verbes très généraux, et donc peu informatifs, nous avons exclu les verbes dont l'index *TF-IDF* est égal à 0.

	Nb d'unités étiquetées	**Étiquetés correcte-ment**	**Présents dans le diction-naire *CRISCO***	**Manquants dans *CRISCO***
Verbes	1094	1029	1009	20
Noms	2464	2074	2049	25

Table 5.13: Les noms et les verbes du corpus reconnus par CRISCO.

	Total des mots	**0**	**1**	**2–4**	**5–12**	**13–20**	**>20**
Verbes	1009	42	46	137	318	213	253
Noms	2040	141	205	408	648	325	313

Table 5.14: Distribution des nombre de synonymes par mot dans le corpus.

Mais cette opération n'est pas suffisante pour constituer les classes sémantiques parce que la plupart des verbes sont polysémiques (d'autant plus qu'en français il y a dix fois moins de verbes que de noms), et parce que le dictionnaire ne distingue pas, de façon

explicite, les différents types de similarité sémantique, notamment la hiérarchie (ou subsomption) et l'équivalence, qui sont réalisées par les prédicats différents et ont des propriétés différentes en théorie logique : l'équivalence est symétrique et transitive, mais pas la hiérarchie (subsomption).

Pour cette raison, l'étape suivante est la recherche des « *synonymes réciproques* » dans chacune des listes de synonymes.

Définition. Deux mots, *a* et *b* sont les « *synonymes réciproques* » si, et seulement si, dans le dictionnaire, *a* figure parmi la liste des synonymes de *b* et *b* figure dans la liste des synonymes de *a*.

Par exemple le dictionnaire CRISCO donne quatre synonymes au verbe *brancher* : {*joindre, orienter, pendre, rattacher*}. Mais, dans ce dictionnaire, le verbe *pendre* n'a ni *brancher* ni *rattacher* parmi ses synonymes ; les seuls synonymes réciproques du verbe *brancher* sont {*joindre, orienter*}.

La justification du choix d'un bon critère pour évaluer la similitude sémantique de deux mots est non-triviale (*Nokel et Loukachevitch* (2013)). Dans le but de quantifier et mesurer le degré de synonymie entre verbes, nous avons testé la mesure de *Cosinus* (cf. la formule 5.2).

$$simCos(v_i, v_j) = \frac{V_i^c \cap V_j^c}{\sqrt{\|V_i\| \times \|V_j\|}} \tag{5.2}$$

Ici $V_i^c \cap V_j^c$ est le nombre de co-occurrences de verbe v_i et v_j avec le même concept. $\|V_i^c\|$ et $\|V_j^c\|$ sont les co-occurrences de ces verbes avec les autre noms du corpus.

L'algorithme est le suivant :

Soit :

On part d'un corpus étiqueté des textes spécialisés **Textes** ;

On cherche l'ensemble des classes sémantiques des noms {`Concept`} ;

On cherche la liste des prédicats *PType(a,x,y,z)* où *a* spécifie le type du prédicat.

Tâches à résoudre et objectif à atteindre : établir l'ensemble initial {***Z***} des verbes, i.e. des références lexicales pour pouvoir identifier les prédicats. Regrouper les verbes du corpus autour des éléments initiaux de {***Z***}.

Résultat à la sortie : L'ensemble des classes sémantiques des verbes {Verbes} représentant les relations entre les concepts.

L'algorithme :

On récupère dans le corpus tous les triples SVO.

On sauvegarde séparément le sous-ensemble des triplets S_cVO_c où le sujet S_c et l'objet O_c sont des entités des classes sémantiques {Concept}.

Á l'aide du dictionnaire de synonymes, on récupère dans le **Texte** les candidats synonymes pour chaque verbe *z* de chaque prédicat *PType(a,x,y,z)*. Á la sortie de cette procédure, on obtient des listes de synonymes pour chaque verbe initial.

On quantifie la similarité entre les verbes de chaque liste provisoire de synonymes. Les verbes pour lesquels la valeur est supérieure à un seuil prédéfini sont ajoutés dans la classe sémantique.

Certains résultats sont présentés dans la table 5.15.

L'étiquette de classe	**Total des synonymes**	**Dans la classe**, $\alpha > 0.25$	**Membres de la classe**
causer	24	6	*provoquer, déclencher, engendrer, entrainer*
protéger	27	3	*veiller, réserver*
contrôler	20	4	*examiner, inspecter,surveiller*

Table 5.15: Sélection des éléments des classes sémantiques.

Notons que la valeur du coefficient α a été établie de manière heuristique : on a testé plusieurs valeurs pour un bon arbitrage entre la précision et le rappel. La graphique 5.7 illustre les résultats pour le verbe *protéger* et ses synonymes.

Les verbes et les noms très spécifiques

Á l'issue de l'étape d'analyse syntaxique et de l'étiquetage morphologique du corpus, nous avons rencontré certains problèmes ; ils peuvent être séparés en trois catégories : le bruit dans les textes (dans certaines mots des lettres ont été déformées ou perdues

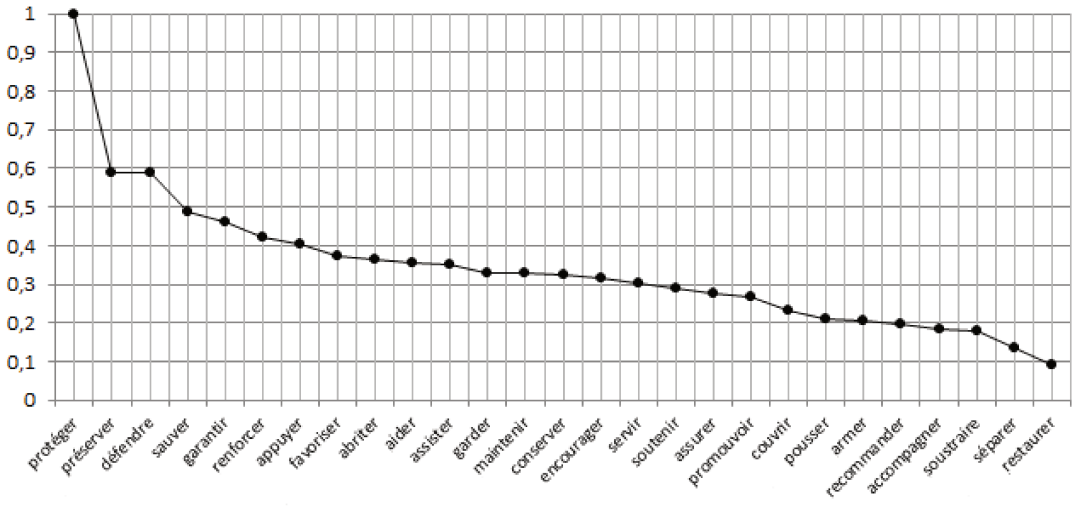

Figure 5.7: Quantification de la similarité entre le verbe *protéger* et ses synonymes.

lors de la transformation des PDF en TXT (on estime le taux de ces erreurs à environ 3%) ; les erreurs de l'analyseur qui étiquette mal certains mots (environ 1% des mots) ; les erreurs de l'analyseur qui étiquette les mots incorrects ; les lacunes du dictionnaire (certains mots corrects sont absents de sa base lexicographique). Bien que ces mots absents du dictionnaire soient peu nombreux et qu'ils n'aient pas été pris en compte dans nos expérimentations, leur analyse a montré que ce sont le plus souvent des termes particuliers, très spécifiques du domaine. Ainsi parmi les 66 mots qui ont été étiquetés par TreeTagger en tant que verbes mais rejetés par le dictionnaire des synonymes, on trouve 14 verbes corrects, soit 21%, qui font partie de la terminologie de la Sécurité Radiologique (cf. la table 5.16) .

cartographier, cloner, corréler, dépressuriser, désactiver, dimensionner, impacter, fiabiliser, hydrogéner, médeciner[11]*, mondialiser, normer, redéployer, sédimenter*

Table 5.16: Liste des verbes spécifiques au domaine, présents dans le corpus et absents du dictionnaire.

Parmi les 307 mots, qui ont été **étiquetés par TreeTagger** en tant que noms mais **rejetés par le dictionnaire**, on trouve 186 mots qui ont été étiquetés correctement dont 56, soit 18%, appartiennent à la terminologie de domaine (cf. la table 5.17) .

chimiothérapie, collimation, conductivité, contremesures, dangerosité, délétion, écotoxicité, exposure, imagerie, nucléide, etc.

Table 5.17: Liste des noms spécifiques au domaine, présents dans le corpus et absents du dictionnaire.

Si on choisit la stratégie « *bottom-up* » pour la construction d'une ontologie, ces mots, rares et très spécifiques, peuvent servir de bons indicateurs dans le corpus afin de démarrer le processus.

5.5.3 Méthode des patrons terminologiques

Description de la méthode

L'hypothèse de travail de cette méthode est que le lexique du domaine peut être repéré dans le corpus spécialisé à l'aide de l'analyse linguistique. En disposant d'une liste de termes génériques et en découvrant empiriquement les structures syntaxiques fréquentes dans lesquelles ces termes apparaissent, on peut élargir le noyau d'ontologie par de nouveaux termes, formant la taxonomie, *Orobinska et al.* (2013).

Par exemple le mot *dose* fait partie du lexique du domaine de la radioprotection. Autour de lui se forment les termes variés tels que *dose efficace*, *dose efficace collective* etc., Fig. 5.8.

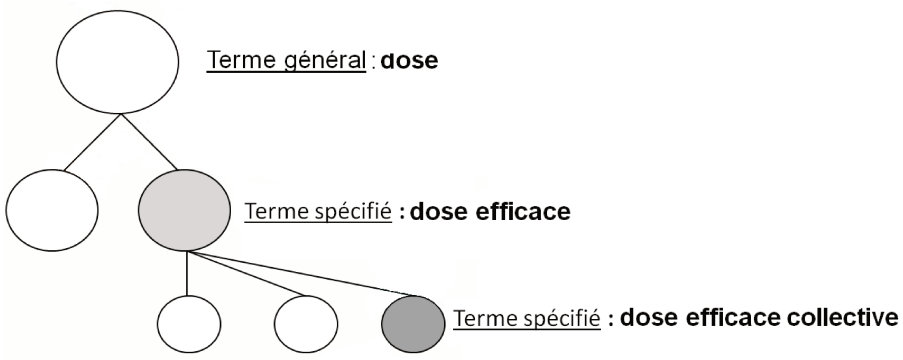

Figure 5.8: Taxonomie formée par « déploiement » autour du mot *dose*.

Notons que l'ontologie de noyau est une structure qui se compose de l'ensemble des identifiants de concepts, de la relation de subsomption entre les concepts qui est transitive, réflexive et antisymétrique (ordre partiel) et de l'ensemble des attributs de concepts, *Cimiano et al.* (2005), *Ganter et Wille* (1997).

Selon *Cabré et al.* (1998), les termes sont formés par des structures syntaxiques hiérarchisées. Et pour enrichir le noyau d'ontologie, il est possible d'utiliser des patrons terminologiques, que nous définissons comme la structure morpho-syntaxique avec l'un des termes génériques à la tête de chacun d'eux. Notre objectif est d'établir ces patrons.

Les patrons terminologiques sont formés de deux manières : à partir de l'analyse des fréquences de structures syntaxiques dans le corpus ; puis à partir de l'analyse syntaxique des termes du glossaire de domaine.

Les fragments de phrases qui correspondent aux patrons sont extraits automatiquement du corpus, puis validés par l'expert. Par construction, tous les fragments extraits contiennent des termes génériques qui forment le noyau d'ontologie : un des termes génériques est le radical de chaque terme nouveau.

Après validation, les termes dérivés de la même racine forment une taxonomie partielle. Ils sont ajoutés dans l'ontologie en tant qu'entités de concepts correspondants.

Le schèma exposant le processus d'extraction de termes est présenté dans la figure 5.9. L'implémentation de la méthode a été réalisée en Java. Le module *Prétraitement* effectue la conversion des fichiers PDF en format "texte" ; puis le nettoyage des textes obtenus élimine les fragments qui contiennent des caractères autres que des lettres, chiffres ou signes de ponctuation ; le module *étiquetage* munit chaque mot des tags de « parties du discours » (cf. section 5.4.2) ; le module *Formation des patrons* recense tous les patrons terminologiques ; le module *Extraction* permet de récupérer dans le corpus les fragments qui correspondent aux patrons, et il forme les taxonomies pour chaque terme-racine (cf. Fig. 5.10, Fig. 5.11, Fig. 5.12).

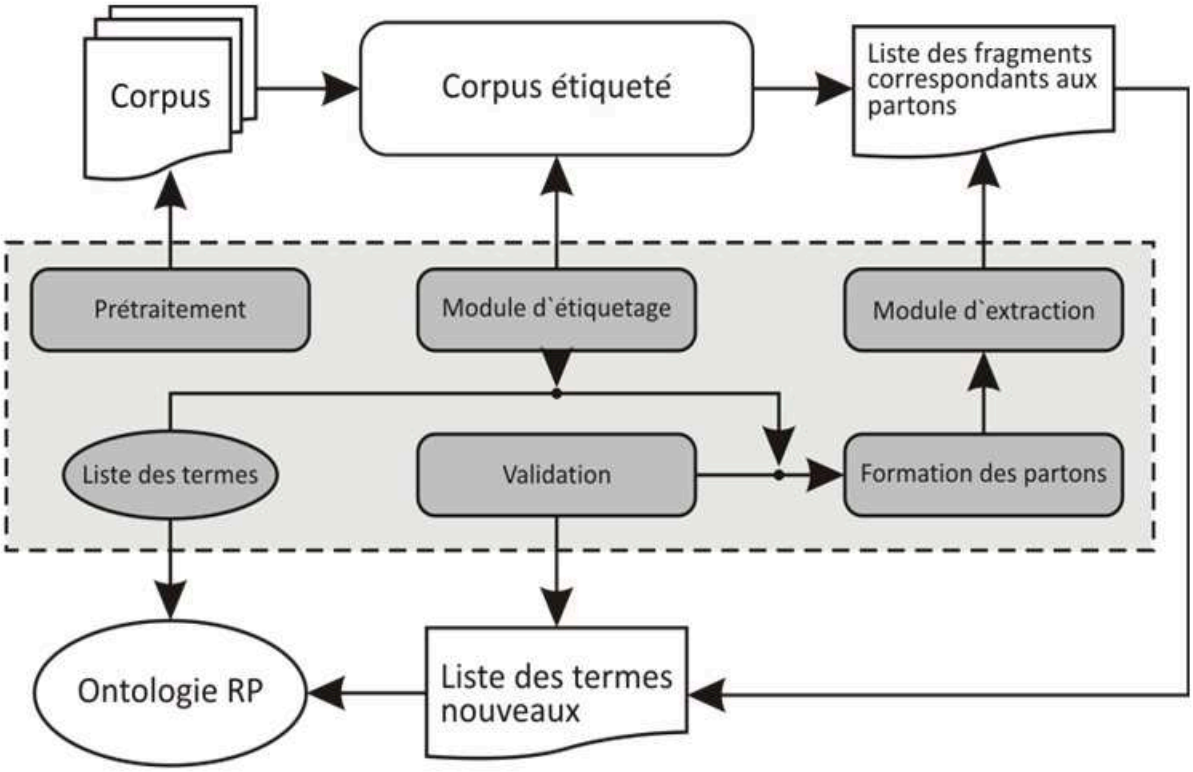

Figure 5.9: Schéma de notre système d'enrichissement de l'ontologie par la terminologie dérivée de la liste des termes génériques.

Une intervention humaine est nécessaire pour la validation définitive des termes-candidats détectés dans le corpus à partir des patrons terminologiques ; pour cela, les résultats sont présentés aux experts du domaine puis ajoutés à l'ontologie initiale s'ils sont validés.

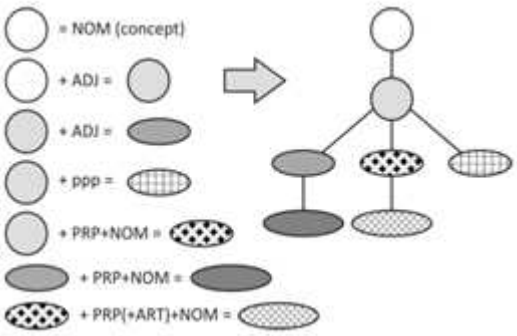

Figure 5.10: Taxonomie avec la racine *NOM+ADJ*.

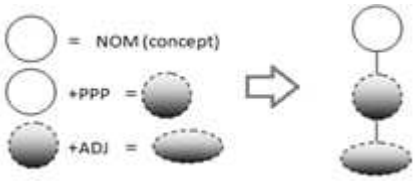

Figure 5.11: Taxonomie avec la racine *NOM+PP*.

Les résultats

Au départ les patrons sont des N-grams de balises grammaticales qui ont remplacé les mots dans le corpus ; nous avons utilisé des grams de taille *N* variant de 2 à 6. Nous avons extrait du corpus tous les fragments de phrases correspondants à ces N-grams. La sélection des patrons potentiellement pertinents a été faite à partir de la liste initiale de termes génériques (cf. section 5.1). Pour l'expérimentation, nous avons sélectionné 38 noms de cette liste.

Nous avons retenu les patrons pour lesquels au moins 70% des phrases correspondantes contiennent un de ces 38 termes. Le seuil a été choisi expérimentalement. Les résultats de cette partie de l'expérimentation sont présentés dans le Tableau. 5.18.

Après avoir réalisé l'analyse syntaxique des termes rassemblés dans le glossaire sur la radioprotection, nous avons ajouté 7 nouveaux patrons Tabl. 5.19.

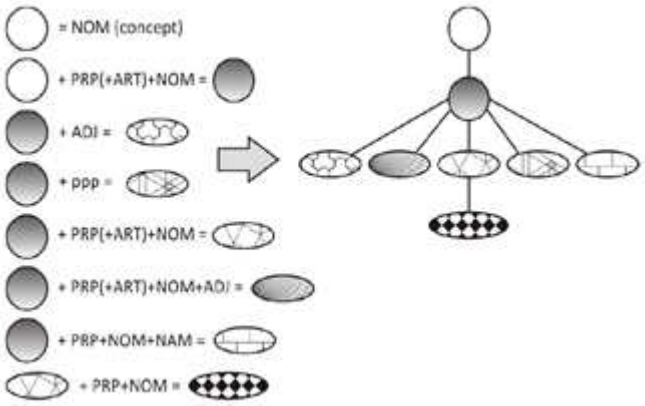

Figure 5.12: Taxonomie avec la racine *NOM+PRP+NOM*.

N	Patron	% des termes génériques dans le patron
1	$NOM + ADJ$	100%
2	$NOM + PRP + NOM$	100%
3	$NOM + PRP + NOM + ADJ$	92%
4	$NOM + PRP + NOM + PRP + NOM$	92%
5	$NOM + ADJ + PRP + NOM$	92%
6	$NOM + ADJ + ADJ$	84%
7	$NOM + PPP$	84%
8	$NOM + PRP + NOM + PPP$	76%
9	$NOM+PRP+NOM+PRP+ART+NOM$	74%
10	$NOM + ADJ + PPP$	71%
11	$NOM+PRP+ART+NOM+PRP+NOM$	71%

Table 5.18: Score de validations des patrons terminologiques formés automatiquement.

La liste finale, avec des exemples de termes correspondant à chaque patron terminologique, est présentée dans le Tabl. 5.20.

Les scores d'évaluation des candidats-termes extraits au moyen des patrons sont donnés dans le Tabl. 5.21 ; dans les colonnes, on lit les informations suivantes :

Patron contient la liste des patrons terminologiques ;

Nb d'expressions dans le corpus est le nombre total de fragments du corpus qui correspondent aux patrons ;

N	Patron	% des termes génériques dans le patron
1	$NOM+PRP+NOM+PRP+NOM+ADJ$	58 %
2	$NOM + ADJ + ADJ + PRP + NOM$	53%
3	$NOM+ADJ+PRP+NOM+PRP+ART+$ NOM ou $NOM + ADJ + PRP + NOM +$ $PRP + NOM$	39%
4	$NOM + PRP + NOM + PRP + ART +$ $NOM + ADJ$	39%
5	$NOM + PPP + ADJ$	37%
6	$NOM + PRP + NOM + PRP + ART +$ $NOM + PRP + NOM$	37%
7	$NOM + PRP + NOM + PRP + NOM +$ NAM	18%

Table 5.19: Patrons terminologiques formés à partir du glossaire.

Nb de fragments contenant un terme et **Taux de fragments contenant un terme** fournissent le nombre et le taux de fragments où entre au moins un des termes de la liste de départ ;

Nb de termes validés et **Taux de termes validés** donnent le nombre et le taux des termes corrects correspondant aux patrons.

Par exemple, ligne 1 du tableau Tabl. 5.21 : dans le corpus, on rencontre 14 759 expressions différentes correspondant au patron $NOM + ADJ$; parmi elles, 1 296 contiennent un mot de la liste initiale, soit 8,8% ($1296/14759 = 0,088$). Parmi ces 1 296 expressions, 558 sont acceptées par l'expert comme des termes qui compléteront l'ontologie, soit 43% ($0,43 = 558/1296$). Notons que la pertinence d'un patron terminologique dépend de sa taille N ; les patrons plus longs sont, en général, plus pertinents.

Par la suite nous utiliserons les définitions suivantes :

- *taxonomie partielle* : la taxonomie de chaque terme, formée par ses descendants (termes dérivées) ; ainsi chaque terme générique est le « sommet » de sa taxonomie partielle ;
- *racine d'un patron terminologique* : la structure minimale linguistique à laquelle un terme du domaine peut correspondre. Nous distinguons trois types de racines : $NOM + ADJ$, $NOM + PPP$ et $NOM + PRP + NOM$; l'ensemble des termes correspondant à une racine forment le niveau *I* de la taxonomie d'un concept ;
- *terme-descendant* le terme dérivé, formé à partir d'une racine ; les termes-descendants forment les niveaux *II* et *III* de chaque taxonomie partielle.

N	Patron	Exemple illustratif
1	$NOM + ADJ$	*zone urbaine*
2	$NOM + ADJ + ADJ$	*accident nucléaire grave*
3	$NOM + ADJ + ADJ + PRP + NOM$	*dose efficace professionnelle par an*
4	$NOM + ADJ + PPP$	*brûlure radiologique étendue*
5	$NOM + ADJ + PRP + NOM$	*effets stochastiques des rayonnements*
6	$NOM+ADJ+PRP+NOM+PRP+ART+NOM$ ou $NOM + ADJ + PRP + NOM + PRP + NOM$	*effet tardif du rayonnement dans le tissu*
7	$NOM + PPP$	*zone surveillée*
8	$NOM + PPP + ADJ$	*dose absorbée individuelle*
9	$NOM+PRP+ART+NOM+PRP+NOM$	*sûreté sur la gestion du déchet*
10	$NOM + PRP + NOM$	*réacteur à eau*
11	$NOM + PRP + NOM + ADJ$	*entreposage du déchet radioactif*
12	$NOM + PRP + NOM + PPP$	*cumul de dose absorbée*
13	$NOM+PRP+NOM+PRP+ART+NOM$	*action de prévention de la pollution*
14	$NOM + PRP + NOM + PRP + ART + NOM + ADJ$	*coefficient de risque pour l'effet nocif*
15	$NOM + PRP + NOM + PRP + ART + NOM + PRP + NOM$	*seuil de dose pour le risque de mortalité*
16	$NOM + PRP + NOM + PRP + NOM$	*durée de vie du réacteur*
17	$NOM+PRP+NOM+PRP+NOM+ADJ$	*site de stockage des éléments combustibles*
18	$NOM + PRP + NOM + PRP + NOM + NAM$	*fonctionnement du réacteur de type BWR*

Table 5.20: Liste finale de nos patrons terminologiques avec des exemples.

Les patrons terminologiques réunis autour de chaque racine permettent de former les taxonomies à trois niveaux. Ceci est illustré dans les figures Fig. 5.10, Fig. 5.11, Fig. 5.12.

Un exemple qualitatif et le récapitulatif quantitatif sont présentés dans les tables 5.22 et 5.23 respectivement.

Notons deux faits :

1. Les ensembles de termes retenus aux différents niveaux ne sont pas forcément emboîtés , par exemple le fragment *exposition continue* n'est pas retenu au niveau *I*, car non spécifique du domaine, mais *exposition continue au rayonnement* avec la structure $NOM + ADJ + PRP + NOM$ est retenu au niveau *II* pour enrichir l'ontologie ;
2. Le nombre des termes directement descendant du niveau *I* vers les niveaux *II* et *III* est inférieur au nombre des termes correspondant à chaque patron plus général

parce que les termes du niveau supérieur ne possèdent pas nécessairement de termes-fils.

Les exemples des instances qui peuplent les concepts sont présentés dans l'annexe **B** sur la page 130.

5.5.4 Règles de reconnaissance

Par peuplement d'ontologie, on entend le processus d'instanciation de la base des connaissances liée au noyau d'ontologie. Cette étape, pour laquelle nous proposons une technique basée sur l'utilisation des règles de reconnaissance, finalise traditionnellement le processus de construction d'une ontologie.

Dans la section 5.5.3, nous avons présenté une méthode permettant de repérer les termes à l'aide de patrons morpho-syntaxiques. On a montré que les patrons, composés seulement des étiquettes morphologiques sont déjà capables d'aider à détecter les termes, surtout les termes déployés, composés de plus de trois mots. La pertinence des résultats est significativement renforcée si on ajoute dans les patrons un nom défini dans le lexique de domaine, soit au début, soit comme deuxième élément du patron *NOM+PRP+NOM*.

Á l'aide de deux autres méthodes (cf. les sections 5.5.1 et 5.5.2) nous avons étendu des concepts et des relation associatives par des items lexicaux. Par exemple {***Risque***} $\supset$ {*danger, menace*}, {***Causer***} $\supset$ {*provoque, déclencher, engendrer*}. Ce niveau d'analyse était suffisant pour installer le noyau d'ontologie ; mais, pour l'instancier, il faut avoir un mécanisme de reconnaissance des termes complets, et nous proposons un tel mécanisme en combinant les trois méthodes exposées ci-dessus sous forme de règles de reconnaissance. Le principe de formation des règles de reconnaissance se formule comme suit :

Ayant défini un certain prédicat binaire et l'un de ses arguments, on trouve la valeur du deuxième argument

On peut l'illustrer par un exemple sur le prédicat représentant la relation causale qui se rapporte à la catégorie des relations fonctionnelles (cf. la table 4.3). Le prédicat de cause se déclare sous forme *PFun(a, x, y, z)*. Ici, le paramètre *a* spécifie le type concret de la relation fonctionnelle, $a = \langle Cause \rangle$; *x* est un terme signifiant la cause et *y* est la conséquence d'un événement ; *z* est une des références lexicales indiquant le type de relation. La plage de valeurs de *z* est très variée : $z \in$ {*à cause de, à la suite de, après un, puisque, ...* }. Dans nos recherches, nous nous sommes concentrés sur les verbes en tant qu'indicateurs du rapport de cause à effet entre *x* et *y*.

Exemple illustratif. Soit la phrase suivante, issue des « Recommandations de la Commission internationale de protection radiologique ».

*D'autres manières de considérer les **effets des rayonnements** peuvent par conséquent s'avérer être plus utiles aux espèces non humaines, notamment en ce qui concerne les effets qui **provoquent** une mortalité précoce, ou une morbidité, ou encore une diminution du taux de reproduction*.

Explications. La construction du noyau d'ontologie nous a permis de créer trois ensembles de résultats, à savoir les classes sémantiques des noms représentant des concepts (cf. la section 5.5.1), les classes sémantiques des verbes représentant des relations entre eux (cf. la section 5.5.2) et la liste des patrons morphologiques susceptibles de dépister dans le corpus les termes composés (cf. la section 5.5.3). En utilisant conjointement ces données, on trouve que la phrase d'exemple correspond à la réalisation linguistique du prédicat *PFun(a, x, y, z)* pour le cas où $a = \langle Cause \rangle$.

Alors,

z=provoquer parce que *povoquer* $\in$ {***Causer***}

x=effet de rayonnement parce que *rayonnement* $\in$ {**Rayonnement**} et il correspond au patron *NOM+PRP+NOM* ;

$\Rightarrow$ $y = \{$*mortalité précoce*, *diminution du taux de reproduction*$\}$ parce que ces deux éléments correspondent aux patrons *NOM+ADJ* et *NOM+PRP+NOM+PRP+NOM* respectivement. Deux nouvelles instances peuplent le concept `Dommage` parce qu'à l'étape de formation du cadre prédicatif, la relation causale a été établie entre les concept `Rayonnement` et `Dommage`.

Au niveau linguistique, dans la phrase qui contient un des concepts à gauche d'un des prédicats binaires du noyau d'ontologie, le fragment à droite est un candidat-terme si sa structure syntaxique correspond à l'un des partons morpho-syntaxique. Ce candidat-terme est subordonné au concept qui est associé au même prédicat à sa droite ; et inversement.

5.5.5 Synthèse des résultats

Finalement, à l'issue nos démarches, nous avons obtenu la ressource terminologique, sous forme d'une taxonomie de termes instanciant les 10 concepts du noyau ; et six types de relations sémantiques ont été définies. Les concepts et les relations sont présentés sous forme de classes sémantiques des synonymes. Les classes de concepts comprennent les termes-synonymes les plus génériques pour notre domaine. Ce sont des termes composé d'un seul mot. Les classes sémantiques des relations comprennent les verbes-synonymes. Dans notre cas, les relations ne sont pas hiérarchisées. Les concepts sont instanciés par les termes les plus spécifiques qui sont mis en taxonomies partielles, par principe de déploiement syntaxique. Au total, 1850 termes ont été retenus. Notons que les termes sélectionnés ne sont pas tous des synonymes directs du concept auquel ils sont associés ; mais il est garanti que chaque terme caractérise certains des aspects du

concept, autrement dit qu'il en est un attribut. La spécification des attributs des concepts n'est pas rentrée dans nos recherches actuelles ; c'est un objectif pour les travaux à venir.

5.6 Conclusion du chapitre 5

La recherche de connaissances dans les textes en langage naturel est un sujet actuel pour l'ingénierie d'ontologie. Il est possible d'améliorer les systèmes d'information dédiés à la construction d'ontologies à condition d'intégrer des modules de modélisation linguistique, notamment pour l'obtention d'informations sur les dépendances structurelles dans les phrases et sur les propriétés sémantiques et grammaticales de mots.

Nous avons proposé une stratégie générale incluant, dès le début de la construction d'une ontologie de domaine, des méthodes à la fois robustes, efficaces et pas trop complexes. Ici, nous avons présenté les méthodes aboutissant à ces résultats en plusieurs étapes, et des résultats d'expérimentation sur de gros corpus de textes professionnels.

Contrairement à l'approche répandue qui commence par l'extraction massive de termes-candidats, nous proposons de commencer par l'installation d'une liste limitées de termes généraux servant de noyau d'ontologie. Le premier objectif est alors d'élargir cette liste par des synonymes, soit, dans l'optique de la logique formelle, d'établir l'ensemble des éléments liés par la relation d'équivalence. Á la sortie de cette première étape, chaque terme initial est remplacé par la classe sémantique de ses synonymes, ce qui permet de passer au niveau des concepts d'ontologie. Á la deuxième étape, les concepts sont liés par des relations sémantiques qui, à leur tour, sont exprimées à l'aide des classes sémantiques de verbes. La disponibilité de deux ensembles de classes sémantiques, celui des concepts et celui des relations, permet de formuler les règles de reconnaissances pour faire "apprendre" l'instanciation de la base des connaissances d'ontologie. Les patrons morpho-syntaxiques assurent une bonne pertinence aux nouveaux candidats-termes.

L'approche peut se généraliser : la méthode a été testée sur deux langues, français et russe, qui sont assez différentes.

L'expérience a montré que la meilleure voie pour résoudre le problème de l'ambigüité des mots, si le corpus est de taille moyenne ou grande, est l'utilisation d'un bon dictionnaire de synonymes lisible par machine.

Les méthodes basées sur les informations linguistiques additionnelles ont des limites ; la plus critique est leur dépendance à la qualité des outils appliqués, particulièrement à l'exactitude des analyseurs syntaxiques et à la complétude des dictionnaires de synonymes utilisés.

N	Patron	Nb d'expressions dans le corpus	Nb de fragments contenant un terme	Taux de fragments contenant un terme	Nb de termes validés	Taux de termes validés
1	$NOM + ADJ$	14 759	1 296	8,8%	558	43%
2	$NOM + ADJ + ADJ$	1 998	356	17,8%	160	45%
3	$NOM+ADJ+ADJ+PRP+NOM$	179	34	19,0%	7	21%
4	$NOM + ADJ + PPP$	1 434	299	20,8%	30	10%
5	$NOM+ADJ+PRP+NOM$	3 556	425	11,9%	204	48%
6	$NOM+ADJ+PRP+NOM + PRP + ART + NOM$ ou $NOM+ADJ+PRP+NOM+PRP+NOM$	119	35	29,4%	25	71%
7	$NOM + PPP$	5 294	582	11,0%	58	10%
8	$NOM + PPP + ADJ$	187	46	24,6%	14	30%
9	$NOM+PRP+ART+NOM+PRP+NOM$	1 898	151	8,0%	30	20%
10	$NOM+PRP+NOM$	16 201	1134	7,0%	590	52%
11	$NOM + PRP + NOM + ADJ$	5 344	160	3,0%	80	50%
12	$NOM + PRP + NOM + PPP$	1 850	166	9,0%	17	10%
13	$NOM + PRP + NOM + PRP + ART + NOM$	1 670	200	12,0%	81	40%
14	$NOM + PRP + NOM + PRP + ART + NOM + ADJ$	219	56	25,6%	39	70%
15	$NOM + PRP + NOM+PRP+ART+NOM+PRP+NOM$	192	53	27,6%	27	50%
16	$NOM + PRP + NOM+PRP+NOM$	4 382	219	5,0%	88	40%
17	$NOM + PRP + NOM + PRP + NOM + ADJ$	868	95	10,9%	57	60%
18	$NOM + PRP + NOM + PRP + NOM + NAM$	75	14	18,7%	4	30%

Table 5.21: Résultats de l'extraction des candidats-termes extraits au moyen des patrons terminologiques.

	I	***I + II***	***I + II + III***
$NOM + ADJ$	*effet néfaste*	*effet néfaste d'exposition*	*effet néfaste d'exposition pour la santé*
$NOM + PPP$	*dose absorbée*	*dose absorbée individuelle*	
$NOM + PRP(+ART) + NOM$	*action de prévention*	*action de prévention de la pollution*	*action de prévention de la pollution par le rejet radioactif*

Table 5.22: Exemples de taxonomies de termes dérivés.

	$NOM + ADJ$	$NOM + PPP$	$NOM + PRP(+ART) + NOM$
I	558	58	620 [12]
$I + II$	238	13	380
$I + II + III$	7	–	4

Table 5.23: Nombres d'inductions de taxonomies partielles à partir de trois racines syntaxiques.

Concept à gauche (domaine)	**Prédicat**	**Patron à droite (range)**	**Candidat-terme (instance)**
lésion	provoquer	*NOM+ADJ+ADJ*	*brûlure radiologique localisé*
surveillance	contrôler	*NOM+PRP+NOM+ADJ*	*rejet des déchets radioactifs*
rayonnement	entraîner	*NOM+ADJ +PRP[ART]+NOM*	*contamination significative de l'environnement*
risque	réduire	*NOM+ADJ*	*rejets radioactifs*

Table 5.24: Exemples de règles de reconnaissance.

Chapitre 6

Conclusion et discussion

6.1 Résumé du travail

Les deux buts de l'ingénierie d'ontologie sont l'analyse des informations présentées sous forme de textes et l'extraction des données relevant d'un domaine particulier. L'amélioration de la fonctionnalité des systèmes d'apprentissage d'ontologies passe par la réalisation de modules capables de détecter dans un corpus les propriétés des concepts, ceci grâce à tout un arsenal de moyens d'analyse syntaxique et sémantique qui sont maintenant disponibles en ligne.

Selon *Murphy* (2012), « (machine learning is) *. . .a set of methods that can automatically detect patterns in the data and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainly* ».

Sans contester les avantages des méthodes statistiques appliquées à l'analyse des textes, notons néanmoins que ces méthodes ont des limites objectives parce que aucune règle ne couvre tous les cas. Par exemple, les verbes qui se terminent en -er ou -ir n'appartiennent pas tous au premier ou au deuxième groupe (Exemple : aller, courir). La seule alternative est de lister les « exceptions ». Ceci est d'autant plus vrai lorsqu'il s'agit de définir les sens des mots.

En théorie, il est possible de constituer le champ lexical d'une notion (i.e. de trouver les mots qui ont la même signification, synonymes) par extraction de tous les contextes où les sens possibles se distinguent. Un contexte correspond à l'ensemble des champs lexicaux [1] des mots qui rentrent dans le cadre sémantique d'une notion.

Mais la mise en pratique de cette idée, pour l'apprentissage d'une ontologie, demande des calculs si longs et un corpus spécialisé si grands que cela n'est guère envisageable. Finalement, il est préférable d'utiliser les ressources lexicales destinées aux recherches d'informations, et de trouver un équilibre entre les méthodes basées sur les patrons cal-

1. Dans les termes que nous adoptons dans ce mémoire, ce sont les classes sémantiques

culés et les méthodes s'appuyant sur la désambiguïsation explicite sémantique à l'aide des dictionnaires ou thésaurus.

Au cours de notre travail, nous avons essayé de suivre cette voie. Nous avons proposé, et mis en oeuvre, un algorithme cohérent pour la construction de l'ontologie du domaine de la sécurité radiologique. Avec notamment l'apprentissage des patrons morphosyntaxiques et d'installation de taxonomies partielles de termes, et la formation de classes sémantiques représentant les concepts et leurs relations. Toutes les méthodes sont intégrées, en partant d'une liste limitée de termes généraux, définie préalablement avec l'expert de domaine.

L'implémentation de cette approche a demandé de l'installation de deux corpus spécialisés dans le domaine de la protection radiologique, en français et en russe, comprenant respectivement 1 500 000 et 600 000 unités lexicales.

Une large synthèse sur l'état de l'art a précédé l'étape expérimentale. Elle couvre les divers aspects de l'apprentissage d'ontologies : les fondations théoriques de la représentation des connaissances, la modélisation de la langue naturelle, l'extraction des termes et des relations, la phase de la conceptualisation et le panorama des outils disponibles.

6.2 Perspectives de travail

Nous envisageons plusieurs pistes possibles pour continuer ces travaux à l'avenir.

Premièrement, nous souhaitons perfectionner les méthodes basées sur le contexte et étudier les structures syntaxiques les plus aptes à détecter la présence du lexique de domaine dans une phrase. Dans les expérimentations actuelles, nous ignorions les propriétés grammaticales des verbes telles que le temps ou le mode (actif ou passif). Par ailleurs, les verbes modaux peuvent servir de balises à partir desquelles on peut trouver les nouveaux candidats-termes ; ceci est motivé par l'observation du style des textes utilisés ; ici, il s'agit des normes prescrivant des actions en cas de risque d'exposition aux rayonnements atomiques, ou des règles en matière de contrôle de sécurité.

Prenons par exemple, les phrases suivantes :

La probabilité d'être exposé, le nombre de personnes exposées et le niveau de leurs doses individuelles ***doivent tous rester aussi faibles*** *qu'il est raisonnablement possible.*

Si une femme exposée professionnellement a déclaré être enceinte, des contrôles supplémentaires ***doivent être considérés*** *afin d'atteindre un degré de protection pour l'embryon et le fœtus du même ordre que celui assuré pour les personnes du public.*

Les doses, qu'elles soient fortes ou faibles, peuvent provoquer des effets stochastiques (cancer ou effets héréditaires), qui ***peuvent être observés*** *en tant qu'augmentation statistiquement détectable de l'incidence de ces effets survenant longtemps après l'exposition.*

On voit que, dans chaque phrase, les groupes nominaux placés avant le verbe modal font partie du vocabulaire du domaine. L'intérêt d'examiner les structures des phrases contenant des formes telles que **doit être**, ou **peut être** est liée au fait qu'elles sont les mêmes dans les différentes langues. Ceci qui peut permettre de généraliser les méthodes proposées.

Un deuxième axe de recherches peut être l'analyse des structures prédicatives minimales suffisantes pour désambiguïser le contexte (comme proposé par T. Zolotova, *Zolotova* (2011)). Par exemple, le changement de sens d'un verbe en fonction du changement de son schéma propositionnel. Comparer : *revenir à* et *revenir de*, *traiter en* et *traiter de* etc. Nous avons écrit un article sur ce sujet en 2012, *Orobinska* (2012) et nous souhaitons approfondir cette piste.

L'utilisation des ressources lexicales nous semble indispensable dans les méthodes d'apprentissage d'ontologies. Mais jusqu'à aujourd'hui on n'a presque pas de dictionnaires conçus spécifiquement pour faciliter la fouille de texte et la rendre plus intelligente. Les développements de WordNet vont dans cette direction ; notons que le thésaurus en russe РуТез, construit sur les principes de l'association des mots, a été pensé pour ces fins dès le début de son élaboration. L'équipe du Centre des Recherches en Informatique de l'université de Lomonosov (Научно-исследовательский вычислительный центр МГУ) nous propose une collaboration afin de tester cet outil sur notre corpus. Cela nous permettrait de voir si les résultats présentés ici peuvent être améliorés.

Cette direction de recherche nous permettra aussi de faire progresser les outils d'évaluation des résultats, car, bien que la tâche d'évaluation fasse partie du cycle de vie d'une ontologie, de nombreux progrès restent à faire. C'est important car cela permet d'améliorer - la qualité des ontologies elles-mêmes, - l'interopérabilité entre les systèmes, et - l'élargissement de leur champ d'application.

Ceci dit, dans le cadre des travaux réalisés, la validation des résultats par un expert reste actuellement une étape très importante.

Annexe A

Titre	Référence	Objectif et champ d'application	Caractéristiques principales
ASIUM, *Faure et Nédellec* (1998)	Laboratoire de Recherche en Informatique (LRI)de l'Université Paris-Sud (première vertion en 1999)	Trouve les relations taxonomiques entre les termes dans les textes complets en français sans les annoter	Utilise les algorithmes basés sur le regroupement conceptuel de textes français (conceptual clustering)
DL-Learner	L'Université de Leipzig, *Lehmann* (2007)	La source ouverte de Framework de Machine Learning, supportant l'OWL et la logique descriptive	DL-Learner comprend les nombreux algorithmes basés sur la programmation génétique ; le raffinement des opérateurs de support des fonctionnalités nombreuses de OWL, y compris le type de données de soutien, et un algorithme adapté pour l'ingénierie de l'ontologie avec une forte polarisation sur des concepts courts et lisibles.

DODDLE (II, OWL)	l'université de de Shizuoka *Morita et al.* (2008)	L'environnement interactive pour le développement des ontologies de domaine	L'assistance des outils de l'apprentissage des relations taxonomiques et non-taxonomiques, portant sur les méthodes statistiques, l'exploitation d'un dictionnaire numérique (WordNet) et les textes spécialises d'un domaine.
GATE, famille	Natural Language Processing Group, l'université de Sheffield, *Cunningham et al.* (2011)	La solution open source du cycle de vie complet pour traitement de texte	L'environnement intégré de développement pour traitement de la langue naturelle. GATE comprend l'ensemble exhaustif des plagins et son propre système d'extraction des informations largement utilisé dans les recherches.
medSynDiKATe	Université de Fribourg-en-Brisgau, *Hahn et al.* (1999)	L'acquisition ds connaissances à partir des textes complets tels que les rapports technologiques et médicaux.	L'apprentissage progressif des termes, des concepts et des relations basé sur l'analyse de deux niveaux : le niveau de sentence et niveau de texte. Il utilise les nombreux axiomes linguistiques et conceptuels tenant équilibre entre la généralisation et le raffinement.
NeOn	IST-2005-027595, la*Poveda et al.* (2009)	L'environnement mufti-plateforme à source ouverte du niveau de l'état de l'art	Fournit la bibliothèque de modèles d'ontologie pour faciliter les solutions de modélisation qui peuvent être appliquées à la résolution des problèmes récurrents lors de design d'ontologie.

OntoGain	Université technique de Crète, *Drymonas et al.* (2010)	Le système de l'acquisition non-supervisée d'ontologie à partir des textes non-structurés orienté à l'extraction des termes multi-mots.	L'acquisition des relation taxonomiques et non-taxonomiques basé sur la classification agglomérante hiérarchisée, AFC et sur les algorithmes probabilistes. L'OntoGain permet la transformation de l'ontologie obtenue en déclarations standards d'OWL
OntoGen	*Fortuna et al.* (2006)	L'éditeur semi-automatique spécialisé sur l'édition	L'interface d'utilisateur interactive ; les méthodes supervisées et non-supervisées de révélation des concepts ; l'extraction des mots-clé et visualisation conceptuelle.
OntoLT	DFKI, Allemand, *Buitelaar et Sintek* (2004)	L'objectif du projet est de faciliter l'intégration des méthodes de l'analyse linguistique dans la pratique de l'ingénierie d'ontologie	La réalisation des règles de mappage permettant de constituer la correspondance entre les entités linguistiques dans le texte et une classe/slot d'une ontologie.
SVETLAN'	LRI, l'Université de Paris-Sud, *de Chalendar et Grau* (2000)	La construction de l'hiérarchie des concepts	L'outil d'appui pour la construction et le développement une ontologie par biais d'apprentissage de l'hiérarchie des substantifs de corpus. Permet d'apprendre la sémantique de domaines à travers des unités thématiques ; permet de classer les noms à l'aide de l'analyse des relations entre les noms et les verbes.

TermExtractor	L'extraction de la terminologie référencée de manière consensuelle dans le corpus des textes spécialisés	En entrée, le logiciel prend un corpus de textes spécialisés et,après l'analyse syntaxique, renvoie la liste des candidates termes qui sont « syntaxiquement plausibles »(ex. nom-adjectif, nom-nom etc.)

Annexe B

Liste de verbes liant les termes initiaux.

améliorer
approcher
assurer
atteindre
causer
comprendre
compromettre
concerner
conclure
conduire
conserver
constituer
contribuer
contrôler
courir
diagnostiquer
différer
dispenser
effectuer
éliminer
englober
entraîner
évaluer
exister
faciliter
fournir
garantir
incendier
induire
limiter
localiser
maintenir
mesurer
obtenir
passer
peser
présenter
produire
provenir
provoquer
qualifier
rater
réaliser
réduire
relever
rendre
répondre
reposer
risquer
signifier
subir
superviser
traiter
travailler
trouver
utiliser

Exemples des étiquettes correspondant aux concepts initiaux et leurs entités.

Dommage = {*dommage, dégradation, destruction, détérioration, détriment, lésion, perte*}	dommage radiologique, dommage au source radioactif ; dommage au matériel génétique ; dommage au tissu du vaisseau ; lésion cellulaire ; lésion cellulaire ; lésion cutané permanent ; détriment radiologique ;
Exposition= {*exposition, exposé*}	exposition chronique ; exposition contrôlable ; surexposition accidentel ; exposition thérapeutique ; exposition normal ; exposition durée ; exposition planifiée ; exposition localisée ; radioexposition limitée ; exposition prolongée ; exposition élevée ; exposition délibérée ; exposition planifiée ; surexposition localisée ; exposition aux rayonnements ;
Contrôle= {*contrôle, inspection, révision, sondage, surveillance*}	contrôle périodique ; contrôle finale ; contrôle administratif ; contrôle technique ; contrôle radiamétrique ; contrôle neutronique ; contrôle environnemental ; contrôle efficace du déchet ; contrôle annuel réglementaire ; contrôle individuel du travailleur ; inspection technique du source ; inspection réglementaire ; inspection planifiée ; cycle itératif de révision ; examen par sondage ; sondage des activités de maintenance ; sondage des mesures de prévention ; surveillance générale permanente ; surveillance radiologique ; surveillance radiologique du territoire ; surveillance écologique ; surveillance de la dose ; surveillance médicale ; surveillance permanent ; surveillance environnementale ; surveillance multilatérale ; surveillance suivie médicale ; surveillance spécialisée ; activité de surveillance ; zone de surveillance ; activité de surveillance
Population= {*population, habitant, gens, démographie*}	gens actifs ; gens ordinaires ; habitant concerné ; population homogène ; populations humaines du monde ; population adulte ; population subit une exposition ;

Protection={*protection, assistance, abri, secours, sauvegarde, préservation, défense, clôture, recommandation, soutien, appui, tutelle, couvercle, écran, verrouillage, conservation, parrainage, restauration*}	protection radiologique ; protection radiologique médicale ; protection radiologique des travailleurs ; protection radiologique des patients ; protection radiologique du public ; protection volumétrique robuste ; protection intrinsèque ; protection supplémentaire ; protection physique du centrale ;protection physique du combustible ; couvercle épais ; dispositif de verrouillage ; système de verrouillage ; écran de protection contre les rayonnements ; écran de protection radiologique ; clôture des sites ; clôture des centrales électronucléaires ; préservation de l'état du combustible nucléaire ; préservation des barrières artificielles ; installation nucléaire de défense ; organismes de parrainage ; conservation de déchets radioactifs ; conservation de sources radioactives ; radioactifs ; abri limitée ; restauration des sites ; restauration d'un environnement dégradé ; système de refroidissement de secours ; ventilation de secours ;
Rayonnement={*chaleur, lumière, irradiation, radiation, rayon, rayonnement*}	irradiation partielle du corps ; irradiation interne ; irradiation externe globale ; irradiation externe localisée ; irradiation prolongée ; irradiation localisée ; rayon cosmique ;
Risque = {*danger, menace*}	risque héréditaire ; risque radiologique ; risque stochastique ; risque potentiel de exposition ; danger potentiel de contamination ; danger radiologique ; danger radiologique imminent ; danger physique de rayonnement ; menace nucléaire ; menace de référence ; menace grave pour les populations ; menace posée par les rayonnements ;
Sûreté = {*sûreté, garantie, sécurité, précaution*}	précaution adéquat de radioprotection ; garantie globale ; garantie de la sécurité nucléaire des déchets ; garantie de la sûreté ; garanties nucléaire ; garanties nucléaires de l'AIEA ; accord de garantie ; mesure de sécurité ; sécurité extérieure ; formation pratique en sécurité ; sécurité physique du site ; sécurité civile ; régime efficace de sécurité ; sécurité nucléaire ; sécurité nucléaire civile ; sécurité nucléaire des déchets ; sécurité nucléaire organisationnel ; sécurité antiterroriste ;

`Source` = {*source, point, mine, endroit, cause*}	source interne ; source externe ; source interne de photons ; source orpheline ; type particulier de source ; source lumineuse au tritium ; source radioactive ; source typique ; source industrielle de cobalt ; source froide ; source naturelle de rayonnement ; source vulnérable ; source terrestre de rayonnement ; source physique de rayonnement ; source puissante de neutrons ; source externe ambiante ; source médicale de chlorure ; source médicale de cobalt ; source de sélénium ; source de rayonnement ; mine de minerais radioactifs ; mine d'uranium ;
`Personnel` = {*personnel*}	personnel de surveillance ; personnel professionnel ; personnel paramédical ; personnel médical ; personnel technique ; personnel technique auxiliaire ; personnel technique supplémentaire ;

Annexe C

Jusqu'au début des années 2000, une multitude de langages de représentation des ontologies a été proposée. Actuellement on distingue trois conceptions principales : langages formels techniques à base des langues naturelles, langages orientés machines et langages universels, *Kuhn* (2014), *Burakova et al.* (2014).

Les langages de représentation des ontologies ont tendance à évoluer vers un l'équilibre entre l'expressivité, proche des langues naturelles, et la possibilité de calculs. Les premiers langages utilisaient déjà la logique du première ordre. Les langages plus récents permettent la modélisation dans la logique descriptive mais ils s'appuient toujours à la fois sur les cadres sémantiques et sur les réseaux sémantiques qui ne possèdent pas leurs propres constructions formelles.

Langages formels

De nos jours, l'anglais est considéré comme le moyen universel de la communication entre les systèmes techniques, pour leur interopérabilité. Mais, comme toute langue naturelle, l'anglais, même simplifié, n'est pas sans défauts ; l'un des plus graves est l'ambiguïté des mots et, par conséquent l'ambiguïté de l'interprétation des textes, ce qui est inacceptable pour la communication entre les systèmes d'information.

Pour éviter l'ambiguïté, les langages formels utilisent les règles garantissant l'interprétation univoque de la sémantique des textes. On peut citer les règles suivantes :

- la définition anticipée du vocabulaire univoque (les significations des mots utilisés sont connues *a priori*) ;
- l'utilisation des énoncés simples ;
- l'interdiction de mettre le sujet à la fin de la proposition ;
- l'interdiction de mettre le complément direct avant le sujet ;
- l'interdiction d'utiliser de l'ordre inverse des éléments de prédicat composé.

Les langages de cette catégorie sont :

- ***STE***[2] (Simplified Technical English), connu aussi sous l'abréviation ***ASD STE100*** (AeroSpace and Defence Industries Association of Europe) est la spécification de l'anglais conçue pour accroître la clarté sémantique des textes. STE a été créé comme langage réglementé afin d'écrire les documents d'accompagnement dans le domaine aérospatial et de la défense. Il a sa propre grammaire, aux règles syntaxiques rigoureuses, il dispose d'un lexique limité et définit la liste des mots interdits. Dans le STE l'utilisation des temps et des formes de verbes est également limitée. Aujourd'hui, cette norme est utilisée par des compagnies telles que British Aerospace, Airbus, The Boeing Company, Lockheed Martin, Rolls Royce, Dassault et Saab Aerosystems.

- ***STR***[3] (Simplified Technical Russian) est le langage technique russe simplifié, élaboré pour lier la documentation de l'industrie aérospatiale russophone à la documentation en anglais.

- ***Gellish*** est un langage qui sert à l'échange d'information entre les différents systèmes de gestion des processus d'affaires, la description des produits et des services dans toutes les étapes du cycle de vie d'un produit, *Van Renssen* (2005).

Langages « orientés machines »

Les langages de cette catégorie sont les prédécesseurs de tous les autres langages formels. Les premiers essais de création de langages machines pour la réalisation des ontologies datent des années soixante-dix. Ils sont construits sur un outil mathématique, notamment les langages basés sur la logique descriptive, les langages basés sur la logique du premier ordre et les langages basés sur les cadres. Deux langages-pionniers, encore utilisés aujourd'hui, sont KIF et CycL.

KIF (Knowledge Interchange Format) est une langue universelle servant à l'échange de données concernant un domaine d'intérêt. La sémantique de KIF est déclarée de manière explicite et permet des expressions variées moyennant de calcul des prédicats de premier ordre. KIF permet la présentation des connaissances par la description des objets, des relations, des fonctions et des règles de productions, *Martin* (2002).

CycL (Cycorp Language) est un langage formel dont la syntaxe se base également sur la logique de premier ordre. Les unités de vocabulaire du CycL peuvent se grouper en expressions qui servent à former les assertions dans la base de données de Cyc. Soulignons que CycL a été utilisé lors de construction de WordNet ; son code source est ouvert.

2. Toutes les informations sont disponibles sur le site http ://www.asd-ste100.org
3. Les informations disponibles sur le site http ://s1000d.ru/userforum/presentations/Day_3_Track1_04_Simplified_Russian.pdf

Langages universels

Resource Description Framework, RDF

RDF (Resource Description Framework) est un modèle de représentation de données à la base de la conception du Web Sémantique, *Brickley et Guha* (2004). RDF est un instrument universel qui demande cependant un réglage additionnel pour la solution des tâches concrètes. Le moyen principal de spécialisation consiste en extension de RDF à l'aide de dictionnaires comme, par exemple le RDF-Schema ou l'OWL.

La structure de base en RFD est un triplet qui comprend un sujet, un prédicat et un objet (*S,P,O*). L'ensemble de tous les triplets forme le RDF-graphe. Les sujets et les objets sont ses sommets, et les prédicats correspondent aux arcs. La flèche d'un arc est toujours orientée du sujet vers l'objet (cf. la figure 6.1).

Chaque triplet correspond à une assertion liant *S*, *P* et *O*. Le sujet et le prédicat s'identifient à l'aide des URIs. L'objet est une ressource qui peut être identifiée par un URI ou par un RDF-littéral.

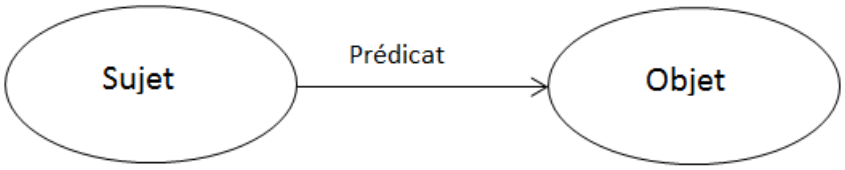

Figure 6.1: Triplet de RDF.

Dans la pratique, deux formes de présentation des graphes de RDF sont utilisées : soit sous forme de document en XML, soit sous forme de la syntaxe N3 (N Triple).

L'ouverture et l'extensibilité de RDF conduisent à des difficultés pour assurer l'intégrité et la cohérence des RDF-descriptions parce que RDF n'interdit pas de produire des déclarations non compatibles avec les autres. C'est un des inconvénients de RDF.

RDFS

Le RDF ne présente aucun mécanisme pour la description des attributs de ressources, ni les relations. Pour pallier cela, on a créé RDFS qui définit les classes, les propriétés et les autres ressources.

RDFS est une extension sémantique de RDF qui met en place les mécanismes pour décrire les groupes de ressources connectées et leurs relations. Toutes les définitions du RDFS sont réalisées sous format RDF. Les nouveaux termes introduits par RDFS tels que « domaine » ou « plage » de propriété sont des ressources de RDF. Les classes et les propriétés des dictionnaires de RDF ressemblent aux types de langages orientés objets tels que Java, mais l'aspect central de RDF est la définition des propriétés, et non des classes.

Les propriétés dans RDF sont définies comme des paires (*domaine*, *range*) où le *domaine* correspond à un ensemble de classes de RDF auxquelles sont appliquées certaines propriétés et le *range* définit l'ensemble acceptable de ressources qui jouent le rôle des valeurs de la propriété. Cela signifie que les classes sont ouvertes parce que leurs propriétés sont définies séparément.

SPARQL (SPARQL Protocol and RDF Query Language) est un langage permettant des requêtes sur des données codées sous forme RDF. **SPARQL** est recommandé par le consortium W3C pour le web sémantique. Les résultats des requêtes en SPARQL peuvent être également présentées sous forme de graphes de RFD.

UML (Unified Modeling Language) est un langage de description graphique, utilisé dans le domaine de l'ingénierie logicielle pour la modélisation des objets. Les désignations graphiques sont utilisées dans l'UML afin de construire un modèle abstrait pour un système d'information (UML-model). L'objectif de l'UML est la visualisation, la documentation et la description du processus de construction des systèmes d'information.

Familles des Web Ontology Languages

OWL (Web Ontology Language) est un language de présentation des ontologies conçu pour traiter l'information stockée dans la toile. OWL peut être envisagé comme doté d'un vocabulaire permettant d'élargir l'ensemble des termes définis par RDFS. La création de l'OWL a été dictée par la nécessité de l'unification de la présentation des connaissances dans la Toile. Deux versions, notamment, ***XOL*** (XML-based ontology exchange language) et ***DAML+OIL*** (DARPA Agent Markup Language et Ontology Inference Laye) ont précédé OWL qui y a puisé ses principes : l'ensemble de primitives tiré des langages basés sur XOL, la sémantique formelle et le mécanisme d'inférence dans le cadre de la logique descriptive, la syntaxe de RDFS garantissant la représentation standardisée.

Le schéma de prédécesseurs d'OWL est présenté dans la figure. 6.2. Chacun des langages du niveau supérieur de cette pyramide continue de se développer, mais le langage le plus répandu est OWL. Dès 2004, il est devenu la recommandation de W3C (World Wide Web Consortioum), *McGuinness et van Harmelen* (2004).

Il existe les trois sous-catégories d'OWL, notamment, *OWL Lite*, *OWL DL* et *OWL Full*. Chaque version est une extension de la spécification précédente, et on peut constater que toute ontologie réalisée en OWL Lite est également une ontologie en OWL DL ; de même toute ontologie en OWL DL peut être considérée comme une ontologie en OWL Full.

OWL Lite est la version la plus simple de la famille OWL ; elle assure la hiérarchie de concepts, et plusieurs contraintes.

OWL DL exploite l'appareil de logique descriptive (d'où son sigle). Cette version d'OWL possède deux propriétés principales : la complétude, qui signifie que toute

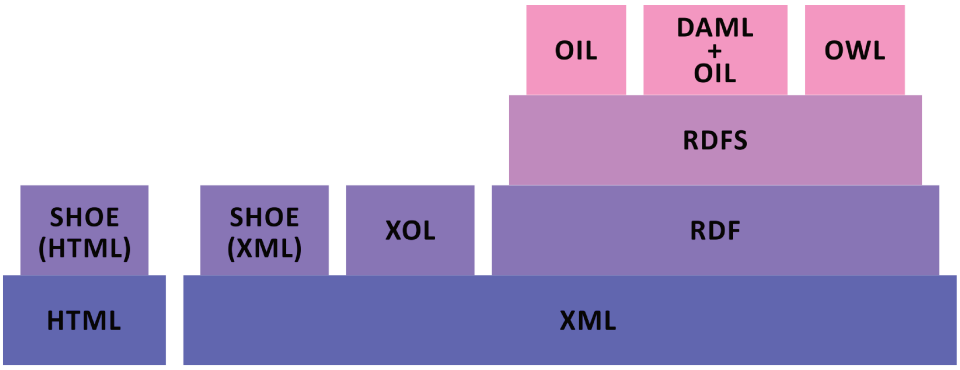

Figure 6.2: La hiérarchie historique des langages de représentation des ontologies.

inférence est calculable, et la garantie de solvabilité, qui signifie que tous les calculs se réalisent en temps fini.

OWL Full assure le maximum d'expressivité, mais ne garantit pas la solvabilité.

Définition des classes

Les classes d'OWL sont définies par l'utilisation de l'élément $owl : Class$ qui est un sous-classe de $rdfs : Class$. Par exemple, dans l'ontologie sur la sécurité radiologique on définit la classe « `Protection` » (cf. la figure 6.3) :

```
<owl:Class rdf:about="&radioprotection;Source">
    <label xml:lang="en">source </label>
    <label xml:lang="fr">source</label>
    <label xml:lang="ru">Источник_Ионизирующего_Излучения</label>
    <comment xml:lang="en">ionizing radiatin source</comment>
    <comment xml:lang="fr">source des raynnement ionisants</comment>
    <comment xml:lang="ru">источник ионизирующего излучения</comment>
</owl:Class>
```

Figure 6.3: Classe « `Protection` ».

Il y a deux classes spécifiques d'OWL : *Thing* et *Nothing*. La première est la super-classe de toute autre classe OWL, et la deuxième est la sous-classe de toute autre classe OWL.

Définition des propriétés et restrictions

Les propriétés en OWL permettent de définir les relation binaires entre les éléments d'une ontologie. On distingue deux catégories de propriétés : les propriétés d'objets qui représentent les relations entre les entités de deux classes, et les propriétés du type de données attribuant des valeurs aux entités.

Ontologies, logique de premier ordre et logique descriptive

L'inconvénient de la logique du premier ordre pour représenter les connaissances ontologiques est qu'elle ne garantit pas la solvabilité, contrairement à la logique descriptive.

La plus petite logique descriptive A fait référence à un concept atomique, tandis que C et D vont référencer des concepts complexes. R correspond à une relation (ou prédicat) binaire pour laquelle on utilise la dénomination *rôle* en logique descriptive.

$$C, D := A \mid \top \mid \bot \mid \neg C \mid C \sqcap D \mid C \sqcup D \mid \forall R.C \mid \exists R.C$$

Les expressions listées ont les significations suivantes :

- $\top$ est l'ensemble de tous les objets (things), couramment appelé *top* ;
- $\bot$ est l'ensemble vide, couramment appelé *bottom* ;
- $\neg C$ est l'ensemble de tous les objets qui ne sont pas membres de la classe C ;
- $C \sqcap D$ est l'ensemble des tous les objets qui sont membres des classes C et D ;
- $C \sqcup D$ est l'ensemble des tous les objets qui sont membres de la classe C ou de la classe D ;
- $\forall R.C$ est l'ensemble des tous les objets qui sont liés aux membres de la classe C par la relation R ;
- $\exists R.C$ est l'ensemble des tous les objets qui sont liés à au moins à un membre de la classe C par la relation R.

Bibliographie

Ackoff, R. (1989), From data to wisdom, *Journal of Applied Systems Analysis*, *16*, 3–9.

Agrawal, R., T. Imielinski, et A. Swami (1993), Mining association rules between sets of items in large databases, in *In : Proceedingd of the 1993 ACM SIGMOD International Conference on Management of Data, Washington DC (USA)*, pp. 207–216.

Ahmad, K., L. Gillam, et L. Tostevin (1999), University of surrey participation in trec8 : Weirdness indexing for logical document extrapolation and retrieval (wilder), in *The Eighth Text REtrieval Conference (TREC-8)*, Gaithersburg, Maryland.

Ananiadou, S. (1994), A methodology for automatic term recognition, in *Proceedings of the 15th Conference on Computational Linguistics - Volume 2*, COLING '94, pp. 1034–1038, Association for Computational Linguistics, Stroudsburg, PA, USA, doi :10.3115/991250.991317.

Anselmo, P., V. Felisa, et G. Julio (2001), Corpus-based terminology extraction applied to information access, in *In Proceedings of Corpus Linguistics 2001*, pp. 458–465.

Apresjan, Y. (1973), *Eléments sur les idées et les méthodes de la linguistique structurale contemporaine : Avec la collab. de S.Golopentia-Eretescu*, Monographies de linguistique mathématique, Dunod.

Aussenac-Gilles, N. (2005), Méthodes ascendantes pour l'ingénierie des connaissances, Habilitation à diriger des recherches, Université Paul Sabatier, Toulouse, France.

Aussenac-Gilles, N., S. Despres, et S. Szulman (2008), The terminae method and platform for ontology engineering from texts, in *Proceedings of the 2008 Conference on Ontology Learning and Population : Bridging the Gap Between Text and Knowledge*, pp. 199–223, IOS Press, Amsterdam, The Netherlands, The Netherlands.

Baader, F., D. Calvanese, D. L. McGuinness, D. Nardi, et P. F. Patel-Schneider (2010), *The Description Logic Handbook : Theory, Implementation and Applications*, 2nd ed., Cambridge University Press, New York, NY, USA.

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.

Barriere, C. (2008), Pattern-based approaches to semantic relation extraction : A state-of-the-art, *Terminology*, *14*(1), 1–19.

Basili, R., A. Moschitti, M. Pazienza, et F. Zanzotto (2001), A contrastive approach to term extraction, in *Proceedings of the 4th Terminology and Artificial Intelligence Conference (TIA)*, France.

Benjamin, P. C., C. P. Menzel, R. J. Mayer, et F. F. et. al (1994), Idef5 ontology description capture method report, *Knowledge based systems, inc method report*, Knowledge Based Systems, Inc.

Biemann, C. (2005), Ontology learning from text : A survey of methods, *LDV Forum*, *20*(2), 75–93.

Blei, D. M., et J. D. Lafferty (2006), Correlated topic models, in *In Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120, MIT Press.

Bloch, I., et H. Maître (2002), Les méthodes de raisonnement dans les images, polycopié du module RASIM de la brique VOIR, *Tech. Rep. 2002TSI36*, Ecole Nationale Supérieure des Télécommunications.

Bloomfield, L. (1933), *Language*, Holt, New York.

Bobrow, D. G., M. J. Stefik, et . Xerox (Palo Alto, CA US) (1983), The loops manual.

Bolshakova, E., N. Loukachevitch, et M. Nokel (2013), Topic models can improve domain term extraction, in *Advances in Information Retrieval, 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings*, *Lecture Notes in Computer Science*, vol. 7814, pp. 684–687, TIERGARTENSTRASSE 17, HEIDELBERG, GERMANY,D-69121, Tiergartenstrasse 17, Heidelberg, Germany, D-69121.

Bouma, G. (2009), Normalized (pointwise) mutual information in collocation extraction, in *From Form to Meaning : Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, vol. Normalized, pp. 31–40, Tübingen.

Bourigault, D., C. Jacquemin, et M. L'Homme (2001), *Recent Advances in Computational Terminology*, Natural language processing, J. Benjamins Publishing Company.

Brickley, D., et R. V. Guha (2004), Rdf vocabulary description language 1.0 : Rdf schema, *W3C Recommendation*, *10*.

Brown, S. W., D. Dligach, et M. Palmer (2011), Verbnet class assignment as a wsd task, in *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11, pp. 85–94, Association for Computational Linguistics, Stroudsburg, PA, USA.

Budanitsky, A., et G. Hirst (2001), Semantic distance in wordnet : An experimental, application-oriented evaluation of five measures, in *In Workshop on WordNet and other lexical resources ; second meeting of the North American Chapter of the Association for Computational Linguistics*.

Buitelaar, P., et P. Cimiano (2008), *Ontology Learning and Population : Bridging the Gap Between Text and Knowledge*, vol. 167, Ios Press Inc.

Buitelaar, P., et B. Magnini (2005), Ontology learning from text : An overview, in *In Paul Buitelaar, P., Cimiano, P., Magnini B. (Eds.), Ontology Learning from Text : Methods, Applications and Evaluation*, pp. 3–12, IOS Press.

Buitelaar, P., et M. Sintek (2004), Ontolt version 1.0 : Middleware for ontology extraction from text, in *The Third International Semantic Web Conference (ISWC2004)*.

Buitelaar, P., P. Cimiano, J. McCrae, E. Montiel-Ponsoda, et T. Declerck (2011), Ontology lexicalisation : The lemon perspective, in *Proceedings of theWorkshops 9th International Conference on Terminology and Artificial Intelligence*, ontology Engineering Group ? OEG.

Burakova, E., N. Borgest, et M. Korovin (2014), Ontology description languages for high-tech filds of applied engineering, *Вестник Самарского государственного аэрокосмического университета им. академика С.П. Королёва (национального исследовательского университета)*, *3 (45)*, 144–158.

Cabré, M., M. Cormier, et J. Humbley (1998), *La terminologie : théorie, méthode et applications*, Collection U. : Série linguistique, Presses de l'Université d'Ottawa.

Charlet, J., A. Baneyx, O. Steichen, I. Alecu, C. Daniel-Le Bozec, C. Bousquet, et M. Jaulent (2009), Utiliser et construire des ontologies en médecine. le primat de la terminologie, *Technique et Science Informatiques*, *28*(2), 145–171.

Chklovski, T., et P. Pantel (2005), Global path-based refinement of noisy graphs applied to verb semantics, in *Natural Language Processing IJCNLP 2005, Second International Joint Conference, Jeju Island, Korea, October 11-13, Proceedings*, pp. 792–803.

Chomsky, N., et M. Braudeau (1969), *Structures syntaxiques : "Syntactic structures" . Traduit de l'anglais par Michel Braudeau*, éditions du Seuil.

Church, K., et W. Gale (1995), Inverse document frequency (idf) : A measure of deviations from poisson, in *Proceedings of the ACL 3rd Workshop on Very Large Corpora*, pp. 121–130.

Church, K. W., et P. Hanks (1990), Word association norms, mutual information, and lexicography, *Comput. Linguist.*, *16*(1), 22–29.

Ciaramita, M., A. Gangemi, E. Ratsch, J. Šaric, et I. Rojas (2005), Unsupervised learning of semantic relations between concepts of a molecular biology ontology, in *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, pp. 659–664, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Cimiano, P., et J. Völker (2005), Text2onto - a framework for ontology learning and data-driven change discovery, in *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB 2005)*, *Lecture Notes in Computer Science*, vol. 3513, edited by A. Montoyo, R. Munoz, et E. Metais, pp. pp. 227–238, Springer, Alicante, Spain.

Cimiano, P., A. Hotho, et S. Staab (2005), Learning concept hierarchies from text corpora using formal concept analysis, *J. Artif. Int. Res.*, *24*(1), 305–339.

Cimiano, P., J. Völker, et R. Studer (2006), Ontologies on demand ? a description of the state-of-the-art, applications, challenges and trends for ontology learning from text.

Cimiano, P., C. Unger, et J. P. McCrae (2014), *Ontology-Based Interpretation of Natural Language*, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers.

Cunningham, H., D. Maynard, et K. Bontcheva (2011), *Text Processing with GATE*, Gateway Press CA.

Daille, B. (1994), *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*, Université de Paris 7.

Daojian, Z., L. Kang, L. Siwei, Z. Guangyou, et Z. Jun (2014), Relation classification via convolutional deep neural network, in *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference : Technical Papers, August 23-29, 2014, Dublin, Ireland*, pp. 2335–2344.

Daudaravicius, V., et al. (2004), Gravity counts for the boundaries of collocations, *INTERNATIONAL JOURNAL OF CORPUS LINGUISTICS*, pp. 321–348.

de Chalendar, G., et B. Grau (2000), Svetlan' or how to classify words using their context., in *EKAW*, pp. 203–216.

de Marneffe, M.-C., B. MacCartney, et C. D. Manning (2006), Generating typed dependency parses from phrase structure parses, in *In proceedings of the International Conference of Language Resources end Evaluation (LREC)*, pp. 449–454.

Declerck, G., A. Baneyx, X. Aimé, et J. Charlet (2012), A quoi servent les ontologies fondationnelles ?, in *23èmes Journées francophones d'Ingénierie des Connaissances (IC 2012)*, pp. 67–82, Paris, France.

DeNicola, A., M. Missikoff, et R. Navigli (2005), A proposal for a unified process for ontology building : UPON, in *Database and Expert Systems Applications, 16th International Conference, DEXA 2005, Copenhagen, Denmark, August 22-26, 2005, Proceedings*, pp. 655–664.

D.G., B., et T. Winograd (1985), An overview of krl, a knowledge representation language, in *Readings in Knowledge Representation*, edited by R. J. Brachman et H. J. Levesque, pp. 263–285, Kaufmann, Los Altos, CA.

Dhillon, I., J. Kogan, et C. Nicholas (2003), Feature selection and document clustering, in *Survey of Text Mining*, edited by M. W. Berry, pp. 73–100, Springer.

Dobrov, B., et N. Lukashevych (2009), Тезаурус РуТез как ресурс для решения задач информационного поиска, in *Труды Всероссийской Конференции Знания-Онтологии-Теории, ЗОНТ-09, Новосибирск, Том.1*, pp. 250–259, Новосибирск.

Dobrov, B., I. V.V., N. Lukashevych, et V. Solovjev (2008), *Ontologies and Tezauruses Онтологии и тезаурусы : модели, инструменты, приложения*, БИНОМ. Лаборатория знаний.

Dretske, F. (2000), *Perception, Knowledge and Belief : Selected Essays*, Cambridge University Press.

Drumond, L., et R. Girardi (2008), A survey of ontology learning procedures, in *WONTO*, *CEUR Workshop Proceedings*, vol. 427, edited by F. L. G. de Freitas, H. Stuckenschmidt, H. S. Pinto, A. Malucelli, et s. Corcho, CEUR-WS.org.

Drymonas, E., K. Zervanou, et E. G. M. Petrakis (2010), Unsupervised ontology acquisition from plain texts : The ontogain system, in *Proceedings of the Natural Language Processing and Information Systems, and 15th International Conference on Applications of Natural Language to Information Systems*, NLDB'10, pp. 277–287, Springer-Verlag, Berlin, Heidelberg.

Dunning, T. (1993), Accurate methods for the statistics of surprise and coincidence, *Computational Linguistics*, *19*(1), 61–74.

Fabre, C., et B. D. (2006), Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques, in *Actes de la 13e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2006)*, pp. 121–129.

Falk, I., D. Bernhard, C. Gérard, et R. Potier-Ferry (2014), Étiquetage morpho-syntaxique pour des mots nouveaux, in *Actes des 21e Conférence TALN2014 (Traitement Automatique des Langues Naturelles)*, edited by B. Bigi, pp. 431–437, Marseille, France.

Faure, D., et C. Nédellec (1998), Asium : learning subcategorization frames and restrictions of selection.

Fellbaum, C. (Ed.) (1998), *WordNet : an electronic lexical database*, MIT Press.

Fillmore, C. J., et B. T. S. Atkins (1992), *Towards a frame-based lexicon : The semantics of RISK and its neighbors*, pp. 75–102, Lawrence Erlbaum Associates, Hillsdale.

Fillmore, C. J., et C. Baker (2010), A frames approach to semantic analysis, *The Oxford Handbook of Linguistic Analysis*, pp. 313–339.

Flati, T., et R. Navigli (2013), Spred : Large-scale harvesting of semantic predicates, in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL*, vol. 1, pp. 1222–1232, Sofia, Bulgaria.

Foo, J. (2012), Computational terminology : Exploring bilingual and monolingual term extraction.

Fortuna, B., M. Grobelnik, et D. Mladenić (2006), Semi-automatic data-driven ontology construction system, *Information Systems*.

Frank, R. (2004), Phrase structure composition and syntactic dependencies.

Frantzi, K., S. Ananiadou, et H. Mima (2000), Automatic recognition of multi-word terms :. the c-value/nc-value method, *International Journal on Digital Libraries*, *3*(2), 115–130.

Frantzi, K. T., et S. Ananiadou (1997), Automatic term recognition using contextual cues, in *In Proceedings of 3rd DELOS Workshop*.

Gangemi, A., G. Steve, et F. Giacomelli (1996), Onions : An ontological methodology for taxonomic knowledge integration, in *ECAI-96 Workshop on Ontological Engineering*.

Ganter, B., et R. Wille (1997), *Formal Concept Analysis : Mathematical Foundations*, 1st ed., Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Gaussier, E., et F. Yvon (Eds.) (2011), *Modèles statistiques pour l'accès à l'information textuelle*, Hermès, Paris.

Geeraerts, D. (2001), Hundred years of lexical semantics, *Versus. Quaderni di studi semiotici*, *88/89*, 63–87.

Geeraerts, D. (2006), *Cognitive Linguistics : Basic Readings*, Cognitive Linguistics Research [CLR], Mouton de Gruyter, Berlin/New York.

Gelbukh, A., G. Sidorov, E. Lavin-Villa, et L. Chanona-Hernandez (2010), Automatic term extraction using log-likelihood based comparison with general reference corpus, in *Proceedings of the Natural Language Processing and Information Systems, and 15th International Conference on Applications of Natural Language to Information Systems*, NLDB'10, pp. 248–255, Springer-Verlag, Berlin, Heidelberg.

Gildea, D., et D. Jurafsky (2002), Automatic labeling of semantic roles, *Comput. Linguist.*, *28*(3), 245–288.

Girju, R., et D. Moldovan (2002), Text mining for causal relations, in *In Proceedings of the FLAIRS Conference*, pp. 360–364.

Girju, R., D. Moldovan, M. Tatu, et D. Antohe (2005), On the semantics of noun compounds, *Comput. Speech Lang.*, *19*(4), 479–496.

Gómez-Pérez, A., et D. Manzano-Macho (2003), A survey of ontology learning methods and techniques, *Deliverable 1.5*, OntoWeb Consortium.

Gross, G. (2008), Les classes d'objets, *Lalies*, *28*, 111–165.

Gross, M. (1975), *Méthodes en syntaxe : régime des constructions complétives*, Actualités scientifiques et industrielles, Hermann.

Gruber, T. R. (1995), Toward principles for the design of ontologies used for knowledge sharing, *Int. J. Hum.-Comput. Stud.*, *43*(5-6), 907–928.

Grüninger, M., et M. Fox (1995), Methodology for the Design and Evaluation of Ontologies, in *IJCAI'95, Workshop on Basic Ontological Issues in Knowledge Sharing, April 13, 1995*.

Guarino, N., et C. A. Welty (2004), *An Overview of OntoClean*, pp. 151–171, Springer Berlin Heidelberg, Berlin, Heidelberg.

GuoDong, Z., S. Jian, Z. Jie, et Z. Min (2005), Exploring various knowledge in relation extraction, in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pp. 427–434, Association for Computational Linguistics, Stroudsburg, PA, USA.

Hahn, U., M. Romacker, et S. Schulz (1999), Discourse structures in medical reports - watch out ! the generation of referentially coherent and valid text knowledge bases in the medsyndikate system, *International Journal of Medical Informatics*, *53*(1), 1–28, cited By 19.

Hailpern, B., et P. Tarr (2006), Model-driven development : The good, the bad, and the ugly, *IBM Syst. J.*, *45*(3), 451–461.

Happel, H.-J., et S. Seedorf (2006), Applications of ontologies in software engineering, in *International Workshop on Semantic Web Enabled Software Engineering (SWESE'06)*, Athens, USA.

Hearst, M. A. (1992), Automatic acquisition of hyponyms from large text corpora, in *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, pp. 539–545, Association for Computational Linguistics, Stroudsburg, PA, USA.

Heid, U. (2007), Valency data for natural language processing : what can the emph Valency Dictionary of English provide ?, in *Valency – Theoretical, Descriptive and Cognitive Issues*, edited by K. G.-V. Thomas Herbst, pp. 365 – 382, Berlin : de Gruyter.

Héon, M., G. Paquette, et J. Basque (2009), Méthodologie assistée de conception d'une ontologie à partir d'une conceptualisation consensuelle semi-formelle, in *IC 2009 : 20es Journées Francophones d'Ingénierie des Connaissances*, p. 61, Tunisia.

Horridge, M., H. Knublauch, A. Rector, R. Stevens, et C. Wroe (2004), A practical guide to building owl ontologies using the protege-owl plugin and co-ode tools edition 1.0, *The University of Manchester*.

Huddleston, R., et G. K. Pullum (2005), *A Student's Introduction to English Grammar*, 322 pp., Cambridge University Press.

Kavalec, M., A. Maedche, et V. Svátek (2004), Discovery of lexical entries for non-taxonomic relations in ontology learning, in *SOFSEM*, vol. 2932, edited by P. van Emde Boas, J. Pokorný, M. Bieliková, et J. Stuller, pp. 249–256, Springer.

Kayser, D. (1997), *La représentation des connaissances*, Collection Informatique, Hermès.

Kipper, K., A. Korhonen, N. Ryant, et M. Palmer (2006), Extending VerbNet with novel verb classes., in *Proceedings of 5th international conference on Language Resources and Evaluation*, Genova, Italy.

Kipper, K., A. Korhonen, N. Ryant, et M. Palmer (2007), A large-scale classification of english verbs, *Language Resources and Evaluation*, doi :10.1007/s10579-007-9048-2.

Kister, L., E. Jacquey, et B. Gaiffe (2011), Liens conceptuels et relations sémantiques : proposition de représentation des connaissances en sciences du langage., in *Ingénierie des connaissances*, pp. 1–3, Chambéry, France.

Kitamura, M., et Y. Matsumoto (1996), Automatic extraction of word sequence correspondences in parallel corpora, in *Proceedings of the 4th annual workshop on very large corpora(WVLC-4)*, pp. 79–87.

Kuhn, T. (2014), A survey and classification of controlled natural languages, *Computational Linguistics*, *40*(1), 121–170.

Kullback, S. (1997), *Information Theory And Statistics*, Dover Pubns.

Kurz, D., et F. Xu (2002), Text mining for the extraction of domain relevant terms and term collocations, in *Proceedings of the International Workshop on Computational Approaches to Collocations*, Vienna.

Lassila, O., et D. McGuinness (2001), The role of frame-based representation on the semantic web, *Tech. rep.*, Knowledge Systems Laboratory Report KSL-01-02, Stanford University, Stanford (USA).

Lehmann, J. (2007), Hybrid learning of ontology classes, in *Proc. of the 5th Int. Conference on Machine Learning and Data Mining MLDM*, *Lecture Notes in Computer Science*, vol. 4571, pp. 883–898, Springer.

Lenat, D. B. (1995), Cyc : A large-scale investment in knowledge infrastructure, *Commun. ACM*, *38*(11), 33–38, doi :10.1145/219717.219745.

Levin, B. (1993), *English verb classes and alternations : a preliminary investigation*, Chicago Press, University.

L'Homme, M.-C. (2004), *La terminologie : principes et techniques*, 282 pp., Les Presses de l'Université de Montréal.

Liu, L., J. Kang, J. Yu, et Z. Wang (2005), A comparative study on unsupervised feature selection methods for text clustering, in *Proc. 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering*, pp. 597–601.

Liu, X., S. Zhang, F. Wei, et M. Zhou (2011), Recognizing named entities in tweets, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies - Volume 1*, HLT '11, pp. 359–367, Association for Computational Linguistics, Stroudsburg, PA, USA.

Léon, J., P. De Brabanter, et J.-M. Fortis (2009), *Histoire Epistémologie Langage : Mathématisation du langage au 20e siècle*, np pp., SHESL.

López, M. F., A. Gómez-Pérez, J. P. Sierra, et A. P. Sierra (1999), Building a chemical ontology using methontology and the ontology design environment, *IEEE Intelligent Systems*, *14*(1), 37–46.

López, M. F., A. G. Pérez, et M. D. R. Amaya (2000), *Knowledge Engineering and Knowledge Management Methods, Models, and Tools : 12th International Conference, EKAW 2000 Juan-les-Pins, France, October 2–6, 2000 Proceedings*, chap. Ontology's Crossed Life Cycles, pp. 65–79, Springer Berlin Heidelberg, Berlin, Heidelberg.

Loukachevitch, N. (2011), *Thesauri in information retrieval tasks Тезаурусы в задачах информационного поиска*, 512 pp., Издательство МГУ Москва.

Maedche, A., et S. Staab (2000), Mining ontologies from text, in *Proc. of Knowledge Engineering and Knowledge Management (EKAW 2000)*, LNAI 1937, Springer.

Makki, J., A.-M. Alquier, et V. Prince (2008), Ontology Population via NLP Techniques in Risk Management, in *ICSWE : Fifth International Conference on Semantic Web Engineering*, vol. 1, pp. 079–085, Heidelberg, Germany.

Malone, J., et H. Parkinson (2010), Reference and application ontologies, *Ontogenesis*.

Martin, P. (2002), *Knowledge Representation in CGLF, CGIF, KIF, Frame-CG and Formalized-English*, pp. 77–91, Springer Berlin Heidelberg, Berlin, Heidelberg.

Mathieu-Colas, M. (2006), Les classes de verbes : syntaxe et sémantique, in *Le traitement du lexique. Catégorisation et Actualisation*, edited by J. B. et Salah MEJRI, pp. 10–24, Université de Sousse (Tunisie) et Université Paris 13, Sousse, Tunisia.

McCrae, J., D. Spohr, et P. Cimiano (2011), Linking lexical resources and ontologies on the semantic web with lemon, in *Proceedings of the 8th Extended Semantic Web Conference on The Semantic Web : Research and Applications - Volume Part I*, ESWC'11, pp. 245–259, Springer-Verlag, Berlin, Heidelberg.

McGuinness, D. L. (2003), Ontologies come of age, in *Spinning the Semantic Web*, edited by D. Fensel, J. A. Hendler, H. Lieberman, et W. Wahlster, pp. 171–194, MIT Press.

McGuinness, D. L., et F. van Harmelen (2004), Owl web ontology language overview, *Tech. rep.*, W3C.

Miller, G. A. (1995), Wordnet : A lexical database for english, *Communications of the ACM*, *38*, 39–41.

Miller, J., et J. Mukerji (2003), Mda guide version 1.0.1, *Tech. rep.*, Object Management Group (OMG).

Minsky, M. (1980), A framework for representing knowledge, in *Frame Conceptions and Text Understanding*, edited by D. Metzing, pp. 1–25, de Gruyter, Berlin.

Minsky, M. (1985), A framework for representing knowledge, in *Readings in Knowledge Representation*, edited by R. J. Brachman et H. J. Levesque, pp. 245–262, Kaufmann, Los Altos, CA.

Mondary, T., S. Després, A. Nazarenko, et S. Szulman (2008), Construction d'ontologies à partir de textes : la phase de conceptualisation, in *19èmes Journées Francophones d'Ingénierie des Connaissances (IC 2008)*, pp. 87–98, Nancy, France.

Morita, T., N. Fukuta, N. Izumi, et T. Yamaguchi (2008), Doddle-owl : Interactive domain ontology development with open source software in java., *IEICE Transactions*, *91-D*(4), 945–958.

Morlane-Hondère, F., et C. Fabre (2012), Étude des manifestations de la relation de méronymie dans une ressource distributionnelle, in *Actes de la Conference de Traitement Automatique des Langues Naturelle, TALN, Grenoble, France*, pp. 169–182.

Moro, A., A. Raganato, et R. Navigli (2014), Entity linking meets word sense disambiguation : a unified approach, *TACL*, *2*, 231–244.

Murphy, K. P. i. (2012), *Machine learning : a probabilistic perspective*, Adaptive computation and machine learning series, MIT Press, Cambridge (Mass.).

Nadeau, D., et S. Sekine (2007), A survey of named entity recognition and classification, *Linguisticae Investigationes*, *30*(1), 3–26, publisher : John Benjamins Publishing Company.

Nakagawa, H., et T. Mori (2003), Automatic term recognition based on statistics of compound nouns and their components, *Terminology*, *9*(2), 201–219.

Navigli, R., et S. P. Ponzetto (2012), Babelnet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, *Artif. Intell.*, *193*, 217–250.

Nayhanova, L. (2008), Основные типы семантических отношений между терминами предметной области, *Известия высших учебных заведений. Поволжский регион. Технические науки*, *1*, 10.

Neches, R., R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator, et W. R. Swartout (1991), Enabling technology for knowledge sharing, *AI Mag.*, *12*(3), 36–56.

Nenkova, A., et K. McKeown (2012), A survey of text summarization techniques, in *Mining Text Data*, pp. 43–76, Springer.

Neveu, F. (2008), Pour une description terminographique des sciences du langage, in *Cahiers du CIEL Langues de spécialité*, edited by P. V. E. Publication de l'Université, p. à paraître, C. Cortès.

Nicola, A. D., M. Missikoff, et R. Navigli (2008), A software engineering approach to ontology building, *Information Systems Journal*, *34*(2), 258–275.

Niles, I., et A. Pease (2003), Linking lexicons and ontologies : Mapping wordnet to the suggested upper merged ontology, in *Proceedings of the International Conference on Information and Knowledge Engineering (IKE 03), LAS VEGAS*, pp. 412–416.

Nokel, M., et N. Loukachevitch (2013), An experimental study of term extraction for real information-retrieval thesauri, in *Proceedings of 10th International Conference on Terminology and Artificial Intelligence*, pp. 69–76.

Nokel, M. A., E. I. Bolshakova, et N. V. Loukachevitch (2012), Combining multiple features for single-word term extraction, *Компьютерная лингвистика и интеллектуальные технологии. По материалам конференции онференции Диалог-2012*, pp. 490–501.

Noy, N. F., et D. L. McGuinness (2001), Ontology Development 101 : A Guide to Creating Your First Ontology.

Noy, N. F., A. Chugh, W. Liu, et M. A. Musen (2006), *The Semantic Web - ISWC 2006 : 5th International Semantic Web Conference, ISWC 2006, Athens, GA, USA, November 5-9, 2006. Proceedings*, chap. A Framework for Ontology Evolution in Collaborative Environments, pp. 544–558, Springer Berlin Heidelberg, Berlin, Heidelberg.

Øhrstrøm, P., H. Schärfe, et S. L. Uckelman (2008), *Conceptual Structures : Knowledge Visualization and Reasoning : 16th International Conference on Conceptual Structures, ICCS 2008 Toulouse, France, July 7-11, 2008 Proceedings*, chap. Jacob Lorhard's Ontology : A 17th Century Hypertext on the Reality and Temporality of the World of Intelligibles, pp. 74–87, Springer Berlin Heidelberg, Berlin, Heidelberg.

Orobinska, O. (2012), Automatic method of domain ontology construction based on characteristics of corpora pos-analysis, *CoRR*, pp. 209–212.

Orobinska, O., J.-H. Chauchat, et N. Charonova (2013), Enrichissement d'une ontologie de domaine par extension des relations taxonomiques à partir de corpus spécialisé, in *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence TIA 2013*, edited by G. A. de Cea et N. Aussenac-Gilles, pp. 129–137.

Orobinska, O. O., et N. V. Sharonova (2011), Метод fca для остроения онтолонии на основе текствовго корпуса (ontology construction from text's corpus with fca), *Бионика интеллекта*, *2(76)*, 129 – 135.

Pang, B., et L. Lee (2008), Opinion mining and sentiment analysis, *Found. Trends Inf. Retr.*, *2*(1-2), 1–135.

Panton, K., C. Matuszek, D. Lenat, D. Schneider, M. Witbrock, N. Siegel, et B. Shepard (2006), *Ambient Intelligence in Everyday Life : Foreword by Emile Aarts*, chap. Common Sense Reasoning – From Cyc to Intelligent Assistant, pp. 1–31, Springer Berlin Heidelberg, Berlin, Heidelberg.

Paquette, G. (1996a), La modélisation par objets typés - une méthode de représentation pour les systèmes d'apprentissage et d'aide à la tâche, *Sciences et Techniques Educatives*, *3*(1), 9–42.

Paquette, G. (1996b), La modélisation par objets typés - une méthode de représentation pour les systèmes d'apprentissage et d'aide à la tâche, *Sciences et Techniques Educatives*, *3*(1), 9–42.

Park, Y., R. J. Byrd, et B. K. Boguraev (2002), Automatic glossary extraction : Beyond terminology identification, in *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pp. 1–7, Association for Computational Linguistics, Stroudsburg, PA, USA.

Perinet, A., et T. Hamon (2013), Hybrid acquisition of semantic relations based on context normalization in distributional analysis, in *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence (TIA 2013)*, pp. 113–122.

Poli, R., M. Healy, et A. Kameas (Eds.) (2010), *Cyc*, pp. 259–278, Springer Netherlands, Dordrecht.

Popper, K. (1998), *La connaissance objective : Une approche évolutionniste*, Champs, Flammarion, Paris.

Popper, K. R. (1979), *Objective Knowledge : An Evolutionary Approach*, revised ed., Clarendon Press, Oxford.

Poveda, M., M. C. Suarez-Figueroa, et A. Gomez-Perez (2009), Ontology analysis based on ontology design patterns, in *WOP 2009 Workshop on Ontology Patterns at the 8th International Semantic Web Conference (ISWC 2009). Proceedings of the WOP 2009.*, edited by 8th International Semantic Web Conference (ISWC 2009), WOP 2009 – Workshop on Ontology Patterns at the 8th International Semantic Web Conference (ISWC 2009).

Punuru, J., et J. Chen (2012), Learning non-taxonomical semantic relations from domain texts, *Journal of Intelligent Information Systems*, *38*(1), 191–207.

Rich, E. (1983), *Artificial Intelligence*, McGraw-Hill, Inc., New York, NY, USA.

Rizoiu, M.-A., et J. Velcin (2011), Topic extraction for ontology learning, in *Ontology Learning and Knowledge Discovery Using the Web : Challenges and Recent Advances*, edited by W. Wong, W. Liu, et M. Bennamoun, chap. 3, pp. 38–61, Hershey, PA : Information Science Reference, doi :10.4018/978-1-60960-625-1.ch003.

Roche, C. (2005), Terminologie et ontologie, *Langages*, *1*(157), 48–62.

Rosario, B., et M. Hearst (2001), Classifying the semantic relations in noun compounds, in *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

Rouane, M. H., M. Huchard, A. Napoli, et P. Valtchev (2007), *A Proposal for Combining Formal Concept Analysis and Description Logics for Mining Relational Data*, pp. 51–65, Springer Berlin Heidelberg, Berlin, Heidelberg.

Rowley, J. (2007), The wisdom hierarchy : representations of the dikw hierarchy, *Information Science*, *33*(2), 163–180.

Schank, R. C., et C. J. Rieger (1985), Inference and the computer understanding of natural language, in *Readings in Knowledge Representation*, edited by R. J. Brachman et H. J. Levesque, pp. 119–139, Kaufmann, Los Altos, CA.

Schmid, H. (2000), Unsupervised learning of period disambiguation for tokenisation, *Tech. rep.*, IMS, University of Stuttgart.

Schutz, A., et P. Buitelaar (2005), Relext : A tool for relation extraction from text in ontology extension, in *Proceedings of the 4th International Semantic WEB Conference (ISWC)*, p. 5.

Sclano, F., et P. Velardi (2007), Termextractor : a web application to learn the shared terminology of emergent web communities, in *Enterprise Interoperability II - New Challenges and Industrial Approaches, Proceedings of the 3th International Conference on Interoperability for Enterprise Software and Applications, IESA 2007, March 27-30, 2007, Funchal, Madeira Island, Portugal*, pp. 287–290.

Serra, I., et R. Girardi (2011), A process for extracting non-taxonomic relationships of ontologies from text, *Intelligent Information Management*, *3*(4), 119–124.

Serra, I., R. Girardi, et P. Novais (2013), PARNT : A statistic based approach to extract non-taxonomic relationships of ontologies from text, in *Tenth International Conference on Information Technology : New Generations, ITNG 2013, 15-17 April, 2013, Las Vegas, Nevada, USA*, pp. 561–566.

Shabanov-Kushnarenko, Y. (1984), *Theory ofIntelligence : Mathematical Tools. Теория интеллекта.Математические средства*, Вища школа, Харьков.

Shannon, C. E. (1948), A Mathematical Theory of Communication, *The Bell System Technical Journal*, *27*(3), 379–423.

Sharonova, N. (2010), *Математические модели знаний и их реализация с помощью алгебропредикатных структур*, 304 pp., Дмитренко Л.Р.

Shortliffe, E. H. (1976), *Computer-based medical consultations : MYCIN*, Elsevier computer science library, Elsevier, New York, Oxford, Amsterdam.

Shustova S., S. E. (2015), Verb valency theory in the russian and western scientific paradigms Теория глагольной валентности в отечественной и западной научной парадигмах, *Вестник Ленинградского государственного университета им. А.С. Пушкина*.

Silva, J. F. D., G. P. Lopes, Q. D. Torre, et M. D. Caparica (1999), A local maxima method and a fair dispersion normalization for extracting multiword units, in *In Proceedings of the 6th Meeting on the Mathematics of Language*, pp. 369–381.

Simperl, E., C. Tempich, et D. Vrandečić (2008), A methodology for ontology learning, in *Proceedings of the 2008 Conference on Ontology Learning and Population : Bridging the Gap Between Text and Knowledge*, pp. 225–249, IOS Press, Amsterdam, The Netherlands, The Netherlands.

Simperl, E. P. B., M. Mochol, et T. Burger (2010), Achieving maturity : the state of practice in ontology engineering in 2009, *IJCSA*, *7*(1), 45–65.

Smadja, F., K. R. McKeown, et V. Hatzivassiloglou (1996), Translating collocations for bilingual lexicons : A statistical approach, *Comput. Linguist.*, *22*(1), 1–38.

Sokirko, A. (2004), Morphology modules at www.aot.ru Морфологические модули на сайте www.aot.ru, in *Компьютерная лингвистика и интеллектуальные технологии : Труды международной конференции Диалог 2004*, Наука Москва.

Sowa, J. F. (Ed.) (1991), *Principles of Semantic Networks : Explorations in the Representation of Knowledge*, The Morgan Kaufmann Series in Representation and Reasoning, Morgan Kaufmann.

Specia, L., et E. Motta (2006), A hybrid approach for extracting semantic relations from texts, in *Proceedings of the 2nd Workshop on Ontology Learning and Population : Bridging the Gap between Text and Knowledge*, pp. 57–64, Association for Computational Linguistics, Sydney, Australia.

Stevenson, M. (2004), An unsupervised WordNet-based algorithm for relation extraction, in *Proceedings of the Fourth International Conference on Language Resources and Evaluation workshop "Beyond Named Entity : Semantic Labelling for NLP tasks", Lisbon, Portugal*.

Tannier, X. (2014), Traitement des événements et ciblage d'information, Habilitation à Diriger des Recherches (HDR).

Tesnière, L. (1959), *Éléments de Syntaxe Structurale*, Librairie C. Klincksieck.

Uschold, M. (2008a), Ontology-driven information systems : Past, present and future, in *Proceeding of the 2008 conference on Formal Ontology in Information Systems : Proceedings of the Fifth International Conference (FOIS 2008)*, pp. 3–18, IOS Press, Amsterdam, The Netherlands, The Netherlands.

Uschold, M. (2008b), Ontology-driven information systems : Past, present and future, in *Proceedings of the 2008 Conference on Formal Ontology in Information Systems : Proceedings of the Fifth International Conference (FOIS 2008)*, pp. 3–18, IOS Press, Amsterdam, The Netherlands, The Netherlands.

Uschold, M., et M. Gruninger (1996), Ontologies : Principles, methods and applications, *Knowledge Engineering Review*, *11*, 93–136.

Van Renssen, A. (2005), *Gellish – A Generic Extensible Ontological Language : Design and Application of a Universal Data Structure*, IOS Press, Incorporated.

Velardi, P., S. Faralli, et R. Navigli (2013), Ontolearn reloaded : A graph-based algorithm for taxonomy induction, *Computational Linguistics*, *39*(3), 665–707.

Vernadski, V. I. (2012), *(Вернадский В.И.) Биосфера и ноосфера*, 576 pp., Айрис-Пресс.

Villaverde, J., A. Persson, D. Godoy, et A. Amandi (2009), Supporting the discovery and labeling of non-taxonomic relationships in ontology learning, *Expert Syst. Appl.*, *36*(7), 10,288–10,294.

Wagner, W., H. Schmid, et S. S. Im Walde (2009), Verb sense disambiguation using a predicate-argument-clustering model, in *Proceedings of the CogSci Workshop on Distributional Semantics beyond Concrete Concepts*, p. 23–28.

Wilson, J., J. Miller, A. Konradi, et F. A. Cucinotta (1997), Shielding strategies for human space exploration, *Tech. rep.*, NASA, NASA Langley Research Center Hampton, US.

Winograd, T. (1972), *Understanding Natural Language*, Academic Press, Inc., Orlando, FL, USA.

Wittgenstein, L. (2005), *Людвиг Витгенштейн. Избранные работы (перевод, подготовка текста, комментарии и приложения)*, 437 pp., Издательский дом "Территория будущего" Москва.

Wong, W., W. Liu, et M. Bennamoun (2007), Determining termhood for learning domain ontologies using domain prevalence and tendency, in *Proceedings of the Sixth Australasian Conference on Data Mining and Analytics - Volume 70*, AusDM '07, pp. 47–54, Australian Computer Society, Inc., Darlinghurst, Australia, Australia.

Wong, W., W. Liu, et M. Bennamoun (2012), Ontology learning from text : A look back and into the future, *ACM Comput. Surv.*, *44*(4), 20 :1–20 :36.

Woods, W. A., et J. G. Schmolze (1992), The kl-one family, *Computers & Mathematics with Applications*, *23*(2), 133–177.

Wüster, E. (1985), *Introduction à la théorie générale de la terminologie et à la lexicographie terminologique*, E. Brent.

Wynne, M. (Ed.) (2005), *Developing Linguistic Corpora : a Guide to Good Practice*, Oxford : Oxbow Books.

Yourdon, E. (1989), *Modern Structured Analysis*, Yourdon Press, Upper Saddle River, NJ, USA.

Zaliznyak, A. A. (1977), *Grammaticheskij slovar' russkogo jazyka Грамматический словарь русского языка. Словоизменение*, 880 pp., Русский язык.

Zhang, W., T. Yoshida, X. Tang, et T. B. Ho (2009), ., *Expert Syst. Appl.*, *8*, 10,919–10,930.

Zhou, L. (2007), Ontology learning : state of the art and open issues, *Information Technology and Management*, *8*(3), 241–252.

Zins, C. (2007), Conceptual approaches for defining data, information, and knowledge, *JASIST*, *58*(4), 479–493.

Zolotova, G. (2011), *Синтаксический словарь : Репертуар элементарных единиц русского синтаксиса*, 356 pp., Едиториал УРСС.