Construction d'une ressource termino-ontologique multilingue pour le domaine de la cuisine et de la nutrition

Nadezda Bebeshina-Clairet

UNIVERSITE PARIS 13

"U.F.R de Santé, Médecine et Biologie Humaine"

N° attribué par la bibliothèque
|_ _|_ _|_ _|_ _|_ _|_ _|_ _|_ _|_ _|_ _|

THESE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PARIS 13

Discipline : Informatique

présentée et soutenue publiquement le 29 janvier 2019

par

Nadia Bebeshina-Clairet

Titre : Construction d'une ressource termino-ontologique multilingue pour les domaines de la cuisine et de la nutrition

---

Directeur de thèse : Sylvie Després
Co-directeur de thèse : Mathieu Lafourcade

JURY

Dr. Hervé Blanchon, MCF HDR, Université Grenoble Alpes, *rapporteur*
Dr. Christophe Roche, PU, Université Savoie Mont-Blanc, *rapporteur*
Dr. Andon Tchetchmedjiev, MCF, IMT Mines Alès, *examinateur*
Dr. Thierry Charnois, PU, Université Paris 13, *examinateur*
Dr. Sylvie Després, PU, Université Paris 13, *directeur de thèse*
Dr. Mathieu Lafourcade, MCF HDR, Université de Montpellier, *co-directeur de thèse*
Dr. François Brown de Colstoun, PDG, Lingua et Machina, *invité*

Résumé

Le présent travail de recherche s'attache à montrer comment un réseau lexico-sémantique multilingue pour un domaine de spécialité peut être construit dans un contexte industriel et comment il peut être utilisé pour accompagner la modélisation ontologique. La construction ontologique est un processus coûteux souvent réalisé manuellement en suivant une démarche descendante. Parallèlement, les ressources de connaissance non ontologiques sont de plus en plus matures et disponibles. Parmi ces dernières, les réseaux lexico-sémantiques peuvent être exploités pour modéliser les connaissances d'un domaine de spécialité donné de façon explicite du point de vue sémantique, y compris en ce qui concerne les raffinements de sens. Lorsque l'on dispose d'un réseau lexico-sémantique de spécialité de ce type, une ressource médiatrice peut être extraite à partir de celui-ci et vue comme une projection d'un modèle de structuration des connaissances donné (en particulier, modèle d'ontologie) sur le réseau. Une telle ressource médiatrice peut servir de support à la construction ontologique par les experts humains dans le cadre d'acquisition des connaissances du domaine et modélisation des structures ontologiques. Ce travail exploite les techniques d'intelligence artificielle avec un accent mis sur l'apprentissage permanent (la ressource décrite dans le présent mémoire est constamment améliorée par les processus qui l'utilisent) et des approches à la construction ontologique à partir des textes qui passent par une étape de structuration de connaissance intermédiaire sous forme d'une termino-ontologie.

Mots clés

Traitement automatique des langues (TAL), intelligence artificielle (IA), sémantique des langues naturelles, modélisation, ressource lexicale, ontologie, termino-ontologie, conceptualisation.

Abstract

The present research work is intended to show how a domain specific lexical semantic network can be built in an industrial context and how it can be used in order to assist ontology modeling. Ontology building is a costly process often done manually. In the same time, non ontological knowledge resources are available and mature. Among those, lexical semantic networks can be exploited in order to model domain specific knowledge in a semantically explicit way with sense refinements and concern multiple languages. Given such domain specific multilingual lexical semantic network, a mediatory terminological and ontological resource can be extracted from this network and viewed as a projection of a given knowledge model (i.e. ontology) onto the domain specific lexical semantic network. Such mediatory resource can support the top down ontology building process as it contains ontological structures that can be used by the human experts during knowledge acquisition and modeling. This work stems from the artificial intelligence approach focused on permanent learning (the network is continuously improved by the processes it is used by) and from text based ontology building approaches.

Keywords

Natural Language Processing (NLP), Artificial Intelligence (AI), natural language semantics, modeling, lexical resource, ontology, terminological ontology, conceptualization.

Résumé étendu

La construction et le maintien des ressources termino-ontologiques interopérables est une tâche longue et coûteuse car elle requiert un effort de conceptualisation difficile à automatiser.

Dans un contexte industriel, l'enrichissement des ressources de connaissance existantes et, en particulier, la construction des ontologies s'oriente vers la mise en œuvre des ressources multilingues de spécialité notamment parce que des ressources terminologiques et lexicales sont de plus en plus abondantes pour les différents domaines tels que la médecine, la biologie, l'agriculture, l'alimentation ainsi que pour les différentes langues. Il s'agit des ressources structurées (dictionnaires, lexiques, glossaires, réseaux lexico-sémantiques) et non structurées (corpus de textes bruts).

Cependant, les ressources structurées disponibles sont essentiellement des ressources langagières qui modélisent la connaissance sur le monde telle qu'elle est exprimée à travers le langage humain et contiennent également la connaissance sur les langues elles mêmes. A ce titre, la valeur ajoutée des ressources structurées disponibles est dans la manière dont elles modélisent la polysémie propre au langage naturel ainsi que les liens entre les différentes langues lorsqu'il s'agit des ressources multilingues. Les utiliser pour créer ou améliorer des ressources termino-ontologiques s'avère difficile car une ontologie est portée par une sémantique formelle qui ne comporte pas de polysémie. Une médiation est alors nécessaire pour exploiter pleinement les ressources langagières dans le cadre des processus de conceptualisation. Lors de la construction ontologique descendante qui part des concepts ontologiques clés pour ensuite définir, éventuellement désigner et caractériser les concepts plus spécifiques, cette médiation serait assurée par les experts humains. Or, ces derniers sont locuteurs d'une ou plusieurs langues naturelles ce qui influence la conceptualisation qu'ils délivrent. Dans les domaines comme la cuisine et la nutrition cet ancrage culturel se manifeste nettement : le mot anglais *stew* désigne en français à la fois *ragoût* et *pot au feu* ce qui équivaut à deux processus distincts de préparation culturellement ancrés (les aliments sont directement bouillis ou revenus dans de la matière grasse puis bouillis). Une ressource termino-ontologique doit rendre compte de ce contraste. Parallèlement, un socle commun de connaissance est aisément ressenti, une ontologie *coeur* interlingue pourrait alors émerger a cours des processus de modé-

lisation translingues.

Dans le présent travail, nous proposons une approche qui utilise un réseau lexico-sémantique multilingue (RLSM) pour enrichir automatiquement une ontologie de référence fournie à l'entrée du processus. L'architecture du RLSM est issue des approches traditionnelles d'Intelligence Artificielle et s'inspire de la structure de RezoJDM, Lafourcade [2011]. Il s'agit d'un graphe orienté, typé, valué construit pour chaque langue. Les nœuds de ce graphe modélisent les objets lexicaux (termes, raffinements des termes) tandis que les relations représentent les relations sémantiques et lexicales entre ces objets. Le lien interlingue se fait via le pivot interlingue. Les données contenues dans ce réseau sont issues des ressources variées (ressources de connaissance existantes partiellement intégrées, corpus). L'ontologie de référence est immergée dans le RLSM (exprimée en termes de formalisme RLSM), puis enrichie grâce à un ensemble de processus d'inférence endogène et pourra être "extraite" dans une autre langue que sa langue de départ. Le pivot interlingue permet également de capter la connaissance partagée dans le cadre d'une modélisation ontologique *coeur*. Cette approche peut être généralisée pour la construction ontologique semi-automatique à partir d'une ébauche d'ontologie.

Remerciements

L'aventure de la thèse s'achève et je tiens à adresser mes remerciements à celles et ceux qui l'ont rendue possible. Je remercie mes directeurs de thèse pour avoir accepté d'encadrer ce travail, pour m'avoir guidée, questionnée et encouragée.

Merci à François Brown de Colstoun pour m'avoir ouvert les portes de son entreprise et pour avoir été le patron à la fois exigeant et bienveillant.

Je remercie Hervé Blanchon et Christophe Roche pour avoir accepté de rapporter sur cette thèse ainsi que pour leurs remarques pertinentes.

Je remercie Guillaume Clairet et Pascal Vaillant pour avoir cru en ce projet et participé à le lancer. Je ne l'oublie pas.

J'adresse mes remerciements chaleureux à tous les membres du laboratoire LIMICS dont j'ai pu apprécier les qualités professionnelles et humaines, malgré la distance géographique et la brièveté de nos rencontres.

Merci à tous les membres de l'équipe Texte et, plus particulièrement, à mes compagnons de thèse Jimmy, Kévin, Davide, Lionel, Mehdi qui ont toujours été là pour moi et qui, au gré des joies et difficultés vécues ensemble, sont devenus des amis.

Merci enfin à ma famille pour m'avoir tant soutenue.

Table des matières

Acronymes

ACO *Aide à la construction d'ontologie*

ASIC *Analyse sémantique des instructions de cuisine*

GWAP *Games With a Purpose*

ISO *International Standard Organization*

RLS *Réseau lexico-sémantique*

RLSM$_{\text{PI}}$ *Réseau lexico-sémantique multilingue avec pivot interlingue*

RWN *Russian WordNet*

SMA *Système multi-agent*

SPG *Score de Poids Global*

TMM *Terme Multi-Mot*

Introduction

Dans le présent mémoire, nous allons nous attacher à la représentation et à l'exploitation des connaissances pour les domaines de la cuisine et de la nutrition. Nous nous focaliserons sur une structure de ressource de connaissance sous forme de réseau lexico-sémantique multilingue avec pivot interlingue (RLSM_{PI}) et nous nous intéresserons à la façon dont ce type de ressource de connaissance non ontologique peut permettre de générer des ressources informelles de médiation destinées à enrichir et outiller le processus de construction ontologique par les experts. Une ressource de médiation est entendue comme une projection d'un modèle de structuration de connaissance donné sur le RLSM_{PI}. Autrement dit, pour une ressource ontologique, il s'agit d'exprimer la connaissance disponible au sein du RLSM_{PI} en termes de hiérarchie des classes et des propriétés de cette ontologie.

Questionnements

Depuis leur apparition, les ressources termino-ontologiques conçues pour un domaine de spécialité font objet d'une construction menée par les experts. Il s'agit souvent d'une construction qui suit un mouvement descendant car elle s'appuie sur la définition des concepts clés, puis détaille ces concepts. Plus précisément, il s'agit d'identifier manuellement les concepts et leur propriétés, les classer sous forme d'arbre et décrire les propriétés, identifier les instances des différentes classes, décrire les instances. Par conséquent, malgré l'apparition des méthodes axées sur l'interopérabilité des ressources (réutilisation, mise en réseau etc.), cette construction demeure longue et coûteuse. En particulier, pour la mener à bien dans un contexte multilingue, il est indispensable de recruter des experts du domaine maîtrisant plusieurs langues car les différences de conceptualisation potentielles doivent être prises en compte. Par ailleurs, supposée être indépendante d'une langue donnée, la conceptualisation dans le cadre de construction d'ontologie reste influencée par la langue de ses concepteurs humains. Ce biais est perceptible lorsque l'on s'intéresse aux ontologies qui modélisent le domaine de l'alimentation humaine.

Question 1

Comment réduire l'effort humain nécessaire à la conceptualisation ?

Parallèlement à cette tendance, des ressources textuelles (corpus) et des ressources de connaissance structurées sous différentes formes sont devenues relativement abondantes notamment lorsqu'il s'agit de la connaissance générale : WordNet (Fellbaum [1998]), FrameNet (Ruppenhofer et al. [2006]), ConceptNet (Speer and Havasi [2012]), RezoJDM (Lafourcade [2007]). C'est également le cas de certains domaines de spécialité : biologie, médecine. En ce qui concerne les ressources lexicales, un phénomène de concentration autour de quelques modèles saillants s'est produit. Ainsi, pour les ressources lexico-sémantiques monolingues, il est possible de distinguer les architectures basées sur les *synsets*, sur les *raffinements* et les fonctions lexicales, sur les *frames*.

Question 2

Comment bénéficier des ressources structurées non ontologiques de connaissance générale dans le cadre de la construction d'ontologie ?

L'état de l'art sur l'évolution des ressources ontologiques existantes vers le multilinguisme laisse apparaître le fait que l'alignement et l'enrichissement des ontologies afin de les rendre multilingues se base sur le traitement automatique des étiquettes de ses classes. Or, les étiquettes d'ontologie font partie de la dénotation formelle et, à ce titre, n'appartiennent pas à une langue naturelle. En particulier, à l'intérieur d'une ressource ontologique, les étiquettes ne sont pas polysémiques. Dans les cas où le mot du lexique correspondant à une étiquette d'ontologie donnée serait polysémique, la projection du modèle d'ontologie la considère non ambiguë.

Question 3

Comment obtenir des alignements entre les différentes langues par émergence (sans passer par la traduction des étiquettes) et refléter les différences de conceptualisation propres à des contextes socioculturels différents ?

Objectifs

Compte tenu de ces problématiques générales, nous nous sommes fixés un ensemble d'objectifs pour tenter d'y répondre :

— **construire un réseau lexico-sémantique multilingue avec pivot interlingue** (RLSM_{PI}) par des méthodes non contributives (sans recours à la contribution directe ni aux *GWAP* (*Games with a Purpose*, jeux contributifs) et, par cette construction, expliciter les différentes approches à cette construction qui peuvent être mises en œuvre dans un contexte industriel ;

— **proposer un protocole d'évaluation** quantitative et qualitative (par la tâche) du RLSM_{PI} ;

— **exploiter RLSM_{PI}** pour enrichir une ontologie de référence, proposer des méthodes d'assistance à la construction d'ontologie par les experts et outiller la modélisation à partir d'une ébauche d'ontologie.

Idées guides

Le présent travail se situe dans le cadre des méthodes d'intelligence artificielle centrées sur le calcul pour la construction de ressources de connaissance. Il ne s'agit pas d'un travail sur la constitution d'une terminologie de spécialité. Les modèles de représentation sous forme de terminologie (dont terminologie multilingue), termino-ontologie ou sous forme d'un ensemble de structures ontologiques conformes à un modèle de référence constituent des projections de ces modèles sur le contenu du RLSM_{PI}.

La ressource en elle-même est structurée et améliorée grâce aux techniques issues d'intelligence artificielle symbolique. Il s'agit notamment des techniques d'inférence de relations, d'explicitation des contenus (sens des mots, méta-informations, etc.) sous forme de termes (nœuds) et de relations (arcs) et d'organisation des différents processus en tant que processus indépendants.

Compte tenu de cette orientation scientifique, les idées principales qui ont guidé l'approche à la construction du RLSM_{PI} et son exploitation sont les suivantes :

— *complétude de la langue naturelle*, autrement dit, toute langue naturelle peut tout exprimer, seuls divergent les moyens lexicaux et morpho-syntaxiques qui sont employés ;

— *amélioration continue du réseau lexico-sémantique* grâce aux processus qui l'utilisent ;

— *caractère englobant des structures sémantiques* car elles peuvent révéler les structures conceptuelles (taxonomie) ou terminologiques (taxonomie informelle, terme et ses variantes).

Contexte de réalisation

Contexte scientifique. Sur le plan scientifique, ce travail s'est déroulé en coopération entre deux laboratoires de recherche

— Laboratoire d'informatique médicale et d'ingénierie des connaissances en e-Santé (LIMICS). Il s'agit d'une unité de recherche interdisciplinaire en informatique et en informatique médicale.

— Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), équipe TEXTE. L'équipe TEXTE développe des modèles et des outils pour analyse automatique, syntaxique et sémantique, du langage naturel ainsi que pour la constitution des ressources de connaissance.

Contexte industriel. Le projet de recherche a fait l'objet de la convention CIFRE 2016/0433. Les travaux se sont déroulés en collaboration avec l'entreprise Lingua et Machina dont l'activité est centrée autour des solutions de traduction spécialisée et de gestion des contenus multilingues. Nos travaux se sont situés dans le cadre d'amélioration de l'outil Libellex développé pour la structuration et la gestion des terminologies multilingues.

Organisation du manuscrit

Le présent manuscrit est divisé en cinq chapitres qui répondent aux problématiques de construction d'un réseau lexico-sémantique multilingue et d'exploitation de cette ressource pour la construction des structures ontologiques et les différentes sorties possibles de cette ressource.

Le premier chapitre introduit, dans un premier temps, les problématiques d'interopérabilité des ressources de connaissance et, en particulier, des ressources langagières. Ensuite, il définit les différentes méthodes que nous avons pu distinguer sur la base de l'état de l'art en ce qui concerne la construction et la mise à jour des ressources de connaissance. Enfin, il décrit les principales ressources de connaissance existantes et pose les problèmes auxquels nous tentons de répondre et qui correspondent à l'utilisation des ressources langagières et des méthodes issues du domaine de l'intelligence artificielle pour outiller la construction termino-ontologique.

Le deuxième chapitre est centré sur les différents processus de construction d'une ressource de connaissance compte tenu des différentes méthodes qui peuvent être employées et décrit les expérimentations quant à l'amorçage et au peuplement de la ressource. Une attention particulière est portée au processus d'inférence translingue des relations sémantiques et des raffinements glosés.

Le troisième chapitre se focalise sur l'identification des structures ontologiques à l'intérieur du RLSM_{PI} qui s'appuie sur l'immersion d'une ontologie de référence.

Il met en évidence comment ces structures peuvent être enrichies au sein du RLSM_{PI} afin de permettre une sortie sous forme de structures ontologiques.

Le quatrième chapitre présente les différentes formes d'évaluation quantitative et qualitative du RLSM_{PI}. Sur le plan quantitatif, il met en évidence les modalités d'évaluation considérant la distribution des termes et relations du RLSM_{PI} en fonction de leur poids, la couverture du pivot interlingue et la gestion de la polysémie. Sur le plan qualitatif, un protocole d'évaluation par la tâche est proposé via, notamment, des tâches comme l'analyse sémantique (générale et focalisée), validation translingue des contributions en attente (pour les ressources qui intègrent la peuplonomie dans leur modèle de construction).

Le cinquième chapitre présente les implémentations qui mettent en perspective les outils basés sur le RLSM_{PI}. Il est question, d'une part, des outils d'assistance à la construction d'ontologie destinés aux ontologues et aux experts et, d'autre part, des méthodes basées sur l'inférence des relations pour outiller la construction des structures ontologiques destinées à alimenter une ressource informelle adossée à une ontologie existante ou la modélisation à partir d'une ébauche d'ontologie à alimenter.

Compte tenu de la variété des sujets traités, chaque partie comporte son propre lexique et fait référence à son propre état de l'art.

Chapitre 1

Contexte et problématiques

Le présent chapitre aborde les problémes liées à l'interopérabilité des ressources, à la structure et à la construction des ressources multilingues et aux applications liées à l'analyse des recettes de cuisine. Il introduit également les méthodes génériques pouvant être utilisées pour la construction des ressources de connaissance et détaille les principales ressources de connaissance existantes et leur particularités.

Termes et notations utilisés dans le chapitre 1

ressource : ensemble de moyens nécessaires pour accomplir un ensemble de tâches.

ressource de connaissance : ressource qui regroupe et/ou structure des connaissances.

ressource langagière : ressource de connaissance qui modélise la connaissance sur le monde telle qu'elle est exprimée à travers le langage humain. À ce titre, elle inclue à la fois la connaissance d'ordre ontologique et sémantique et la connaissance d'ordre purement lexical et spécifique à une langue donnée. Toute ressource de connaissance n'est pas forcément une ressource langagière.

terme : *(Terminologie [a])* désignation verbale d'un concept général dans un domaine spécifique. *(RLS)* item lexical (vocable, expression polylexicale).

concept : *(Terminologie [b])* unité de connaissance créée par une combinaison unique de caractères.

donnée : *(Terminologie)* informations représentées sous une forme conventionnelle convenant à la communication, à l'interprétation, au stockage et au traitement.

NB : La norme ISO/DIS 1087 concernant les travaux terminologiques est actuellement en cours d'élaboration. Nous utilisons les définitions issues de l'ancienne norme

ISO-1087 1 et 2 (aujourd'hui annulée).

a. Standard ISO 1087-2.
b. Standard ISO 1087-1.

1.1 Contexte général

Les applications de TALN à l'analyse des textes de spécialité nécessitent constamment de nouvelles méthodes d'analyse sémantique et de construction termino-ontologique.
L'analyse sémantique est nécessaire pour l'extraction des termes et des structures sémantiques en vue d'annotation sémantique, de traduction automatique, de résumé automatique des textes mais aussi, et c'est à ce titre que nous l'évoquons, l'acquisition permanente des connaissances en vue d'amélioration d'une ressource de connaissance. La désambiguïsation du sens, l'identification des relations sémantiques et des rôles prédicatifs (ex. agent, patient, instrument, lieu etc), la définition de la structure *qualia* d'un terme, l'identification des traces de concept dans les textes (explicitation des concepts), la détection des événements peuvent être cités parmi les tâches d'analyse sémantique généraliste. C'est ainsi, de façon large, qu'elle sera entendue dans le présent manuscrit.
La construction termino-ontologique est traditionnellement considérée comme une tâche longue et très coûteuse car elle est souvent conduite manuellement par une groupe d'experts dans une démarche descendante, en partant des concepts les plus génériques d'un domaine de spécialité et en spécifiant ces concepts. Le consensus entre les experts est à la base de cette démarche. Ainsi, le maintien et l'augmentation de même que le passage d'une langue à l'autre comme cas particulier de l'augmentation constituent des barrières importantes quant à l'interopérabilité des ressources termino-ontologiques de spécialité. Simultanément, dans un domaine de spécialité tel que la nutrition ou la gastronomie, un socle commun de connaissances partagées par plusieurs traditions gastronomiques et reflété sur le plan linguistique peut être aisément pressenti. Une méthode d'aide à la construction termino-ontologique est alors souhaitée afin de faciliter cette démarche et la gestion de granularité entre différentes langues dans le cadre multilingue de spécialité.
La construction d'une ressource multilingue pouvant servir de support à la construction termino-ontologique implique l'intégration des données existantes. Incontestablement, les textes en langue naturelle sont une source très abondante de données de spécialité. De même, sachant que leur intérêt dépend du domaine de spécialité, doivent être considérées d'une part les ressources lexicales et, d'autre part, les ressources terminologiques et ontologiques préexistantes. De son côté, l'architecture de la ressource multilingue doit permettre l'augmentation permanente à la fois en circuit ouvert à savoir par intégration des ressources et par analyse sémantique et en circuit fermé soit par des méthodes endogènes de peuplement telles que l'inférence (création de nouveaux éléments à partir des

éléments déjà présents dans la ressource) ou d'intégrer la ressource .

1.2 Interopérabilité des ressources

La conception des ressources interopérables est le fil rouge de l'ingénierie de connaissances moderne. L'interopérabilité des ressources implique que l'on puisse connecter entre elles des ressources possiblement issues de formalismes différents en se basant sur un ensemble de principes formels de conversion. Lorsqu'il s'agit des ressources langagières, l'interopérabilité sous-entend de façon implicite l'existence d'une sémantique du monde indépendante d'une langue donnée et de toute structure lexicale, grammaticale ou syntaxique prédéfinie et que cette sémantique puisse être encodée de manière consensuelle.

Exemple 1.1

$< wikicat\_Norwegian\_ballet\_dancers >$

`rdfs:subClassOf` $< wordnet\_ballet\_dancer\_109834699 >$

`rdfs:subClassOf` $< wordnet\_dancer\_109989502 >$ [a].

a. exemple cité d'après Rebele et al. [2016]

Formaliser les relations sémantiques de manière consensuelle (Princenton WordNet Fellbaum [1998]) permet de concevoir les ressources interopérables monolingues et multilingues qui peuvent suivre et étendre le même formalisme (*Russian WordNet*, Loukachevitch [2016], *Open Multilingual WordNet*, Bond and Foster [2013]) et être liées à des ressources qui suivent un modèle différent. Ainsi, dans YAGO (Suchanek et al. [2007]) les feuilles de l'arbre des catégories de Wikipedia sont liés à des ensembles de synonymes de PWN (exemple 1.1).

Définition 1.1

Interopérabilité est la capacité que plusieurs systèmes ou ressources puissent communiquer et opérer ensemble sans ambiguïté, sans conflit de système ou de format de contenu.

De façon intuitive, la distinction entre l'interopérabilité de format, de modèle et de contenu peut être aisément perçue. L'interprétation de cette problématique faite dans le cadre des expérimentations concrètes relève à la fois de la nature de la ressource conçue ou rendue interopérable et des objectifs d'ordre applicatif visés.

Premièrement, s'impose l'*interopérabilité de format*. Ainsi, de nombreux travaux ont été et continuent à être menés sur différents aspects des formats basés sur le format XML et utilisés, par exemple, dans le cadre de construction terminologique.

Exemple 1.2

Type de ressource	Exemples de formats
corpus annoté	Extremely Annotational RDF Markup (EARMARK) [a] NLP Interchange Format (NIF) [b] Lexical Markup Framework (LMF) [c]
terminologie	SKOS (Simple Knowledge Organization System) [d] TBX (Term-Based Exchange) [e] TMX [f]
ontologie	OntoLex (*lexicon model for ontologies*) [g]

a. `http://www.essepuntato.it/lode/owlapi/http://www.essepuntato.it/2008/12/earmark`
b. `http://persistence.uni-leipzig.org/nlp2rdf/`
c. http ://www.lexicalmarkupframework.org/
d. `https://www.w3.org/2004/02/skos/`
e. `http://www.tbxinfo.net/`
f. `http://xml.coverpages.org/tmxSpec971212.html`
g. `https://www.w3.org/community/ontolex/wiki/Final_Model_Specification`

Le développement des standards et des formats ouverts tels que `ontolex` concerne les ressources lexicales et l'ancrage linguistique des ressources ontologiques. Ainsi, `ontolex` étend le format `owl` afin de permettre de détailler la lexicalisation des concepts d'ontologie. De façon similaire, `SKOS` est le format de référence pour les terminologies, les taxonomies ainsi que les autres ressources portées par un schéma informel. Les travaux sur l'interopérabilité de format concernent également les ressources langagières semi-structurées telles que corpus annotés car l'interopérabilité des annotations relève en grande partie de celle du format utilisé pour les encoder. Elle implique aussi l'interopérabilité conceptuelle (vocabulaire partagé). Les formats de référence pour l'annotation des corpus ont été proposés dans le cadre du Modèle Ouvert d'Annotation des Données *(Open Annotation Data Model)* [1], *NLP Interchange Format* (NIF), *Extremely Annotational RDF Markup* (EARMARK).

Deuxièmement, il est possible de considérer l'*interopérabilité de modèle* soit l'*interopérabilité représentationnelle*, d'après la définition proposée par Witt et al. [2009]. Les modèles de deux ressources de connaissance peuvent être interopérables

— "par conception" (deux modèles définis de façon indépendante livrent une représentation de connaissance similaire) ;

1. `http://www.openannotation.org/spec/core/`

— "par référence" (deux modèles A et B sont similaires car B intègre A partiellement ou totalement) ;
— "par interconnexion" (deux modèles dont les structures peuvent différer mais les éléments sont connectées via les faisceaux de (hyper)liens).

L'interopérabilité par référence à un modèle pré-existant est de plus en plus fréquente notamment par nécessité de limiter les coûts de construction et de mise à jour de ces ressources de connaissance et, en particulier, des ressources langagières. La figure 1.1 donne un aperçu partiel de l'interopérabilité par intégration ou par référence des principales ressources de connaissance utilisés actuellement. Les ressources comme YAGO (Suchanek et al. [2007]) sont absents de ce diagramme car leur conception et leur mise à jour se basent sur l'utilisation des données semi-structurées (Wikipedia) que l'on peut trouver sur le web, leur modèle est conditionné par les données largement disponibles. À titre d'exemple, sur le plan représentationnel, toute ressource basée sur les associations lexicales est interopérable avec la ressource RezoJDM, Lafourcade [2007]. Cette ressource est interopérable avec BabelNet (et al. [2012]), DBNary (Sérasset [2014]) etc. par interconnexion.

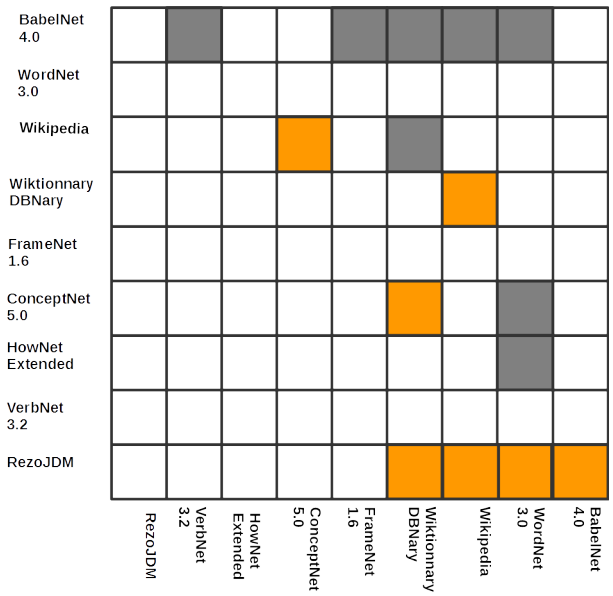


FIGURE 1.1 – Diagramme ressources et interopérabilité. En gris - interopérabilité par référence, en orange - interconnexion (Web Sémantique, liens spécifiques).

Troisièmement, vient l'interopérabilité de contenu. À ce niveau surgissent de nombreuses problématiques car le terme "contenu" englobe à la fois la sémantique des composants (le contenu formel) et la sémantique des données contenues dans telle ou telle ressource. À ce titre, les ressources langagières, notamment lexico-sémantiques bien que présentant quelques similarités de modèle ou de format avec les ontologies, ne sont pas nécessairement interopérables en termes de sémantique des composants car les relations de base telles que la relation d'hyperonymie peuvent ne pas vérifier les mêmes contraintes formelles dans ces deux types de ressources.

L'effort de construction du Web Sémantique traduit la volonté de peupler le Web avec le contenu qui possède une sémantique formelle. Ceci donne aux agents automatiques la possibilité de *raisonner à propos du contenu du Web et produire une réponse intelligente face aux situations non rencontrées précédemment.* [2]. Le partage des données ouvertes induit la nécessité de connecter et rendre interopérables les ressources exprimées en langues différentes.

Les données peuvent diverger ostensiblement selon la langue des ressources, leur qualité de peuplement et leur couverture de même que la qualité d'alignement entre les différentes langues dans le cadre des ressources multilingues. L'intégration des ressources existantes apparaît comme le moyen courant permettant de garantir l'interopérabilité de contenu des ressources langagières. Ainsi, les ressources comme ConceptNet (Speer and Havasi [2012]) ou BabelNet (et al. [2012]) intègrent WordNet (Fellbaum [1998]). Dans le contexte industriel, l'intégration totale ou partielle des ressources existantes est fréquemment choisie. Pour des ressources ontologiques, il est souvent impossible d'intégrer directement une ressource langagière pour des raisons formelles. Par conséquent l'amélioration d'interopérabilité termino-ontologique passe par l'enrichissement des ontologies existantes notamment par introduction de nouvelles langues ou par la mise en réseau des ontologies. Cet enrichissement d'ontologie est un processus long et coûteux car, traditionnellement, il requiert la participation des experts et peut difficilement être automatisé.

Deux familles d'approches cohabitent quant à la conception des ressources de connaissance : approches centrées sur le calcul et celles centrées sur l'expert.

Les approches centrées sur le calcul mettent en oeuvre des différentes structures de données telles que textes en langue naturelle, plongements lexicaux, graphes servent de corpus d'apprentissage dans le cadre d'un système apprenant. La peuplonomie et l'acquisition contributive est également utilisée. L'enjeu central de ce type d'approche est l'acquisition et la représentation de la polysémie des items lexicaux ainsi que de leur sémantique. Il s'agit de structurer ces connaissances de la façon la plus explicite possible en représentant les éléments du modèle sous forme discrète. Cette représentation peut être redondante. Ce type d'approche sous-tend les ressources telles que réseaux lexico-sémantiques, graphes conceptuels, etc.

> **Définition 1.2**
>
> Quelques définitions des ressources centrées sur le calcul.
>
> (1) **Réseau sémantique** : graphe orienté et étiqueté composé d'un ensemble d'objets (nœuds), d'une ensemble de liens entre ces objets (arcs orientés et étiquetés) et d'un ensemble d'opération d'exploitation (mécanisme de raisonnement) ;

2. D'après Lassila and McGuinness [2001]

(2) **Réseau lexico-sémantique** : réseau sémantique où les nœuds correspondent aux objets (items) lexicaux et les arcs correspondent aux relations sémantiques et lexicales ;

A travers ces définitions, l'on constate l'approche intuitive à la structuration de la connaissance ainsi que l'absence de définition unique des éléments. La structuration est guidée par la sémantique.

Les approches issues de la logique sont axées sur la notion de *consistance logique*. Il s'agit au contraire de réduire les redondances et de structurer la connaissance de manière abstraite en intégrant les caractéristiques les plus pertinentes dans un modèle qui exprime une conceptualisation partagée sur un domaine de connaissances plus ou moins spécifique.

Définition 1.3

Quelques définitions des ressources issues de l'approche à la structuration des connaissances basées sur la logique :

(1) **Ontologie**

1. une ***conceptualisation d'un domaine à laquelle un ou plusieurs vocabulaires peuvent être associés***. Définie avec un objectif donné, une ontologie exprime un point de vue partagé par une communauté donnée. Une ontologie est représentée dans un langage dont la sémantique permet de garantir les propriétés de celle-ci en termes de consensus, cohérence, partage et réutilisation (d'après Roche [2003]) ;
2. une ***spécification explicite d'une conceptualisation*** (d'après Gruber [1995])

(2) **Terminologie** (d'après ISO-1087-1 [a]) :

1. ensemble des désignations appartenant à une langue de spécialité ;
2. science étudiant la structure, la formation, le développement, l'usage et la gestion des terminologies

(3) **Ontoterminologie** : terminologie dont le système conceptuel est une ontologie formelle (d'après Roche [2007]) [b].

(3) **Graphe conceptuel** : graphe fini, connecté, non orienté et bipartite dont les nœuds du premier type sont appelés "concepts" et les nœuds du deuxième type sont appelés "relations conceptuelles" (d'après Sowa [1976]).

a. `https://edisciplinas.usp.br/pluginfile.php/312608/mod_resource/content/1/ISO_1087-1_2000_PDF_version_\%28en_fr\%29_CPDF.pdf`

b. `http://ontoterminology.com/`

Ce type d'approche est à l'origine de la construction des ontologies, termino-ontologies, terminologies, qui s'inscrivent dans le cadre d'une approche "normative de la communication et de l'échange d'information" comme remarqué par Roche [2007].

Avec ces deux approches, l'approche intuitive semble s'opposer à une approche normative. Outre le modèle et la sémantique de ses composants, la différence entre ces deux familles d'approches réside également dans la façon de concevoir l'optimisation d'accès aux informations contenues dans la ressource. Cette optimisation concerne le coût de l'inférence d'un objet implicite par rapport au coût de la recherche d'un objet explicitement représenté (relation, objet lexical, concept etc.). Les ressources ontologiques explicitent uniquement les propriétés essentielles des objets, les informations pouvant être obtenues par raisonnement ne sont pas explicitement représentées. Les ressources qui traduisent les différents paradigmes de modélisation issues de l'intelligence artificielle peuvent contenir des informations redondantes (par exemple, toutes les formes de surface d'un terme, ses synonymes, ses variantes). Ainsi, lors du parcours de ces ressources, ces objets sont directement accessibles.

Le socle commun pour ces deux groupes approches à la structuration de la connaissance inclut le langage naturel dont le référentiel est utilisé par toutes les approches ainsi que la connaissance générale sur le monde qui sous-tend toute modélisation d'un domaine de spécialité.

1.3 Domaine de l'alimentation comme contexte applicatif

L'alimentation est une activité humaine culturellement ancrée. Cet ancrage détermine les particularités de sa terminologie dont notamment la prédominance du *code* sur le *texte* soit présence de concepts implicites. À titre d'exemple, le verbe *blanchir* renferme la notion de "exposer brièvement à la chaleur". La terminologie de la nutrition, quant à elle, interagit moins avec le fond usuel de la langue. Cependant, elle aussi est culturellement ancrée notamment en ce qui concerne l'interprétation de la quantité des nutriments à la jonction avec la terminologie de la cuisine. Les quantités consommées ne sont pas les mêmes selon les pays et l'origine des produits, la composition des recettes qui portent le même titre peut différer ostensiblement. Par ailleurs, le contexte dans lequel se trouvent les textes de cuisine est un contexte multilingue. De nombreux concepts de ce domaine et de celui de la nutrition participent au fondement du discours collectif désormais multilingue, à l'entente et à l'identification sociale. Ce discours collectif est exprimé à travers les textes de spécialité dont le texte principal est la recette de cuisine. Dans le contexte d'*analyse sémantique* de celle-ci, la représentation de la connaissance implicite et, plus particulièrement, des événements implicites prend toute son importance. Par exemple, dans l'énoncé "*faire revenir les oignons*", les événements implicites sont "éplucher les oignons", "rincer les oignons", "prendre une planche à découper", "utiliser un couteau", "faire chauffer de l'huile dans une poêle" etc. Une des solutions possibles à cette problématique peut être le calcul des structures sémantiques basé sur les éléments explicités au sein d'une ressource de connaissance. Ce calcul peut permettre

d'enrichir (augmenter) le contenu fourni en entrée et de proposer par ce biais des événements typiques possibles pour un objet donné. Une autre thématique dans ce même cadre est la problématique de substitution des ingrédients et des ustensiles afin de pouvoir adapter une recette de cuisine donnée.

Jusqu'à la période récente, les tentatives de modélisation de recette de cuisine en tant qu'exemple de texte procédural la considéraient comme une structure "immuable" dont on analyse le contenu textuel sans chercher à expliciter les concepts implicitement présents. Le défi de fouille de textes 2013 (DEFT 2013), les travaux tels que Dufour-Lussier et al. [2014] et Gaillard et al. [2015] dans le cadre du projet Taaable (Cordier et al. [2014]) sont des exemples de grande qualité construits autour de ce concept de recette statique.

Cependant, la recette de cuisine telle qu'elle existe notamment sur le Web a autant de lectures possibles que d'utilisateurs. Tel un réseau dynamique, le texte de la recette de cuisine est constamment "réécrit" en fonction des restrictions alimentaires, préférences et disponibilités. La connaissance nécessaire à l'analyse d'une recette de cuisine englobe la connaissance sur la *composition* (dont la composition nutritionnelle) et la connaissance sur la *transformation*.

En dehors des approches focalisées sur les réseaux de saveurs qui traitent les recettes comme des "sacs d'ingrédients" dont notamment Ahn et al. [2011], les travaux existants se concentrent sur l'analyse des recettes de cuisine en tant que suites d'instructions. De nombreuses approches implémentent différentes types d'apprentissage supervisé. Mori et al. [2012] utilisent les données annotées pour extraire des structures prédicat-arguments des recettes en japonais afin de représenter la recette comme un flux. Dans le paradigme *Semantic Role Labeling*, Malmaud et al. [2014] proposent un processus décisionnel markovien dans lequel les ingrédients et les ustensiles sont propagés à travers l'ordre temporel des instructions. D'autres auteurs se placent dans le paradigme de raisonnement à partir des cas comme Dufour-Lussier et al. [2012] et Müller and Bergmann [2015].
Les méthodes existantes d'analyse des textes de cuisine convergent sur la nécessité de disposer d'un contexte riche autour des termes présents dans les corpus à analyser. Ce contexte se présente notamment comme :

— méta-langage spécifique *ex.* projet SOUR CREAM (Tasse and Smith [2008]), SIMMR (Jermsurawong and Habash [2015] [3]) ;

— structures dynamiques telles que vecteur d'état latent chez (Malmaud *et al.*, *op. cit.*), vocabulaires ou ontologies construits à la volée ;

— recours à de vastes ressources sous forme de graphe (ontologies, réseaux sémantiques) et leur exploitation par projection.

Parmi les projets qui ont porté sur la construction des ressources et applica-

3. *Simplified Ingredient Merging Map in Recipes*

tions spécifiques autour de l'alimentation, PIPS [4] et OASIS [5] furent les premiers projets d'envergure au niveau européen axés sur la promotion de l'alimentation saine et des systèmes de conseil dans ce domaine. Certains travaux ont porté sur la génération des menus pris comme un problème d'optimisation à plusieurs niveaux (MenuGene Pinter et al. [2012]) [6]), d'autres sur la proposition des menus alternatifs (Semanticook, Akkoç and Cicekli [2011]). Le système Taaable (Cordier et al. [2014]) [7]) basé sur les cas a également évolué dans ce sens en intégrant les valeurs nutritionnelles ainsi que les restrictions alimentaires et la provenance géographique des recettes. Cependant, dans ce paradigme la représentation formelle (graphe orienté acyclique) des recettes en termes de propriétés sémantiques est requise. La base de connaissances du domaine et le formalisme choisi dans le cadre de ce système permettent en effet d'exprimer les incompatibilités basées sur la composition nutritionnelle et la relation de subsomption ex. *sans alcool*, *sans cholestérol*, *sans gluten*, *goutte*, *végan*, *végétarien*, *sans noix*.

Dans le cadre de *construction ontologique*, la recette de cuisine évoque les questionnements suivants :

— Représentation de la composition de la recette soit représentation des "valeurs" qui correspondent aux ingrédients et de leurs "modifieurs" (appelés *traits*, *attributs*, *qualités*, *caractéristiques* par différents auteurs). ;

— Représentation des processus (transformation des ingrédients, modification de leur état au fil du temps de préparation).

Ces deux axes concernent la conception des patrons de conception d'ontologie (*Ontology Design Patterns*) adaptés à la représentation des différents composants de la recette de cuisine. Sans apporter une réponse exhaustive à ces problématiques, nous proposons dans le présent travail une façon de construire une ressource de connaissance multilingue et de l'utiliser afin d'enrichir les ressources de connaissance ontologiques et d'analyser les textes de cuisine.

1.4 Mécanismes de construction des ressources de connaissance

1.4.1 Intégration

L'**intégration** est un ensemble de méthodes pour l'identification et l'inclusion des données structurées issues des ressources existantes dans une ressource structurée en cours de construction. Dans un contexte industriel de construction d'une

4. http ://cordis.europa.eu/project/rcn/71245_en.html. Modélisation des concepts *régime alimentaire*, *aliment*, *calories* etc.
5. http ://www.oasis-project.eu/
6. http ://www.menugene.com/
7. http ://wikitaaable.loria.fr

ressource de connaissance, l'optimisation des coûts est importante. L'identification des ressources pouvant être intégrées peut alors s'appuyer sur leur disponibilité, leur couverture et leur structure. Selon les langues, les ressources peuvent être plus ou moins abondantes, elles peuvent couvrir un domaine de spécialité de façon plus ou moins satisfaisante. Outre ces aspects liés à la richesse des ressources en ce qui concerne la quantité et la qualité de données qu'elles offrent, l'expressivité, la granularité, la couverture ainsi que la méthode de construction semblent être les critères de décision quant à l'intégration d'une ressource préexistante.

De manière générale, les ressources peuvent être structurées ou non structurées. Dans le contexte multilingue, la connaissance partagée peut trouver son expression à travers les textes de spécialité qui sont des exemples des ressources non structurées. Ces ressources nécessitent des mécanismes d'extraction terminologique et ne sont pas pertinentes pour les processus de l'intégration mais pour celui de l'augmentation (section 1.4.2). Les ressources structurées varient en fonction de plusieurs ensembles de critères :

— **expressivité.** Les types de liens présents dans les ressources structurées (ex. synonymie, hyperonymie etc.) déterminent l'expressivité de la ressource. Ces types de liens sont déterminés par le modèle formel choisi pour une ressource donnée ;

— **couverture.** La couverture d'une ressource structurée est déterminée par sa construction (langues concernées) et par son contexte applicatif (domaine de spécialité particulier, usager précis i.e. lexicographe, traducteur etc.) ;

— **granularité**. Taxonomique ou sémantique, la granularité correspond au degré de détail local par rapport à un critère global (distinction des sens d'usage, profondeur d'une taxonomie).

En termes d'***expressivité***, les différents types de ressources pouvant être intégrées se répartissent comme représenté sur le tableau 1.2 et sur la figure 1.3.

En faisant le rapprochement avec le spectre d'ontologies détaillé notamment par Lassila and McGuinness [2001] qui définit une série de spécifications possibles d'ontologie comme représenté sur la figure 1.4, nous pouvons avancer qu'il existe une corrélation entre l'expressivité et la précision formelle du modèle de représentation des connaissances.

En termes de ***couverture***, les ressources se subdivisent d'une part en ressources de connaissance générale et ressources de *spécialité* et, d'autre part, en ressources *monolingues* et *multilingues*. La plupart des ressources disponibles sont des ressources de spécialité monolingues. Ainsi, le nombre de ressources pouvant être qualifiées de "générales", de "multilingues" et de "générales et multilingues" est relativement restreint. Deux phénomènes peuvent être constatés : la redondance et l'asymétrie de couverture langagière des ressources existantes. De nombreuses

Expressivité (relations) / Type de ressource	lemme	traits syntaxiques POS	traits lexicaux (variantes, synonymes)	traits sémantiques: hyperonyme hyponyme	traits sémantiques: partie-tout matière	traits sémantiques: relations prédicatives (agent, patient..)	traits sémantiques: caractéristique manière	traits sémantiques: relations thématiques (lieu, instrument...)
Ontologie du domaine *(BBC Food, Agrovoc,PIPS)*				■				
Base de données lexicales *(Wiktionnary, DBNary)*	■	■	■					
Ressources lexicales par similarité type WordNet *(RWN,WordNet)*	■	■	■	■	■		■	
Ressources lexicales type RLS *(ConceptNet)*	■	■	■	■	■	■		
Vocabulaire contrôlé *(Composition nutritionnelle des aliments : Okali, Ciqual)*					■			
Mémoire de traduction *(IATE)*	■	■	■					
Léxique, glossaire	■	■						

FIGURE 1.2 – Expressivité des différentes types de ressources de connaissance, en bleu - présence certaine d'information, en gris - présence possible d'un type d'information.

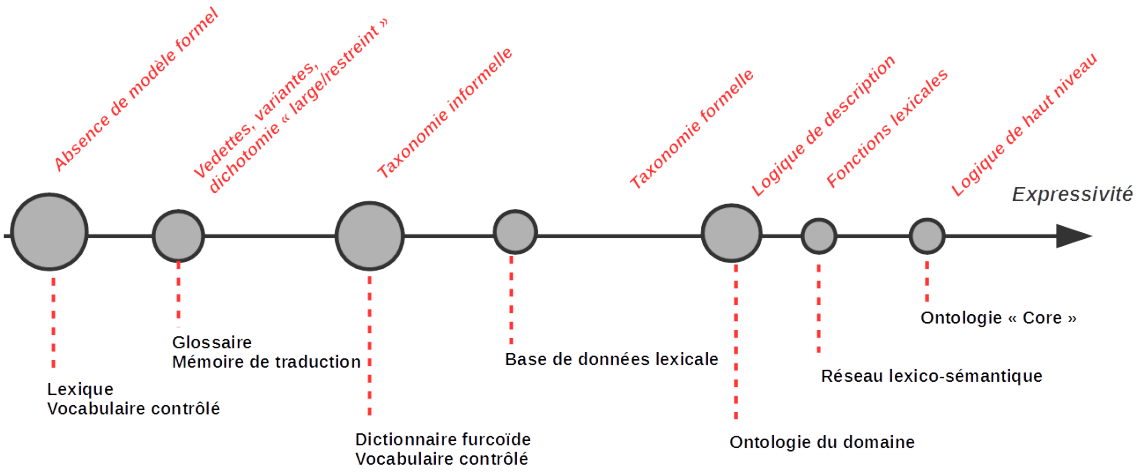


FIGURE 1.3 – Expressivité des différentes types de ressources de connaissance, la taille du point indique la *disponibilité* des ressources

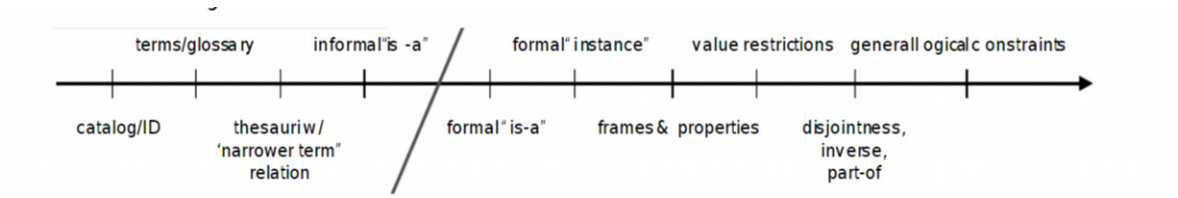


FIGURE 1.4 – Spectre d'ontologie, cité d'après Lassila and McGuinness [2001]

.

ressources de connaissance s'appuient sur les mêmes ressources de connaissance telles que Wikipedia (Wikidata, DBNary), Wiktionnary (DBNary), GeoNames et WordNet. Par conséquent, l'intersection entre les ensembles de données contenues dans ces ressources peut être importante.

En termes de ***granularité***, il est important de distinguer la granularité des composants et la granularité des données. La granularité des composants est liée au choix du modèle qui permet de structurer la connaissance. Par exemple, dans le cadre du réseau lexico-sémantique RezoJDM (Lafourcade [2011]), 138 types de relations (arcs) sont distingués [8], les arcs sont orientés, pondérés et peuvent se voir associer une méta-information qui précise ou contextualise la relation (appelée *annotation*). Dans le cadre du réseau ConceptNet, 37 types de liens ont été modélisés [9], l'information de direction est absente, les arcs sont pondérés.

Pour une ressource langagière, la granularité correspond principalement à la distinction des différents sens au sein des différentes langues telle qu'elle observée. Ainsi, le terme anglais *stew* correspond à la fois à une préparation de type *ragoût* et à une préparation de type *pot-au-feu*. De même, le terme français légume peut couvrir à la fois *legume* (légumineuse) et *vegetable* (légume au sens plus large) en anglais. Même si *légumineuse* existe en français, il s'agit d'un terme plus spécialisé. Dans une certaine mesure, la granularité peut s'exprimer au niveau des relations sémantiques, notamment dans le cas des lieux typiques et des quantificateurs.

1.4.2 Augmentation

L'**augmentation** d'une ressource langagière est entendue dans le présent manuscrit comme un ensemble de méthodes d'ajout de données extérieures (peuplement exogène) de la ressource. Contrairement à l'intégration, l'augmentation ne s'intéresse pas aux problématiques de modèle et ne cherche qu'à ajouter de nouvelles données dans un modèle pré-existant. Dans le cadre de la construction des ressources de spécialité, l'augmentation vise à apporter de la connaissance de spécialité à partir des ressources appropriées telles que des terminologies, listes, corpus spécifiques. Une des sources de données des méthodes d'augmentation est le ***corpus de textes en langue naturelle***. Dans le cadre d'exploitation des corpus, ses principaux enjeux consistent à *identifier* et *extraire* les informations structurées conformément au modèle pré-établi (par exemple, nœuds et arcs pour les ressources sous forme de graphe) à partir des données non structurées (textuelles). La deuxième source de données pour l'augmentation est l'***apprentissage ouvert*** (données acquises par peuplonomie (externalisation ouverte), grâce aux contributeurs humains). Face à ces données contributives, l'enjeu central est la *validation des contributions*. S'il s'agit de la contribution experte, le consensus entre plusieurs experts peut servir à valider les propositions. La troisième source de données à considérer dans le cadre de l'augmentation est l'***apprentissage par la tâche***, notamment la tâche d'analyse sémantique. Une telle analyse se sert des informations déjà présentes dans une ressource langagière, diagnostique les lacunes et les incohérences de la ressource et déclenche

8. `http://www.jeuxdemots.org/jdm-about-detail-relations.php`
9. `https://github.com/commonsense/conceptnet5/wiki/Relations`

les processus d'acquisition des données manquantes via des méthodes de récupération des données externes ou par apprentissage ouvert.
Les auteurs dans Mitchell et al. [2015] définissent les exigences vis-à-vis d'un algorithme d'apprentissage comme la capacité d'apprendre :

- à partir d'une grande variété de types de connaissance ;
- à partir de l'expérience auto-contrôlée ;
- de manière incrémentale, en se servant des connaissances acquises pour acquérir de nouvelles connaissances ;
- de manière auto-réflexive où la capacité de formuler de nouvelles représentations et de nouvelles tâches évite à l'apprenant la stagnation.

Ainsi, dans le cadre de son augmentation, la ressource n'est pas utilisée de manière statique, elle est constamment améliorée via les différents processus qui l'utilisent.

1.4.3 Consolidation

Si l'augmentation correspond à l'*ajout* de nouveaux éléments dans la ressource, la **consolidation** est l'ensemble des méthodes endogènes destinées à *produire de nouveaux éléments* à partir des éléments déjà présents dans une ressource langagière. Aucune donnée extérieure n'est utilisée. Il s'agit des mécanismes de raisonnement, des mécanismes discursifs (comportant des *médiations* soit des ensembles de prémisses). Dans le cadre du présent travail, il s'agit d'inférence des relations sémantiques ainsi que de méta-informations (annotations).

L'inférence consiste à créer (inférer) de nouveaux éléments à partir des informations et structures déjà présentes dans une ressource. Dans le cadre de notre approche, les informations pré-existantes sont de termes et leurs relations présentes dans le réseau. Les structures qui nous intéresseront particulièrement sont les termes jugés similaires et leurs relations d'une part et les termes identifiés comme polysémiques et leur arbres d'usages (sens d'usage), d'autre part.

1.4.4 Alignement

L'**alignement** est l'ensemble de méthodes destinées à harmoniser une ressource langagière qui comporte plusieurs partitions d'éléments afin que les données contenues dans ces partitions soient interopérables. Dans le cadre d'une ressource langagière multilingue, il s'agit de faire en sorte que les vocables et les sens des vocables inclus dans une partition A possèdent un maximum d'équivalences dans une partition B. Dans le cadre du présent travail, l'alignement concerne principalement le calcul des raffinements de sens et, éventuellement, l'inférence translingue des raffinements. L'enjeu principal de l'alignement est la définition et le calcul de proximité entre les ensembles de données contenues dans les partitions à aligner via l'exploration des relations sémantiques.

Dans la section suivante, nous allons détailler les ressources existantes qui illustrent la variété des méthodes de construction et des modèles existants. Nous accorderont de l'attention non seulement aux ressources de spécialité, mais également aux ressources de connaissance générale car la connaissance générale permet de définir la connaissance de spécialité et ne peut pas être considérée séparément de celle ci.

1.5 Ressources existantes

1.5.1 Ressources de connaissances pour la recherche d'information sur le Web fondées sur les entités

Il existe aujourd'hui un certain nombre de ressources basées sur les faits et construites via les techniques non supervisées d'extraction des entités et des faits sur les entités depuis les données disponibles sur le Web.

YAGO2 (*Yet Another Great Ontology*), Suchanek et al. [2007]. YAGO2 est une base de connaissances sémantiques constituée automatiquement à partir d'informations extraites de Wikipédia (catégories, redirections, infobox), WordNet (synsets, hyponymie) et GeoNames [10]. Cette ressource est basée sur les entités et les faits sur ces entités représentés sous forme de liens. Afin de l'intégrer au Web des données, YAGO2 est liée aux ontologies DBpedia [11] et SUMO [12]. YAGO2 a été utilisée par le programme d'intelligence artificielle Watson.

YAGO (actuellement YAGO2) est un projet de l'Institut Max-Planck Informatique, Sarrenbruck, et de l'Université Telecom ParisTech, Paris. A présent, YAGO contient plus de 10 million entités et plus de 120 million de faits sur les entités.

NELL (*Never-Ending Language Learning*), Carlson et al. [2010]. "Lire le Web" ("*Read the Web*") est le nom du projet de recherche qui vise la création d'un système informatique auto-apprenant qui extrait les faits à partir du texte non structuré trouvé sur le Web, puis essaie d'améliorer sa performance de lecture de façon à pouvoir extraire de nouveaux faits.

NELL est un projet de l'Université Carnegie Mellon. A présent NELL contient 50 millions de faits sur les entités candidats (`candidate beliefs`) *obtenus à partir du Web. plus de 2 millions de ces faits ont un score de confiance élevé.*

KnowledgeVault (littéralement : "la voûte de la connaissance"), Dong et al. [2014] est une base de connaissance factuelles probabiliste qui combine des extractions issues des contenus Web (analyse de texte, données tabulaires, structure des pages Web, annotations, etc.) avec des connaissances dérivées de référentiels de connaissances existants. Cette base utilise les méthodes de *machine*

10. `www.geonames.org`
11. `https://wiki.dbpedia.org/`
12. `http://www.adampease.org/OP/`

learning dans le but de fusionner les sources d'informations distinctes. L'information est stockée en format RDF où les types d'entités et les prédicats viennent d'un référentiel (ontologie) prédéfini. KnowledgeVault est structuré de façon similaire à NELL, YAGO, DeepDive. KnowledgeVault contient environ 302 millions de données factuelles. KnowledgeVault se veut un référentiel structuré de connaissance indépendant de la langue.

KnowledgeVault est un projet de Google Inc. depuis 2012. les auteurs dans Dong et al. [2014] annoncent 45 millions d'entités et environ 271 millions de faits avérés sur les entités.

KnowItAll, Etzioni et al. [2005] est une ressource de données factuelles inspirée par les travaux de Hearst [1992] notamment sur l'acquisition automatique des hyponymes depuis des corpus de grande taille (méthode qui s'appuie sur les relations hiérarchiques de WordNet Fellbaum [1998]). Le système de construction de la ressource est un système non supervisé qui utilise un ensemble de huit domaines de connaissance indépendants et des modèles d'extraction afin de générer des faits candidats. Il teste automatiquement la plausibilité des faits candidats en utilisant le score d'information mutuelle (*Pointwise Mutual Information*) et associe une probabilité à chaque fait. KnowItAll a été amorcé avec 50 000 instances de classe.

KnowItAll est un projet de l'Université de Washington depuis 2004.

DeepDive[13], Shin et al. [2015] est un système permettant de créer des données structurées (sous forme d'une base de données) à partir d'informations non structurées (documents textuels, tableaux, images etc.). Il s'agit de plus d'extraire des relations complexes entre des entités et faire des conclusions sur des faits impliquant ces entités. DeepDive est un système basée sur l'apprentissage, il est entraîné grâce à un système de règles et à un système d'annotation des données appelé *Mindtagger*.

DeepDive est un projet de l'Université de Stanford (son développement a été arrêté depuis 2017). La taille de cette ressource en 2014 était d'environ 2,7 millions d'entités et 7 millions de faits sur les entités.

Dans le courant des ressources basées sur les données à grande échelle issues du Web, se distingue **Probase**[14] Wu et al. [2012], une taxonomie probabiliste pour la compréhension du texte. Le monde est utilisé comme modèle. Le but de cette ressource est d'aider la machine à mieux comprendre la communication humaine. La construction de Probase a utilisé un algorithme itératif d'apprentissage pour extraire les paires de termes du texte Web et un algorithme de construction de taxonomie pour connecter ces paires de termes dans une structure hiérarchique. L'espace conceptuel de Probase contient près de 2,7 millions de concepts. C'est un espace conceptuel 8 fois plus grand que celui de YAGO.
Probase est un projet de Microsoft Research.

13. `http://deepdive.stanford.edu/`
14. `https://www.microsoft.com/en-us/research/project/probase/`

Les ressources que nous avons détaillées illustrent des approches efficaces à la conception de ressources interopérables grâce à l'utilisation des données ouvertes. Ces ressources sont utiles notamment dans le cadre de reconnaissance des entités et d'annotation des textes. La conception même de ces ressources les rend axées sur les relations qui peuvent exister entre les entités à travers les faits dont notamment les relations hiérarchiques (hyperonymie, hyponymie) et l'appartenance à une ou plusieurs catégories. Par conséquent, tout un ensemble de relations plus fines échappe à ce type de ressources et nécessite une approche différente. Par exemple, pour le terme *stew* la base NELL donne les précisions suivantes :

visualizablething(76.5%)
food(73.1%)
beverage(57.0%)

soit un ensemble de faits qu'il serait impossible d'exploiter dans le cadre d'analyse d'un texte de spécialité.

De son côté, YAGO exploite les domaines WordNet qui restent très génériques et possède outre les propriétés RDF/RDFS (*subClassOf, type, range, domain, comment* etc.) les relations spécifiques factuelles suivantes :

— Propriétés fonctionnelles relatives à l'emplacement dans l'espace (l'emplacement est un emplacement permanent, géographique etc.) : `isLocatedIn`, `isPlacedIn`, `wasBornIn` etc. ;

— Propriétés relatives au temps (échelle temporelle annuelle) : `diedIn`, `diedOnDate`, `occursIn`, `wasBornIn` etc. ;

— Propriétés relatives au temps relatif : `endsExistingOnDate`, `objectEndRelation`, `occursIn`, `wasBornIn` etc. ;

— Propriétés relatives aux entités telles que : `actedIn`, `hasGender`, `isKnownFor`, `isPlacedIn` etc. ;

— Propriétés relatives aux actions telles que :`created`, `directed`, `edited`, `worksAt` etc.

Comme le montrent ces exemples, la nature des relations basées sur les faits, empêche leur utilisation plus ciblée notamment dans le cadre d'analyse des textes de spécialité ou de construction des ressources de spécialité. Une autre particularité que nous remarquons est le recours à l'humain dans le cadre d'évaluation (YAGO), d'annotation (DeepDive), ainsi que l'utilisation de ressources créées par les experts (utilisation récurrente de WordNet).

1.5.2 Ressources multilingues contributives et leur structuration en format de données liées.

Les ressources contributives multilingues telles que Wikipedia et Wiktionnary qui utilisent un format non interopérable (format *wiki*[15] qui nécessite d'être décodé). Le format fait objet d'une recommandation. Ainsi, une perte d'information due à l'inconsistance de format est inévitable lors de l'extraction des informations depuis les pages Wikipedia vers les formats structurées adaptés au traitement automatique, notamment traitement automatique des langues.

Wikipedia est une encyclopédie collaborative qui couvre aujourd'hui plus de 250 langues pour lesquelles elle est d'une richesse variable. Les éditions les plus importantes contiennent plusieurs millions d'articles. Ainsi, Wikipedia est une ressource très utilisée en tant que source de données non structurées (utilisation en tant que corpus), de données structurées (extraction des taxonomies à partir des catégories Wikipedia), extraction des terminologies etc.

DBpedia (Bizer et al. [2007]) est une ontologie OWL qui contient les données Wikipedia compatibles avec ce format. Cette ressource est le fruit du projet communautaire et universitaire qui vise à extraire et exploiter les données contenues dans Wikipedia en les rendant ainsi accessibles au traitement automatique grâce à l'utilisation des standards du Web Sémantique.

Wiktionnary est un dictionnaire collaboratif. Des particularités de format sont parfois introduites dans les recommandations fournies aux contributeurs des différente éditions.

DBNary (Sérasset [2012]) est une extraction depuis Wiktionnary dans un format d'ontologie RDF en utilisant le vocabulaire OntoLex-Lemon (McCrae et al. [2017]). Il s'agit d'une ressource interopérable avec d'autres ressources qui utilisent le même format. DBNary utilise également des extensions. Un travail sur la désambiguïsation et l'alignement par sens basé sur les gloses a été effectué par Tchechmedjiev [2016].

FreeBase (Bollacker et al. [2008]) est une base de connaissances construite de façon collaborative grâce à un moteur de structuration de connaissance efficace. FreeBase a été récupéré par Google, c'est une base de connaissance structurée utilisée notamment dans le cadre de création de KnowledgeVault.

15. Le *wiki* est une forme particulière de site Web qui peut être édité par tous. Le *format wiki* est un format textuel spécifique qui contient le marquage wiki (*wiki markup*). Les caractères tels que asterisks, apostrophes etc. y ont une fonction spéciale qui peut dépendre de leur position. Dans le cadre de utilisation du format *wiki* pour encoder les pages Wikipedia, la structure d'une page Wikipedia est en partie imposée : infobox, titre, résumé, sommaire, entêtes etc.). Cependant, le format de contenu fait objet d'une simple recommandation.

1.5.3 Ressources interlingues

UNLKB (*Universal Networking Language Knowledge Base*) [16] est une base de données interlingue issue de l'approche *UNL (Universal Networking Language)*. L'idée qui sous-tend cette approche est que l'information exprimée à travers les langues naturelles peut être exprimée formellement en tant que réseau sémantique. Ce réseau est constitué de trois types d'éléments : les mots universels (*Universal Words*), les relations universelles (*Universal Relations*) et les attributs universels (*Universal Attributes*). Les mots universels appartiennent à un langage intermédiaire pivot. Il est possible de traduire entre le langage naturel et le langage intermédiaire grâce à un convertisseur et un déconvertisseur. Les mots exprimés en langage naturel sont traduits en langage intermédiaire par un convertisseur. Les nœuds du réseau de type UNL sont des unités discrètes exprimées sous forme de noms, verbes, adjectifs et adverbes en langue naturelle. Tout autre contenu sémantique est représenté sous forme d'attributs ou de relations. Si une valeur sémantique donnée ne peut pas être exprimée sous forme d'un mot du lexique dont la catégorie grammaticale est une catégorie ouverte, cette valeur ne peut pas être exprimée sous forme de mot universel quelle que soit sa réalisation dans d'autres langues naturelles.

1.5.4 Construction experte

Réseau Lexical du Français (RLF, ATILF [2017]). Il s'agit d'un modèle du lexique formel du français contemporain. Le lexique du français est modélisé sous forme de graphe dont les nœuds sont principalement les sens lexicalisés dans la langue et les arcs représentent des liens paradigmatiques et syntagmatiques standardisés. Le standard adopté est celui du système des fonctions lexicales (d'après les principes décrits par Polguère [2009] afin de mettre en œuvre les principes introduits par Jolkovsky and Mel'čuk [1967] et tout au long des travaux de Igor Mel'čuk).

WordNet (Fellbaum [1998]) est un réseau lexical de l'anglais qui couvre le domaine de la langue générale. WordNet est organisé autour des ensembles de synonymes et des relations lexicales, mais aussi sémantiques (hyperonymie, hyponymie, méronymie) qui peuvent exister entre ses ensembles. la construction de cette ressource a bénéficié des travaux psycholinguistiques sur le fonctionnement de la mémoire humaine. Ainsi chaque ensemble de synonymes (*synset*) correspond à un sens (décrit par une glose spécifique). WordNet est une ressource très utilisée pour l'acquisition des relations sémantiques, des taxonomies, pour l'analyse des textes. Elle souvent intégrée dans d'autres ressources (ConceptNet, YAGO, BabelNet). D'abord basées sur la traduction automatique, des travaux spécifiques ont été menés pour produire des WordNets pour d'autres langues de même que les ressources multilingues dont les exemples sont RuWordNet (Loukachevitch [2016]), EuroWordNet (Hirst [1999]).

16. `http://www.unlweb.net/`

HowNet (Dong et al. [2010]) [17] est une base de données de connaissance générale qui a mis en évidence les relations inter-conceptuelle et inter-attribut entre les concepts dont les traces sont mises en évidence lors de l'analyse des équivalences entre les lexiques chinois et anglais. Contrairement à WordNet dont la conception a été celle d'une base de données lexicale (conçue pour être consultée par les utilisateurs humains), HowNet est un système de connaissance générale basé sur les concepts et orienté machine.

FrameNet est une ressource fondée sur la théorie des cadres sémantiques (Ruppenhofer et al. [2006], Baker et al. [1998]). Un cadre sémantique correspond à un "ensemble d'attributs, de valeurs associées et de contraintes" (d'après Rich and Knight [1990]). La nature de ces éléments est syntactico-sémantique. La construction experte par les lexicographes et les linguistes a permis d'annoter sémantiquement des corpus anglais et d'en extraire des phrases exemples qui ont permis de définir les cadres. De nombreux projets basés sur cette approche ont vu le jour : FrameNet multilingue [18], ASFALDA (FrameNet pour le français) [19] etc. Le lexique hiérarchisé **VerbNet** [20](Kipper et al. [2000]) combine les cadres FrameNet et les synsets de WordNet pour classer les verbes anglais afin de former une arborescence.

Parmi les ressources construites par les experts, plusieurs exploitent UNL pour construire des bases lexicales monolingues reliées entre elles par le pivot interlingue. Les mots universels (UW) garantissent l'ancrage ontologique des bases lexicales ainsi structurées. Notamment, Dikonov [2013] [21] décrit le développement des bases lexicales sous forme de réseau sémantique dont la structure (ontologique !) est basée sur l'ontologie SUMO. La ressource obtenue dans le cadre de ces travaux a un format complexe dû à l'ajout de multiples liens (notamment vers WordNet) et des informations sémantiques variées. Cette complexité de format peut rendre l'utilisation d'UNL problématique dans le cadre de la construction des ressources de spécialité et dans un contexte industriel.

Le projet **Papillon** (Tomokiyo et al. [2000]) exploite également une architecture avec pivot interlingue artificiel. Dans le cadre de ce projet, une base lexicale multilingue a été construite de façon collaborative. L'architecture de cette base est une architecture avec un pivot. Les entrée des dictionnaires monolingues sont reliées entre elles par des liens interlingues (également appelées acceptions, Sérasset [1994]) formant le dictionnaire pivot. Une hiérarchie des acceptions interlingues permet de se prémunir d'un éventuel contraste artificiel [22].

17. `http://www.keenage.com/html/e_index.html`
18. `https://framenet.icsi.berkeley.edu/fndrupal/node/5548`
19. `https://sites.google.com/site/anrasfalda/`
20. `https://verbs.colorado.edu/~mpalmer/projects/verbnet.html`
21. `https://github.com/dikonov/Universal-Dictionary-of-Concepts`
22. Le **phénomène contrastif artificiel** correspond à une perte d'information discriminatoire lors d'un alignement qui ignore ces distinctions lorsqu'on utilise un pivot naturel (langue naturelle). En effet, une langue naturelle aura une conceptualisation et des lexicalisations divergentes par rapport aux autres langues contenues dans la ressource multilingue (Sérasset [1994]).

La construction experte a pour inconvénient d'être très coûteuse et longue car elle nécessite la participation de nombreux experts. À titre d'exemple, la construction de WordNet a duré près de 25 ans et a engendré un coût de plusieurs millions de dollars américains. L'alternative à ce mode de construction des ressources est l'externalisation ouverte soit l'utilisation des ***jeux avec un but***[23] pour l'acquisition lexicale et sémantique dans le cadre de la construction d'une ressource donnée.

Construction par externalisation ouverte

ConceptNet (Speer and Havasi [2012]) est un réseau lexico-sémantique qui a été créé, à l'origine, grâce aux jeux en ligne dans le cadre du projet *Open Mind Common Sense* lancé en 1999 au MIT Media Lab. Aujourd'hui cette ressource continue à croître par intégration et connexion avec d'autres ressources collaboratives (DBPedia) ou créées par les experts (WordNet). Les jeux avec un but (Verbosity) ne semblent plus être au centre du modèle d'acquisition actuel de ConceptNet.

RezoJDM (Lafourcade [2007]) est un réseau lexico-sémantique du français construit grâce à un ensemble de jeux avec un but. Il s'agit d'un graphe orienté, typé et pondéré dont les nœuds représentent les items lexicaux et les arcs - les relations sémantiques et lexicales entre ces items. Cette ressource a inspiré nos travaux de construction d'un réseau lexico-sémantique multilingue et sera présentée en détail dans le chapitre 2.

Ressources obtenues automatiquement

Parmi les ressources construites automatiquement, **BabelNet** (et al. [2012]) est la ressource majeure par sa couverture (15 millions de synsets, 284 langues) et par la richesse de son écosystème. Construit à partir de WordNet, BabelNet complète les synsets avec les mots des autres langues grâce aux liens de pages multilingues présents dans Wikipedia. Un système de traduction automatique est utilisé afin de compléter les définitions manquantes.

Outre WordNet et Wikipedia, BabelNet a été construit par intégration automatique de ressources telles que OmegaWiki[24], GeoNames[25], FrameNet[26] etc. BabelNet met en œuvre une architecture avec un pivot naturel (la langue anglaise).

D'autres ressources construites automatiquement ont été détaillées par Tchechmedjiev [2016] :

23. ***Games With a Purpose*** (GWAP) est également le terme courant pour cette méthode de construction.
24. `http://www.omegawiki.org`
25. `http://www.geonames.org/`
26. `https://framenet.icsi.berkeley.edu/fndrupal/`

1. ressources lexico-sémantiques sous forme de réseau :
   (a) Uby (Gurevych et al. [2012]) intègre en grande partie les mêmes ressources que BabelNet et le format Lexical Markup FrameWork (LMF) [27] ;
   (b) OpenMultilingualWordNet (Bond and Foster [2013]) ;
2. graphes de traduction : PanLex (Kamholz et al. [2014]), PanDictionary (Mausam et al. [2009])

1.5.5 Ressources spécialisées liées à l'alimentation.

Ressources terminologiques

Parmi les ressources terminologiques liées à l'alimentation, beaucoup sont de petite taille. Parmi celles-ci, on distingue deux thésaurus/vocabulaires contrôlés : Langual et Agrovoc.

L'appellation **Langual** [28] correspond à "Langua aLimentaria". Il s'agit d'une méthode automatique pour décrire, capturer et extraire les données sur les aliments. La construction de cette ressource a commencé dans les années 1970 au Centre de Sécurité Alimentaire et Nutrition Appliquée aux États-Unis. Il s'agit aujourd'hui d'une base de connaissances qui regroupe les informations sur la composition nutritionnelle des aliments. Ces informations sont reprises et adaptées en français dans le cadre du projet Ciqual [29]. La version actuelle de Langual (Langual^{TM}) contient 12 605 descripteurs.

Agrovoc est "un vocabulaire contrôlé couvrant tous les domaines d'intérêt de l'Organisation des Nations Unies pour l'alimentation et l'agriculture (FAO), notamment ceux ayant trait à l'alimentation, la nutrition, l'agriculture, la pêche, la foresterie, l'environnement, etc. Il est publié par la FAO et révisé par une communauté d'experts" [30].

Outre ces deux ressources spécialisées, une base de mémoires de traduction de l'Union Européenne **IATE** (*Interactive Terminology for Europe*) [31] contient un sous-ensemble de lexiques alignés pertinents pour la construction d'une ressource que nous abordons. Cette base est utilisée par les institutions européennes depuis 2004 dans le but de collecter et gérer les terminologies spécialisées.

La ressource construite dans le cadre de nos travaux fait appel directement à ces ressources terminologiques spécialisées directement (IATE, Ciqual, Agrovoc) ou indirectement (Langual).

27. Défini par le standard international pour le traitement automatique des langues et dictionnaires machine.
28. `http://www.langual.org/`
29. `https://ciqual.anses.fr/`
30. `http://aims.fao.org/fr/agrovoc`
31. `https://iate.europa.eu/home`

Ressources ontologiques.

Parmi les ressources ontologiques relatives au domaine de spécialité que nous abordons se distinguent :

— des ressources qui servent à décrire les objets (aliments, produits, ustensiles, composition nutritionnelle etc.) et techniques de transformation relatifs à la cuisine et à la nutrition ;

— des ressources qui se focalisent sur un type particulier de maladie et de l'alimentation

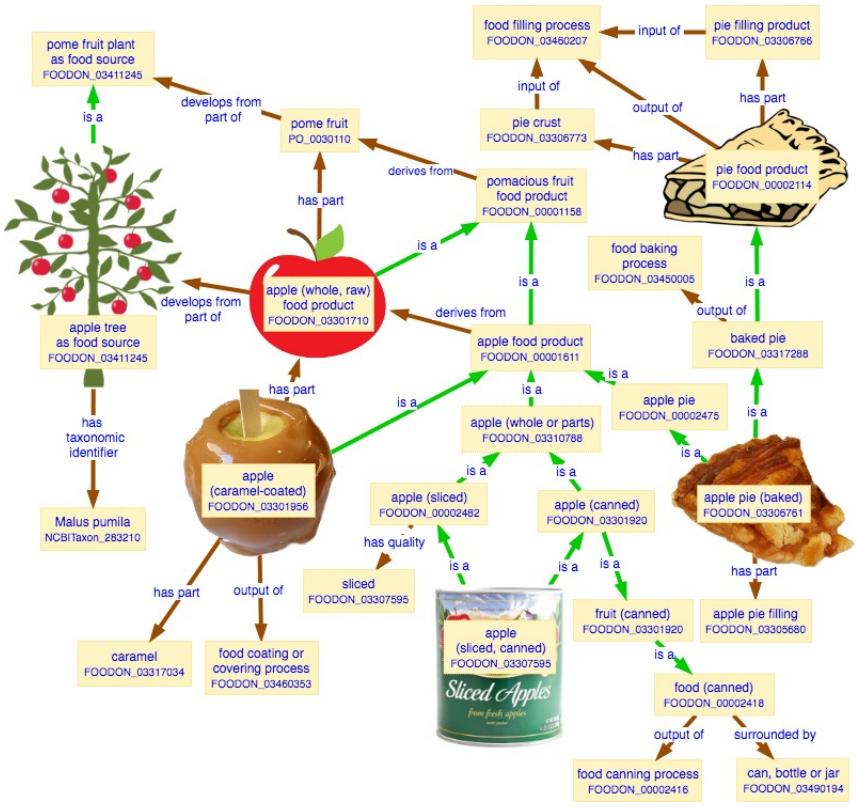
— des ressources qui servent à décrire les régimes alimentaires spécifiques à certaines maladies et l'alimentation qui respecte ces régimes ;

— des ressources qui servent à décrire les recettes de cuisine de façon standardisée.

FoodOn [32] est une ontologie compatible avec BFO (*Basic Formal Ontology*) et issue la construction à partir de Langual depuis 2016. Fin 2018, FoodOn contient 27 051 termes.

Exemple 1.3

Exemple de représentation du concept *pomme* dans FoodOn.



32. https://www.ebi.ac.uk/ols/ontologies/foodon

Dans une certaine mesure, les étiquettes de cette ontologie peuvent correspondre à la polysémie d'usage. Ainsi, « $FOODON : fish\ product$ `derivesFrom` $FOODON : fish\ organism$ ». Par conséquent, nous avons *fish (portion cut)* et *fish (food source)*. FoodOn est un projet très récent, les données sont disponibles en anglais. Les propriétés de type ObjectProperty (relations entre les classes) semblent peu diversifiées : *has ingredient*, *has food substance analog* etc. Il s'agit d'une initiative intéressante.

Food Ontology Knowledge Base (FOKB, Çelik [2015]) développée développée en anglais et en turc contient quatre sous-sections : personne, maladie, produit et ingrédients ou composants alimentaires. Elle est orientée sur la compatibilité plat-régime et exploite les cas d'usage liés au choix des produits alimentaires industriels. FOKB contient 62 classes pour 1 740 individus et 46 propriétés.

MIAM (Desprès [2016]) est une ontologie modulaire pour l'univers de la cuisine numérique (pratique de préparation et de consommation des aliments avec utilisation des appareils intelligents et connectés). Le but applicatif de cette ontologie consiste à permettre l'élaboration automatique de suggestions nutritionnelles permettant à des internautes de s'alimenter de manière équilibrée. Son modèle de connaissance concerne les domaines de la cuisine, de la nutrition, des recettes et du matériel servant à les réaliser. Dans le cadre de ce modèle, une terminologie en français couvrant une bonne partie des termes susceptibles d'être utilisés dans la description d'une recette de cuisine est proposée. Cette terminologie peut servir de support pour l'annotation des recettes. L'enrichissement sémantique d'une recette peut être fait en réalisant d'abord un appariement de sa description textuelle avec la terminologie contenue dans l'ontologie puis en effectuant des inférences sémantiques au niveau du modèle sur les concepts identifiés grâce à la terminologie. **MIAM** est notre ontologie de de référence, elle sera détaillée dans le chapitre 2.

Les ressources qui ciblent les **aliments industriels** représentent un intérêt limité compte tenu de notre démarche. Les ressources ontologiques et lexicales qui s'y consacrent telles que FoodWiki [33] sont relativement abondantes.

Les **recettes de cuisine** telles qu'elles sont représentées à travers le Web [34] ne bénéficient pas encore de structure standardisée malgré les recommandations et l'évolution observée au niveau de la structure `html` des pages Web. BBC Food [35] et `schema.org` fournissent les éléments nécessaires à la représentation structurée et interopérable des recettes de cuisine.

D'autres buts d'ordre normatif ont été poursuivis notamment dans le cadre de construction de l'ontologie **FOOD** (*FOod in Open Data*, Peroni et al. [2016]).

33. `https://bioportal.bioontology.org/ontologies/FOOD_ONTOLOGY`

34. La requête « recette de cuisine » via le moteur de recherche Google renvoie près de 90 millions de résultats en français tandis que la requête « *cooking recipe* » retourne 873 millions de résultats en anglais.

35. `https://www.bbc.co.uk/ontologies/fo`

Cette ontologie a eu comme objectif de structurer la connaissance sur les appellations d'origine protégées en Italie. L'ontologie est structurée autour de 20 catégories de produits soumis à des réglementations spécifiques en Italie.

Outre les ressources terminologiques et ontologiques, il existe de nombreuses ressources lexicales de petite taille pertinentes pour le domaine de la cuisine : lexiques, glossaires, ressources livresques. Certaines de ces ressources sont disponibles en format de données ouvertes comme, par exemple, Kolchin et al. [2015].

1.6 Discussion

Lorsque l'on se penche sur les considérations théoriques qui concernent la construction des ressources de connaissance, les nombreuses ressources existantes et les standards d'interopérabilité, on remarque un certain nombre de contrastes.

Premièrement, le décalage entre la théorie et la pratique de la construction des ressources de connaissance est perceptible. Nous remarquons une distinction très rigoureuse sur le plan terminologique, notamment, en ce qui concerne les deux paradigmes (orientée « logique »" et orientée « calcul ») qui régissent la théorie de la structuration des ressources de connaissance. Cependant, en pratique, de nombreuses ressources s'affranchissent de ces distinctions (en particulier, celles de choix de modèle de type réseau lexical ou ontologie) notamment face à des problématiques de gestion des lexiques multilingues.

Deuxièmement, une certaine forme de répétition en ce qui concerne les solutions proposées peut être remarquée. Les ressources issues de la construction experte telles que WordNet, FrameNet, UNLKB, ainsi que des ressources massives telles que Wikipedia et Wiktionnary sous-tendent ou accompagnent la construction de nombreuses ressources qu'elles soient fondées sur les entités et les faits (YAGO), construites automatiquement à grande échelle (BabelNet) ou ciblées (VerbNet). Cette récurrence est inhérente à la mise en correspondance des ressources notamment grâce aux format de données interopérables. De nombreuses ressources de connaissance générale répercutent involontairement l'asymétrie langagière du Web où de nombreuses langues sont sous-représentées. De ce fait, la construction des ressources sémantiques de connaissance pour ces langues dans le cadre académique ou industriel ne bénéficie pas (ou assez peu) du potentiel des ressources structurées existantes.

Troisièmement, les ressources de spécialité sont peu connectées aux ressources de connaissance générale et souvent monolingues. Formellement, une ontologie construite pour un domaine de spécialité exprime les concepts qui peuvent être désignés ou non par de termes (mots du lexique). L'utilisation des ontologies dans un cadre d'analyse et de classification des textes, de recherche d'information crée la nécessité de "connecter" les ontologies à des lexiques ou des ressources lexicales correspondantes. La frontière entre un mot du lexique utilisé en tant que terme spécialisé (où son sens spécialisé est activé) et son usage général

est avant tout fonctionnelle. Par conséquent, voir les ressources de spécialité comme une projection d'un modèle termino-ontologique donné sur une ressource généraliste qui contient également des connaissances de spécialité issus d'un processus d'acquisition spécifique afin d'y capturer les traces structurées des concepts semble être une piste intéressante.

Conclusion du chapitre

Nous avons exploré les différentes ressources de connaissance qui ont émergé dans le cadre d'approches parfois très diverses. Il est possible de dégager plusieurs grands axes, notamment, quant à la réutilisation des ressources existantes pour la construction et l'exploitation d'une ressource de spécialité multilingue.

Le premier axe serait de se tourner vers les ressources collaboratives, basées sur les données issues de Wikipedia et, plus largement, de Wikidata [36] et notamment sur des extractions de ces données dans des formats interopérables. Étant donné qu'il s'agit des ressources très dynamiques et rendues interopérables, il nous a semblé intéressant de les utiliser dans le cadre d'une démarche de construction d'une ressource multilingue pour un domaine de spécialité. Le deuxième axe serait d'exploiter les ressources structurées existantes. Le troisième axe serait de s'appuyer sur les ressources acquises par externalisation ouverte (peuplonomie).

Ainsi, la construction d'une ressource de spécialité multilingue dédiée à accompagner la construction et l'évolution d'ontologie fait appel à des méthodes variées et peut utiliser les ressources de connaissance générale pour automatiser partiellement l'acquisition des informations et structurer les connaissances du domaine en se basant sur les critères sémantiques riches. Par ailleurs, les ressources non normatives structurées permettent l'appariement des éléments trouvés dans le texte de la recette avec la terminologie contenue dans l'ontologie non plus via l'analyse des formes de surface mais sur la base d'analyse sémantique du texte.

Contributions du chapitre 1

Les contributions du présent chapitre sont les suivantes :

— définir des ensembles de méthodes qui peuvent être utilisées pour la construction d'une ressource lexico-sémantique multilingue ;

— présenter le panorama des ressources existantes, leur spécificités et interconnexions.

36. `https://www.wikidata.org/wiki/Wikidata:Introduction`

Chapitre 2

Construction de la ressource multilingue

Dans le présent chapitre nous allons décrire la méthodologie de création d'une ressource lexico-sémantique multilingue pour le domaine de la cuisine et nutrition. Il s'agira d'une ressource structurée sous forme de graphe orienté, typé et valué. Cette structure de graphe inclura les éléments spécifiques tels que terme, relation, ensemble de types de relations, annotation. La ressource sera développée de manière à être utile pour les tâches courantes de TALN (Traitement automatique des Langues Naturelles) telles que l'analyse sémantique et la conceptualisation. La conceptualisation est ici entendue comme l'identification des processus (séquences d'actions décrites dans les textes enrichies par projection de la ressource multilingue) et objets génériques (ex. ingrédients, ustensiles, lieux, manières).

Organisation du chapitre. Dans un premier temps, nous détaillerons nos deux ressources de référence : le réseau lexico-sémantique RezoJDM (Lafourcade [2007]) et l'ontologie modulaire pour le domaine de la cuisine et de la nutrition MIAM (Desprès [2016]). Dans un second temps, nous décrirons l'architecture de la ressource multilingue proposée. Enfin, nous aborderons les différentes mécanismes de construction de la ressource multilingue.

Termes et notations utilisés dans le chapitre 2

architecture : *(Ressource)* organisation des éléments constitutifs d'une ressource de connaissance en vue d'optimiser l'ensemble de sa conception.

glose : 1.*(RLS)* terme qui sert à "nommer" un sens d'usage (dans le cadre de structuration du lexique autour des raffinements (sens d'usage) des vocables.
2a. *(Lexicographie)* Une définition. 2b. *(Lexicographie)* Un exemple d'utilisation d'une vocable donnée dans un texte.

synset : *(Lexicographie)* ensemble de synonymes ou quasi-synonymes.

2.1 Ressources de référence

La tâche de construction termino-ontologique dotée d'interopérabilité (chapitre 1) implique l'existence d'une ou plusieurs ressources de référence. Dans le cadre de nos expériences, pour permettre l'exploitation des connaissances de sens commun ainsi que des connaissances linguistiques dans le cadre de la construction termino-ontologique, nous disposons de deux ressources de référence : le réseau lexico-sémantique du français RezoJDM issu du projet JeuxDeMots construit au sein du laboratoire LIRMM (Lafourcade [2007]) et l'ontologie modulaire pour le domaine de la cuisine et de la nutrition MIAM (Desprès [2014]) construite au sein du laboratoire LIMICS. Ces ressources sont en français.

RezoJDM est un réseau lexico-sémantique construit par peuplonomie via le jeu JeuxDeMots [1] et les jeux annexes [2] depuis 2007. Cette ressource est un graphe orienté, typé et pondéré. À ce jour, RezoJDM contient 2,7 millions de termes représentés sous forme de nœuds du graphe et 240 millions de relations (arcs).

Le parti pris lors de la conception de RezoJDM a été celui de maintenir toutes les informations au sein de la ressource y compris les redondances (cycles) et les informations dites "négatives" ou reconnues "fausses" par les utilisateurs sous forme de relations inhibitrices à poids nul ou négatif. 880 000 relations inhibitrices sont présentes dans le RezoJDM à l'heure où nous écrivons.

Au sein du RezoJDM, la polysémie est modélisée grâce aux raffinements. Le raffinement sémantique correspond à la modélisation de *sens d'usage* soit la projection d'une *acception* sur un *contexte* particulier. Au sens classique de la lexicographie, une acception correspond à un "*sens variable, nuance sémantique d'un mot suivant ses conditions d'emploi ou d'interprétation*" [3].

Tchechmedjiev [2016] propose une formalisation mathématique complète des acceptions et définit la relation de raffinement comme une relation binaire entre deux classes d'équivalence de sens. D'après Davey and Priestley [1990], la relation de raffinement est une "relation d'ordre partiel strict". En effet, le raffinement d'un terme spécifie le sens de ce terme dans un contexte spécifique donc va du général au spécifique. Le raffinement d'un terme peut être raffiné à son tour et s'inscrit alors dans une structure arborescente, une hiérarchie des raffinements ou hiérarchie des acceptions s'il est question uniquement de l'usage général. Il s'agit d'une relation transitive.

Contrairement à la majorité des ressources de connaissance générale sous forme de graphe qui contiennent un ensemble de types de relations limité, RezoJDM

1. `http://www.jeuxdemots.org`
2. `http://imaginat.name/JDM/Page_Liens_JDMv2.html`
3. `http://www.cnrtl.fr/definition/acception`

contient plus de 100 types de relations [4]. L'état de l'art en ce qui concerne les types de relations disponibles permet de constater une diversité moindre : en dehors de RezoJDM, la ressource qui contient le plus de types de relations sémantiques prédéfinies est ConceptNetSpeer and Havasi [2012] avec environ 25 types de relations.

Les relations d'un RLS tel que RezoJDM représentent explicitement des fonctions lexicales (exemple 2.1). Leurs extrémités sont des entités lexicales.

Exemple 2.1

Parmi les fonctions lexicales (FL) [a] peuvent être distinguées les fonctions syntagmatiques (modélisation d'un lien collocationnel) et paradigmatiques (modélisation d'une dérivation sémantique). Ainsi, **Syn** est la fonction qui permet de modéliser les synonymes ou les quasi synonymes d'une acception (lexie).

$$\textbf{Syn}(\textit{voiture}) = \textit{automobile}, \textbf{fam}\,\textit{auto}, \textbf{fam}\,\textit{bagnole}, \textbf{fam}\,\textit{caisse}$$

La FL **Anti** permet de modéliser les antonymes d'une acception : **Anti**(*petit*) = *grand*.

La FL **Magn** est un exemple d'une FL syntagmatique qui associe à une acception un ensemble d'acceptions ou d'expressions linguistiques qui expriment une intensification :

$$\textbf{Magn}(\textit{chagrin}) = \textit{grand}, \textit{gros>énorme}, \textit{immense}$$

Les FL telles que $\textbf{Oper}_\textbf{i}$, $\textbf{Func}_\textbf{i}$ et $\textbf{Labor}_\textbf{ij}$ dont les noms viennent respectivement du latin *operari* (« faire »), *functionare* (« travailler ») et *laborare* (« travailler ») permettent de décrire les constructions à verbe support.

$$\textbf{Oper}(\textit{coup}) = \textit{asséner}, \textit{donner}, \textbf{fam}\textit{flanquer}.$$
$$[\textit{asséner/donner/flanquer un coup} \equiv \textit{frapper}]$$

Les FL $\textbf{Real}_\textbf{i}$, $\textbf{Fact}_\textbf{i}$ et $\textbf{Labreal}_\textbf{ij}$ permettent d'exprimer les verbes sémantiquement pleins (par exemple, les verbes de « réalisation », « satisfaction ») qui contrôlent les mêmes structures syntaxiques qui les FL de verbes supports.

La figure 2.1 présente de nombreux exemples de la FL $\textbf{S}_\textbf{loc}$ qui modélise les dérivés nominaux circonstantiels (« endroit typique » où se déroule la chose dénommée par la lexie ou l'action concernant cette lexie peut se passer.

$$\textbf{S}_\textbf{loc}(\textit{loup}) = \textit{forêt}, \textit{conte}, \textit{bergerie}$$
$$\textbf{S}_\textbf{loc}(\textit{sel}) = \textit{salière}, \textit{aile de poulet}$$

a. Exemples inspirés et cités d'après Polguère [2002] et Jousse [2007].

Les *fonctions lexicales* ont été formulées dans le cadre de la *théorie sens-texte* (TST, Jolkovsky and Mel'čuk [1967]), la valeur de l'application de la fonction lexicale f_{lex} à l'item I est un ensemble d'items lexicaux $\{I_1, I_2, ...I_n\}$. Chaque

4. `http://www.jeuxdemots.org/jdm-about-detail-relations.php`

item de la valeur de la fonction lexicale $f_{lex}(I)$ remplit à peu près le même rôle vis-à-vis l'item I. Ainsi, il s'agit d'une fonction multivoque et approximative. Par conséquent, *elle ne peut pas être considérée comme une fonction mathématique et fait partie uniquement de la théorie linguistique*. Cependant, les formalismes fondés sur les fonctions lexicales permettent de décrire les entrées lexicales de façon suffisamment rigoureuse non seulement pour l'usage humain, mais aussi pour un usage machine.

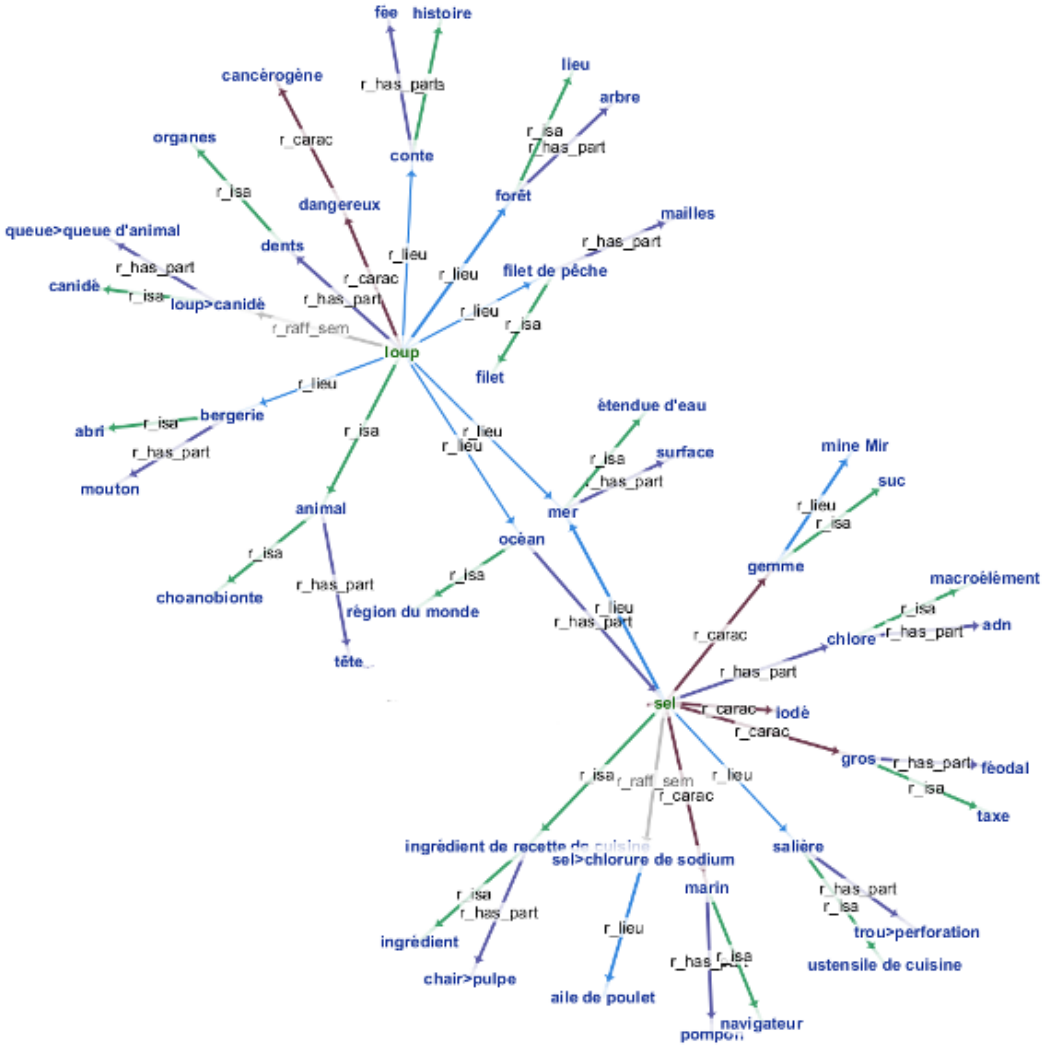


FIGURE 2.1 – Représentation des termes *sel* et *loup* au sein du RezoJDM. On remarque que ces deux termes sont polysémiques. On note aussi la richesse de l'ensemble des types de relations qui participent à expliciter la sémantique des termes.

MIAM (figure 2.3) est une termino-ontologie modulaire pour l'univers de la cuisine numérique construite par une communauté d'experts en suivant la méthodologie collaborative de construction ontologique NeOn [5]. Son but applicatif est de fournir la connaissance nécessaire à l'élaboration de suggestions nutritionnelles généralistes. Ainsi, elle contient une partie importante des termes pouvant être utilisés pour l'annotation des recettes de cuisine.

5. `www.neon-project.org`

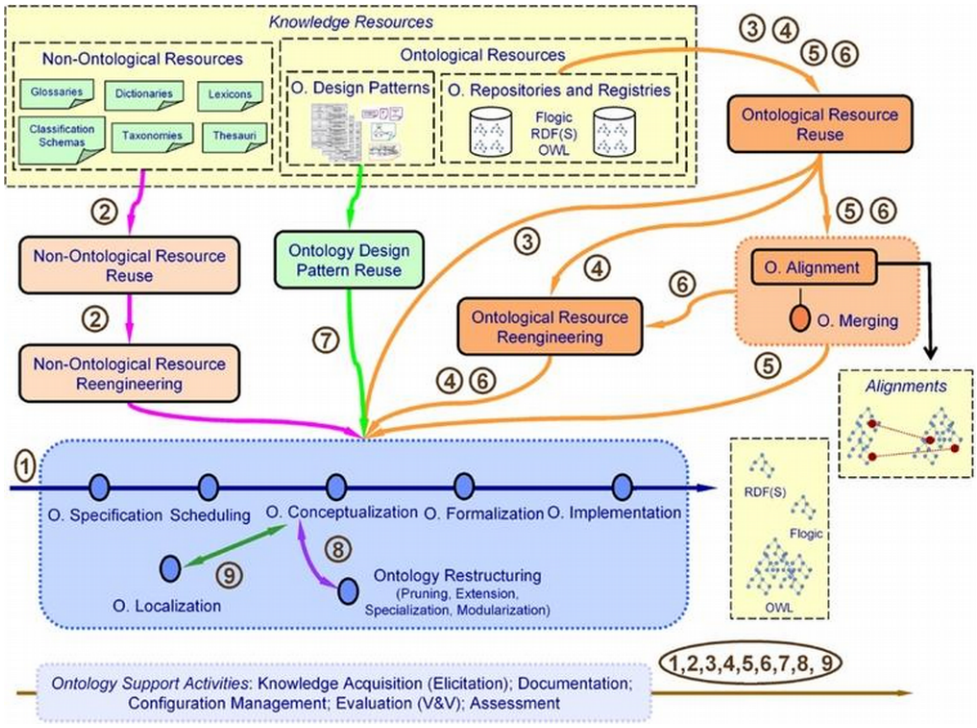


FIGURE 2.2 – Méthodologie de construction d'ontologie NeOn.

La méthodologie NeOn (figure 2.2) utilisée pour la construction de MIAM est axée sur la réutilisation des ressources existantes et leur intégration dans le processus de construction d'ontologie. Plusieurs scénarios ont été définis dans ce cadre :

1. *spécification → implémentation* : développement sans réutilisation des ressources existantes ;
2. *réutilisation des ressources non ontologiques*[6]. Les ressources de ce types doivent être intégrés dans l'ontologie ;
3. *réutilisation des ressources ontologiques*. Dans le cadre de ce scénario, les ressources ontologiques sont prises en tant que tout, sans les intégrer dans l'ontologie en cours de construction. Il s'agit d'une mise en réseau des ontologies ;
4. réutilisation et ré-ingénierie des ressources ontologiques existantes ;
5. *réutilisation et fusion des ressources ontologiques*. Ce scénario concerne le cas où plusieurs ressources ontologiques sont disponibles pour un même domaine d'activité ;
6. *réutilisation, fusion et ré-ingénierie des ressources ontologiques* existantes, ce scénario est similaire au précédent mais la décision peut être prise de restructurer l'ensemble ;
7. *réutilisation des patrons de conception d'ontologie*[7] (*ODPs*, *Ontology Design Patterns*) ;

6. Non-ontological resources (NORs).
7. `http://ontologydesignpatterns.org/wiki/Main_Page`

8. *restructuration des ressources ontologiques* (réduire, étendre, spécialiser une ressource existante ;
9. *localisation* des ressources ontologiques (adaptation à d'autres langues et cultures).

MIAM a été construite selon une combinaison de scénarios. La participation des experts du domaine de la cuisine et de la nutrition a joué un rôle important dans la construction de MIAM (conformément au premier scénario spécifié par la méthodologie NeOn). Parallèlement, des ressources non ontologiques ont été réutilisées (LanguaL, ressources terminologiques, glossaires, lexiques variés).

Aliment est le modèle clé de l'ontologie MIAM (figure 2.3) car l'annotation des recettes en termes de MIAM nécessite avant tout de modéliser la composition, les caractéristiques sensorielles des aliments.

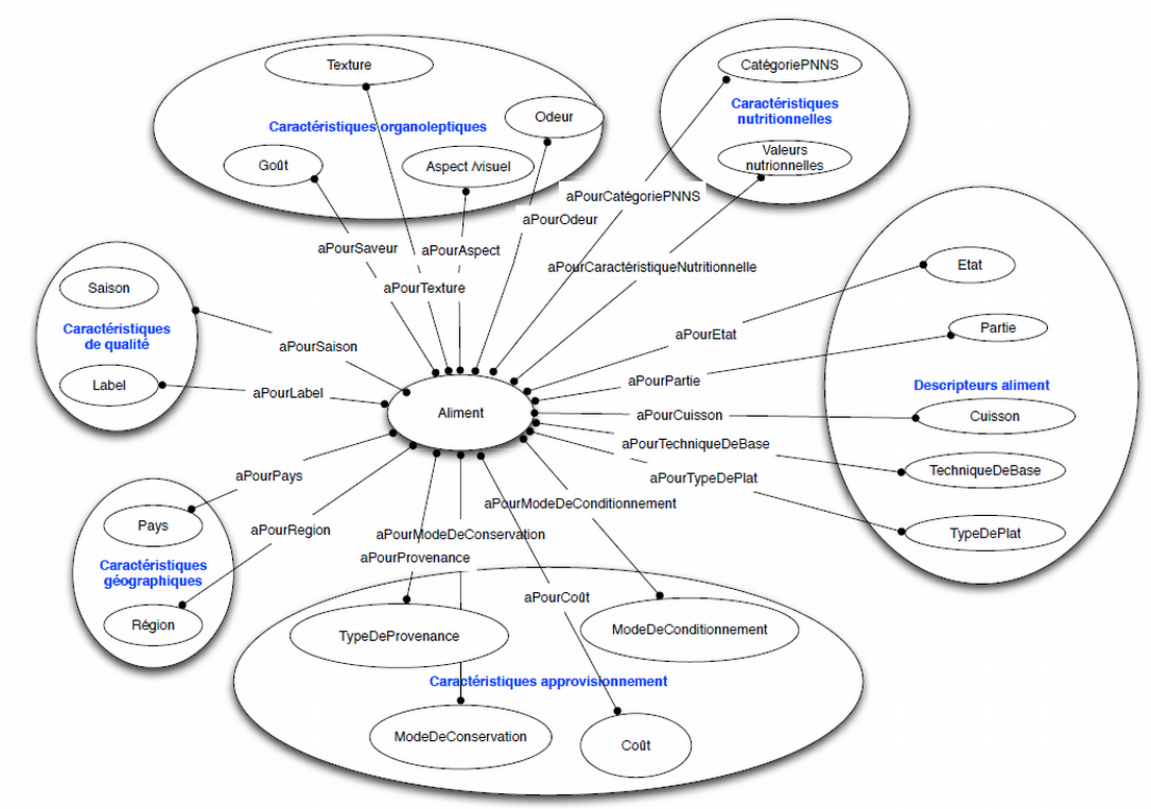


FIGURE 2.3 – Modèle de l'*Aliment* , ontologie MIAM.

Le modèle de connaissance de l'ontologie regroupe les connaissances expertes sur les aliments, les transformations, les actions culinaires, les plats représentatifs de la tradition culinaire française, les recettes utilisées pour réaliser ces plats.

Les connaissances contenues dans l'ontologie MIAM sont exprimées sous forme de hiérarchie des concepts (classes d'ontologie) et des individus (qui instancient les concepts d'ontologie). Outre les relations hiérarchiques (en particulier, les relations qui expriment l'héritage telles que, par exemple, `subClassOf`), l'ontologie MIAM contient des relations qui se déclinent en *Object Properties* (propriétés

à valeur objet) qui sont des relations transversales entre les concepts) et *Data Properties* (propriétés à valeur donnée, valeurs littérales associées aux concepts).

Compte tenu de la spécificité de MIAM et du cadre méthodologique qui sous-tend la construction des ressources ontologiques actuelles, mais aussi de la pertinence incontestable des ressources de type "réseau lexico-sémantique" face aux différentes problématiques d'analyse des textes en langue naturelle et d'autres applications TAL [8], nous avons délimité le cadre de nos expériences comme suit :

1. Notre démarche consiste à définir et à construire une ressource de connaissance multilingue sous forme de réseau lexico-sémantique qui permettrait d'accompagner la construction termino-ontologique multilingue et qui serait compatible avec le modèle de connaissances de l'ontologie de référence MIAM ;
2. en conformité avec le scénario 2 de la méthodologie NeOn, nous utilisons une ressource non ontologique pour appuyer la construction ontologique ;
3. en divergence avec la méthodologie NeOn, au lieu d'inclure les informations issues de la ressource de connaissance multilingue que nous construisons dans l'ontologie, nous procédons par immersion d'ontologie dans une ressource structurée non ontologique. Cette immersion permet de projeter le modèle d'ontologie sur une ressource non ontologique structurée et ainsi découvrir et capturer les structures lexico-sémantiques qui peuvent contribuer à la construction, ré-ingénierie et localisation d'ontologie.

La divergence que nous évoquons est inévitable car le cadre méthodologique NeOn sous-entend l'utilisation des ressources peu structurées et peu expressives. Or, actuellement, de nombreuses ressources de connaissance structurées sont disponibles et exprimées en utilisant les formats interopérables (par exemple, le format OntoLex, chapitre 1).

Mondary [2011] distingue les "trois plans de conceptualisation" qui sont le *corpus*, le *modèle linguistique* au sens large et le *modèle du domaine*. Dans notre approche à la construction d'une ressource langagière multilingue, nous nous rapprochons des cette distinction en amorçant notre ressource grâce aux extractions depuis le corpus spécialisé, en choisissant l'architecture d'un réseau lexico-sémantique (représentation de la connaissance telle qu'elle est reflétée dans une langue naturelle) et en visant les tâches d'enrichissement d'ontologie et de conceptualisation à partir d'une ressource non ontologique.

Tout au long de nos expériences nous nous focaliserons sur le modèle *Aliment* et sur le module de MIAM SensoMIAM qui est consacré aux caractéristiques sensorielles des aliments.

8. De nombreux travaux qui ont concerné RezoJDM, BabelNet, ConceptNet témoignent de l'intérêt que ce type de ressources représente notamment dans le cadre de la désambiguïsation des lexiques, de l'analyse sémantique, de la conceptualisation etc.

2.2 Architecture du RLSM$_{\mathbf{PI}}$

Malgré le cadre applicatif de nos expériences et notre démarche de construction spécifique au domaine de cuisine et nutrition, son architecture se veut indépendante du domaine et de la langue. L'architecture proposée est celle d'un Réseau Lexico-Sémantique Multilingue avec Pivot Interlingue (RLSM$_{\text{PI}}$). Il s'agit d'une ressource sous forme de graphe qui comporte un certain nombre de particularités de modèle et de structure.

2.2.1 RLSM$_{\mathbf{PI}}$ en tant que graphe

Le modèle de représentation des connaissances du RLSM$_{\text{PI}}$ s'inspire de celui du réseau lexico-sémantique RezoJDM, Lafourcade [2011]. Cependant, des changement y ont été apportés compte tenu du caractère multilingue de la ressource à mettre en œuvre. Il s'agit d'un graphe *orienté*, *typé* et *valué* (ses arcs sont caractérisées par des poids et des annotations). Ce graphe comporte k sous-graphes correspondant à chacune des k langues couvertes par la ressource et un sous graphe spécifique qui remplit la fonction de *pivot interlingue*.

Définition 2.1

En tant que graphe, RLSM$_{\text{PI}}$ comporte un ensemble de termes T et un ensemble de relations R. $RLSM_{\text{PI}} = \{T, R\}$. Une relation $r \in R$ est un sextuplet

$$r =< s, t, type, v, l_s, l_t >$$

où

- s et t correspondent respectivement à la source et cible de la relation ;
- *type* correspond à son type (ex. *r_ isa, r_ carac, r_ object*) ;
- v correspond à une valuation (méta-information) associée à la relation. Il peut s'agir de poids, d'une annotation ;
- l_s et l_t correspondent aux labels qui explicitent l'appartenance des termes source et (respectivement) cible à un des sous-graphes de la ressource.

L'ensemble des termes du RLSM$_{\text{PI}}$ comporte des partitions qui correspondent aux différentes langues. Les éléments de ces partitions sont reliés uniquement via le pivot interlingue.

$$\texttt{en:flavor} \xleftarrow{r\_covers} \texttt{en:flavor in:flavor} \xrightarrow{r\_covers} \texttt{fr:flaveur}$$

En termes de leur *arité*, les relations présentes dans le RLSM$_{\text{PI}}$ sont principalement des relations binaires. Une relation binaire $r_{type}(s, t)$ soit $(s, t) \in r_{type}$ peut être :

— **symétrique** (cas d'une relation de synonymie, traduction, antonymie : *viande de veau* $\xrightarrow{r\_syn}$ *veau* > *viande*) ou **anti-symétrique** (cas d'une relation de raffinement : *veau* $\xrightarrow{r\_refinement}$ *veau* > *viande*) ;

— **transitive** (hyperonymie et hyponymie : *veau* > *viande* $\xrightarrow{r\_isa}$ *viande blanche* $\wedge$ *viande blanche* $\xrightarrow{r\_isa}$ *viande* $\Rightarrow$ *veau* > *viande* $\xrightarrow{r\_isa}$ *viande* ou **non-transitive** (cas de *salade verte* $\xrightarrow{r\_location}$ *assiette* $\wedge$ *assiette* $\xrightarrow{r\_location}$ *lave − vaisselle*) ;

— **réflexive** (cas d'une relation de type *r_lemma* : *veau* $\xrightarrow{r\_lemma}$ *veau*).

En tant que ressource destinée à être parcourue par les algorithmes de recherche d'information et d'inférence, RLSM$_{\text{PI}}$ est un graphe dont le diamètre est très réduit (graphe dit ***petit monde***) qui peut être exploré en faisant soit l'hypothèse de *monde ouvert* soit l'hypothèse de *monde fermé*. Ces deux hypothèses telles qu'elles sont entendues dans le présent manuscrit portent sur la façon d'appréhender le caractère incomplet des ressources langagières à savoir la présence de *silences* (absence d'informations pertinentes) et de *bruit* (présence d'informations non pertinentes ou fausses). En faisant l'hypothèse de *monde ouvert*, nous considérons qu'une *information absente est vraie*, l'inférence peut inclure la recherche d'information à l'extérieur du réseau. Dans l'hypothèse du *monde fermé* (par exemple, lors de l'analyse sémantique), le réseau « se ferme », il est considéré comme « complet » et « suffisant », les *silences sont considérés comme « faux »*, l'inférence est réalisée à partir des éléments déjà présents dans le graphe au moment t.

2.2.2 RLSM$_{\text{PI}}$ en tant que ressource multilingue

Les problématiques de construction du RLSM$_{\text{PI}}$ se rapprochent des problématiques d'alignement entre plusieurs ressources lexicales en langues différentes. En effet, compte tenu de la quantité importante des ressources pré-existantes, la ressource multilingue est amenée à en intégrer un certain nombre. Par conséquent, le choix de l'architecture pour la construction d'une ressource lexico-sémantique multilingue de spécialité repose sur le choix qui pourrait être fait pour l'alignement des ressources lexicales susceptibles d'être intégrées partiellement ou totalement dans celle-ci. Il s'agit de l'une des deux approches suivantes :

— *architecture par transfert*. Ce type d'architecture résulte de l'alignement des ressources deux par deux, les liens de correspondance sont ainsi des liens de correspondance directs ;

— *architecture par pivot*, c'est-à-dire une architecture qui va "mettre en relation les éléments équivalents venant des différentes ressources à aligner" [9]. « Un pivot est une généralisation sur les représentations sémantiques des éléments alignés issus des différentes ressources » (Witt et al. [2009]).

9. Tchechmedjiev [2016], op.cit. p.22

Tchechmedjiev [2016] fournit une analyse détaillée de ces deux familles d'approches à l'alignement des ressources.

Le choix a été fait en faveur d'une architecture avec pivot interlingue afin de se doter d'une structure qui, à terme, permettra de contourner le problème de contraste en termes de granularité (existence ou absence de certains sens).

Pour éviter les difficultés inhérentes à la construction d'un pivot artificiel dont notamment la nécessité aligner N sens simultanément, le pivot interlingue du RLSM$_{PI}$ est amorcé à la manière d'un pivot naturel (en utilisant l'édition anglais de DBNary comme ensemble de données d'amorçage). Le pivot évolue ensuite vers un pivot interlingue de façon incrémentale. Certains sens présents dans le pivot se retrouvent fusionnés, d'autres émergent à partir des structures observées dans les partitions de la ressource.

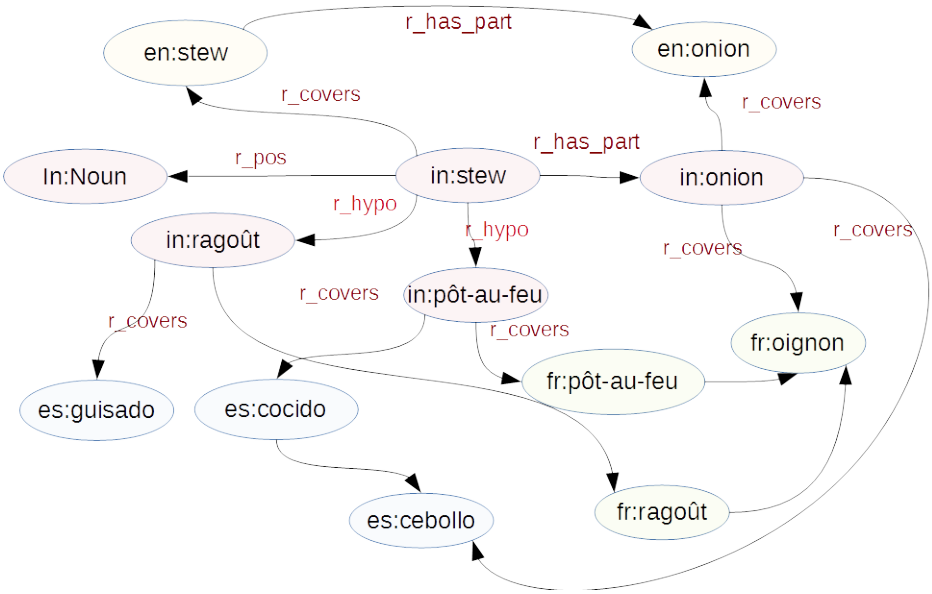


FIGURE 2.4 – Architecture du RLSM$_{PI}$. Les préfixes *in*, *en*, *fr*, *es*, *ru* dénotent l'appartenance des termes à ses différents sous-graphes.

Exemple 2.2

En russe et en anglais, il existe des lexicalisations différentes pour exprimer le concept d'un animal et de la chair comestible de cet animal.

porc (animal) : **pig** en anglais et **свинья** en russe
porc (chair) **pork** en anglais et **свинина** en russe

Si l'on choisit les concepts et les sens lexicalisés du français comme pivot naturel, on devrait utiliser **porc** pour relier ces termes ce qui aboutirait a des propositions fausses dans le cadre de la tâche de traduction.

Viser le pivot interlingue permet de réduire progressivement le *phénomène contras-*

tif artificiel défini par Sérasset [2012] comme « une perte d'information discriminatoire liée à une conceptualisation et lexicalisation qui divergent entre les langues » propre à l'utilisation d'un pivot naturel (exemple 2.2). L'utilisation de la ressource, notamment dans le cadre d'aide à la construction terminoontologique favorise également le choix du pivot artificiel interlingue.

2.3 Construction du $RLSM_{PI}$

Dans le cadre de la démarche de construction du $RLSM_{PI}$, les données structurées issues des extractions depuis les corpus de textes en langue naturelle de même que les ressources terminologiques serviront à amorcer la ressource et à guider sa construction. Les données structurées issues des ressources pré-existantes telles que RezoJDM [10] (Lafourcade [2011]), DBNary [11] (Sérasset [2014]), WordNet (Fellbaum [1998]), ConceptNet [12] (Speer and Havasi [2012]), RWN [13] (Loukachevitch [2016]) seront intégrées dans la ressource notamment dans l'optique d'interopérabilité. L'alignement entre les termes des différentes langues sera basé sur un dictionnaire (DBNary, terminologies bilingues), mais aussi sur la comparaison des structures sémantiques observée dans les différents sous-graphes du réseau.

Dans le choix de notre ***stratégie d'amorçage*** du $RLSM_{PI}$, nous avons considéré l'hypothèse de Ramadier [2016] sur la non séparation entre les différents types de connaissance, notamment, la connaissance générale et la connaissance de spécialité. Dans le cadre de cette hypothèse, nous avons fait le choix d'intégrer la connaissance générale disponible dans les ressources existantes choisies pour intégration en utilisant comme points d'ancrage les termes et les relations présents dans les corpus de spécialité, principalement des corpus de recettes de cuisine. Dans un contexte industriel, en présence de listes de termes et de nombreux document peu ou pas structurés, une telle stratégie semble permettre d'optimiser l'amorçage.

2.3.1 Remarques préliminaires

Avant de détailler l'extraction des termes, il est important de souligner que dans le cadre de nos travaux d'extraction et d'amorçage du $RLSM_{PI}$, il ne s'agit pas des termes correspondant à la définition « désignation verbale d'un concept général dans un domaine spécifique » propre à la discipline de terminologie. Nous traitons uniquement des termes tels que *items lexicaux* qui peuvent exprimer une très grande variété d'objets. Il s'agit des usages des mots qui ne correspondraient qu'en partie aux entrées d'un dictionnaire.

10. `http://www.jeuxdemots.org/jdm-about.php`
11. `http://kaiko.getalp.org/about-dbnary/`
12. `http://conceptnet.io/`
13. `https://ruwordnet.ru/en/`

Les **termes multi-mot** (TMM) et les relations sémantiques sont au centre de notre démarche d'extraction depuis les corpus d'amorçage. En effet la valeur ajoutée du $RLSM_{PI}$ est de répertorier les usages et d'expliciter les relations sémantiques qui existent entre ces usages modélisés sous forme de termes du réseau de façon la plus riche possible.

Comme il sera détaillé dans la suite du présent manuscrit, le raisonnement qui utilise $RLSM_{PI}$ repose sur un ensemble de mécanismes d'inférence. Il a été observé dans le cadre de notre ressource de référence RezoJDM que t_r, le temps d'exécution d'une requête pour un ensemble de données D est de l'ordre de $t_r = log_2(|D|)$. Par conséquent, à chaque fois que $|D|$ double, le temps d'exécution d'une requête augmente de 1. t_r correspond au temps d'exécution d'une requête lorsque l'information est directement accessible via une requête unique. Dans le cas contraire, un mécanisme d'inférence est nécessaire. Lors de l'utilisation des règles d'inférence, le nombre moyen des requêtes nécessaire se situe entre 10 et 20 requêtes par règle. Compte tenu de l'augmentation de temps d'exécution des requêtes constatée avec RezoJDM, pour un nombre de requêtes n_r défini sur $[10;\ 20]$, la croissance de la taille de données idéalement nécessaire pour le calcul d'inférence $|D_{perf}| = 2^{n_r}$. Donc, dans le cadre de ce mode d'estimation du temps nécessaire pour un ensemble de requêtes d'inférence, ajouter une requête nécessite de doubler la taille de données $|D_{perf}|$. Pour n_r tel que nous l'avons défini, $|D_{perf}|$ se définirait sur l'intervalle $[1024, 1048576]$. Par conséquent, il est pertinent d'expliciter au maximum les données [14] pour un fonctionnement optimal des mécanismes d'inférence et de raisonnement.

Malgré cette différence de méthodes de conception des ressources de connaissance, il est aisé de faire le rapprochement entre les méthodes que nous définissons dans le présent chapitre et les ensembles d'outils comme *Terminae* (Szulman [2012]). Ce dernier permet de passer des textes à l'ontologie en utilisant les extractions terminologiques et autres outils de TAL pour constituer un ensemble d'éléments textuels (termes et relations) organisés sous forme d'un réseau qui peut être utilisé dans le cadre de construction ontologique experte.

2.3.2 A propos de l'intégration des ressources existantes guidée par le corpus de spécialité

Lorsqu'il s'agit de construire un corpus depuis les données récoltées sur le Web, il n'est plus question de traitement de donnés textuelles non structurées, mais de données semi-structurées. Sa principale subtilité est liée à la nécessaire com-

14. Rendre l'accès aux informations le plus direct : l'ensemble des relations sémantiques est directement attaché à chaque terme (par opposition aux ressources termino-ontologiques dont la construction exclut les redondances pour les raisons de consistance logique notamment grâce à l'héritage des propriétés).

binaison de critères *structurels* [15] et *textuels* [16].

Dans le cadre de notre approche, les propriétés structurelles sont principalement exploitées lors de la récolte des données sur le Web en suivant une procédure classique de parcours du graphe en largeur. Les éléments textuels sont identifiés grâce à un ensemble de sélecteurs [17]. La distinction faite par Dong et al. [2014] entre l'exploitation de l'arbre HTML par opposition aux tableaux HTML [18] pour l'extraction de certaines relations (par exemple, méronymie, hyperonymie) est également pertinente dans notre cas. Les données textuelles récoltées ont été stockées au format `XML` afin de permettre des traitements spécifiques concernant les sous-ensembles de données : titres de recettes, listes d'ingrédients, instructions de préparation, définitions [19].

Bouamor [2014] donne la distinction entre les corpus parallèles, comparables et indépendants [20].

— ***corpus parallèles*** : paires de documents en relation de traduction. D'après Fung and Yee [1998], un corpus parallèle doit posséder les caractéristiques suivantes : mots monosémiques, traduction unique associés à chaque mot, pas de traduction manquante, positions et fréquences des mots en relation de traduction sont comparables. Le caractère monosémique et la condition de traduction unique ne sont généralement pas satisfaits. Les corpus de ce premier type sont rares en ce qui concerne le domaine de la cuisine et nutrition.

— ***corpus comparables*** : corpus qui rassemblent des documents multilingues qui partagent des traits communs (domaine, discours, période etc). D'après Déjean et al. [2002],

> « *Deux corpus de deux langues l1 et l2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue l1, respectivement l2, dont la traduction se trouve dans le corpus de langue l2, respectivement l1.* »

D'après Bowker and Pearson [2002],

15. Les critères structurels sont des critères qui relèvent du balisage du document (notamment de l'organisation et la sémantique des balises XML, HTML et autres langages de description issus du protocole SGML).

16. Les critères textuels sont des critères linguistiques (syntaxe, sémantique)

17. Nous avons utilisé les sélecteurs CSS. Ces sélecteurs sont des patrons qui portent sur l'arbre **html** de la page Web en cours d'exploration.

18. D'après Dong et al. [2014], à travers le Web, il existe 570 millions de tables qui contiennent l'information relationnelle (par opposition à l'utilisation des tables HTML pour le formatage). Selon ces auteurs, les techniques d'extraction d'informations adaptées aux arbres et au contenu textuel ne fonctionnement pas pour l'extraction des relations à partir des tables HTML.

19. A titre accessoire car, à la différence des approches qui implémentent les différentes variantes de l'algorithme *Lesk* (Lesk [1986]), notre approche ne se focalise pas sur les définitions et les calculs de proximité sémantique qui peuvent être faits à partir des définitions des dictionnaires appelées également les "gloses".

20. *Op.cit.* p. 8.

> *« Un corpus comparable est composé d'ensembles de textes, dans des langues différentes, qui ne sont pas des traductions mutuelles. »*

— ***corpus indépendants*** : documents traitant des sujets similaires.

Dans le cadre de nos expériences, il s'agit presque exclusivement des **corpus comparables** et, dans la moindre mesure, des **corpus indépendants**. Compte tenu de la spécificité des textes de cuisine et de leur ancrage culturel, les corpus parallèles sont extrêmement rares pour ce domaine [21]. Au lieu d'être un obstacle à notre travail, l'usage de corpus comparables est une opportunité. En effet, d'après, Johansson [2007], « les corpus comparables permettent de tirer les conclusions sures sur les similarités et les différences entre les langues » [22].

— Ils révèlent des phénomènes propres à une langue, difficiles à remarquer dans un cadre monolingue (c'est-à-dire, des phénomènes relevés dans une langue qui n'apparaissent pas dans une autre).

— Ils mettent en évidence des différences entre langues (d'un point de vue syntaxique, typologique ou culturel) mais soulignent aussi des phénomènes universels.

— Ils éclairent les différences entre textes traduits et textes sources, mais aussi entre textes « natifs » d'une langue et textes traduits.

Des travaux de recherche sont activement menés autour des méthodes d'extraction des lexiques bilingues à partir des corpus comparables depuis les années 1990. Dans leur grande majorité, ils sont issus de la sémantique distributionnelle basée sur l'hypothèse de Harris. Les approches dites « standard » ou « par traduction directe » modélisent le voisinage du terme donné au niveau du segment textuel où il apparaît sous forme d'un *vecteur de contexte*. Ce vecteur est un vecteur lexicalisé et peut parfois correspondre à un ensemble partiellement ordonné de couples (terme, score). Certains auteurs associent l'information lexico-sémantique et information fréquentielle liées au termes de la langue source et/ou cible) élargi des termes à extraire et intègrent la distribution des mots non plus uniquement dans le contexte local [23] mais au niveau du *modèle de la langue*.

Les travaux existant sur l'extraction des lexiques bilingues à partir des corpus comparables diffèrent :

— par **la nature et l'origine des corpus comparables**. Wikipedia apparaît comme corpus fréquemment utilisé pour collecter des corpus comparables pour un domaine de spécialité. Une liste de mots prédéfinie, les liens et/ou les catégories Wikipedia servent de point de départ pour la construction des corpus. Les projets européens et les corpus de textes qui y sont

21. A titre d'exemple, la taille du corpus parallèle français-anglais obtenu à partir de la ressource en ligne *CuisineAZ* (`http://cuisineaz.com` est très faible (811 documents).

22. Op. cit., p.10

23. Comme remarqué par Bouamor [2014], dans le cadre d'extraction des lexiques depuis les corpus comparables,« *il n'est plus possible de s'appuyer sur la distribution des mots dans le document* », *Op.cit.*, p.75).

associés constituent une autre source fréquemment citée pour l'acquisition des corpus comparables pour les langues européennes ;

— par **la manière de construire le pont de traduction** entre langues source et cible choisies pour l'extraction des lexiques bilingues. Un dictionnaire bilingue dit « dictionnaire plat » est fréquemment utilisé. Cette pratique crée la nécessité de désambiguïser les traductions disponibles dans ce dictionnaire qui sert à traduire les vecteurs de contexte afin de les comparer ;

— par **la manière de construire le modèle de la langue**, notamment, par le type de ressources utilisées pour cette modélisation (Web en tant que corpus, des ressources de connaissance structurées telles que WordNet Fellbaum [1998], espace conceptuel ;

— par **la manière de caractériser le contexte afin de sélectionner les termes candidats appropriés**.

La caractérisation des contextes via l'approche standard enrichie grâce à la désambiguïsation sémantique est une approche très prometteuse qui a été proposée par Bouamor [2014]. Le déroulement de cette méthode est présenté sur la figure 2.5. Le même auteur propose une autre méthode de désambiguïsation basée sur

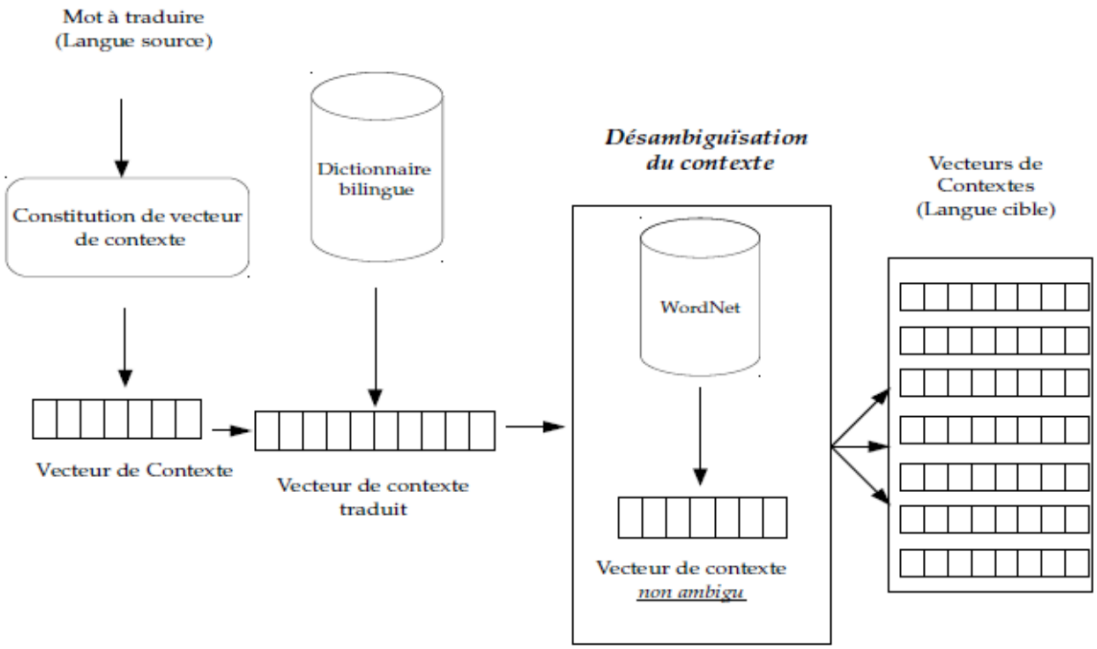


FIGURE 2.5 – Approche à l'extraction des lexiques bilingues à partir des corpus comparables avec désambiguïsation sémantique, image citée d'après Bouamor [2014] *op.cit* p.92.

l'exploitation des titres des articles Wikipedia.

> *« au lieu de considérer l'espace des mots d'un corpus pour la représentation des mots que l'on souhaite traduire, ces derniers sont représentés dans l'espace des titres des articles de Wikipedia »*

À l'instar des autres approches basées sur la construction des vecteurs de contexte, cette approche est concernée par les difficultés inhérentes à ce modèle de structuration, à savoir :

1. l'élément textuel (le contexte du terme donné) reste associé à un espace vectoriel de dimension finie [24] qui dépend de la taille de fenêtre contextuelle choisie ;
2. la production des vecteurs candidats avant désambiguïsation repose sur l'utilisation d'un dictionnaire bilingue. Par conséquent, la qualité de ces vecteurs dépend de la couverture du dictionnaire choisi et, dans le cas d'un dictionnaire classique [25] peut manquer de refléter certains usages ;
3. tous les termes du vecteur de contexte doivent être traduits. Par conséquent, un contraste artificiel peut survenir lors de la construction des vecteurs de contexte candidats ;
4. étant donné que les termes qui constituent un vecteur de contexte doivent vérifier une dépendance syntaxique ou être dans un rapport d'association ou de vraisemblance avec le terme à modéliser, un tel vecteur peut offrir une représentation aplatie du contexte. En effet, dans le modèle vectoriel le typage des relations est global. À savoir, tous les termes entretiennent une relation d'une seule et même nature avec le terme à représenter. Dans le cadre de l'approche standard, cette relation est la relation d'association, une relation sous-typée ;
5. la production des vecteurs non ambigus implique que chaque sens d'un terme donné soit considéré comme indépendant.

Compte tenu de ces aspects ainsi que du fait que la constitution des lexiques bilingues n'est pas l'objectif principal de notre démarche expérimentale, plutôt que d'utiliser les vecteurs de contexte associés à des ressources de connaissance telles que WordNet (Fellbaum [1998]) ou Wikipedia en amont de la construction de notre ressource, nous avons fait le choix de concevoir et appliquer une méthode de construction et d'alignement des lexiques *à l'intérieur de la ressource de connaissance en cours de construction.* Les lexiques sont intégrées dans les sous-graphes du $RLSM_{PI}$ pour ensuite être alignés (lorsqu'un alignement est possible compte tenu des informations contenues dans la ressource au moment de l'essai). correspondants Cette méthode peut être qualifiée de méthode d'alignement des lexiques bilingues « par immersion ». Cette méthode permet à la fois d'augmenter

24. En effet, pour un terme donné, l'ensemble des cooccurrences dans est d'une dimension finie.

25. Dans sa définition générale, un dictionnaire est un objet culturel qui présente le lexique sous forme alphabétique, en fournissant sur chaque terme un certain nombre d'informations. Un dictionnaire peut être plus ou moins structuré, il est généralement destiné à l'usage humain.

la ressource en cours de construction et d'extraire les lexiques bilingues alignés lorsque cela est nécessaire.

Lafourcade [2011] fait la comparaison entre le réseau lexical et le vecteur conceptuel en tant que structure. D'après cet auteur, le réseau lexical est une définition des termes "*en extension par rapport aux autres termes du lexique*" tandis que le vecteur correspond à "*une définition en intention*" [26]. En effet, dans réseau lexico-sémantique le voisinage d'un terme permet de déduire ses caractéristiques tandis que dans un espace vectoriel, un vecteur donné énumère les caractéristiques qui permettent de déduire les voisinages d'un terme. Étant donné que, contrairement aux vecteurs, dans les réseaux lexicaux le typage des relations est un typage local, le voisinage des termes une fois qu'ils sont insérés dans la ressource de connaissance sous forme de RLSM$_{\text{PI}}$ est une structure discrète sémantiquement riche. Ainsi, dans le cadre de notre méthode, les lexiques extraits du corpus comparable sont d'abord traités dans le cadre monolingue dans le but de découvrir les relations sémantiques présentes dans le RLSM$_{\text{PI}}$ ou pouvant être inférées. Le contexte devient un contexte augmenté qui peut intégrer, par exemple, les hyperonymes, les hyponymes, les variantes lexicales, les caractéristiques, les méronymes etc. Lorsqu'il s'agit d'un seul type de relation utilisé pour la construction du contexte, il est possible de parler de la signature telle qu'elle a été définie par Lafourcade [2011] [27]. Les alignements sont produits grâce à un pivot interlingue.

Pour appliquer notre méthode par immersion, il est nécessaire de collecter le corpus de comparable approprié et de modéliser les expressions polylexicales à l'intérieur de la ressource. La sections suivantes détaillent la façon dont les différentes méthodes de construction des ressources de connaissance introduites dans le chapitre 1 ont été appliquées à la construction d'un réseau lexico-sémantique multilingue avec pivot interlingue (RLSM$_{\text{PI}}$).

2.3.3 Corpus utilisés et méthode d'amorçage

Le tableau 2.1 résume les corpus construits et utilisés. Dans le cadre de l'amorçage et de l'augmentation du RLSM$_{\text{PI}}$, nous travaillons avec les corpus comparables dont la taille est assez réduite. Malgré l'appartenance au même domaine et une structure similaire, cette taille réduite (environ 5 millions de mots) rend les observations concernant les co-occurrences des termes relativement peu fiables. Par conséquent, nous ne nous fixons pas l'objectif d'en extraire des lexiques alignés. Chaque corpus est traité de façon indépendante, l'alignement est fait une fois que les termes et les relations acquis depuis les différents corpus sont intégrés dans le RLSM$_{\text{PI}}$ en cours de construction.

26. *Op.cit.* p.16.
27. *Op.cit.* p.20.

Lang.	EN	FR	RU	ES
Documents	18 337	35 245	34 232	8 269
Nombre de mots	1 900 325	3 172 050	2 567 400	661 520

TABLE 2.1 – Corpus comparables utilisés pour guider l'intégration des ressources de connaissance existantes (mots pleins après lemmatisation).

Le tableau 2.2 résume les différentes sources qui nous ont permis de collecter les documents (recettes de cuisine) pour constituer le corpus. Les travaux sur l'espagnol ont été lancés plus tardivement et de manière « complémentaire » ce qui explique la taille réduite du corpus correspondant.

langue	textes	segments	mots	termes	sources principales
Français	35 245	52 450	3 172 050	7 867	*DEFT 2013, CuisineLibre, RicardoCuisine*
Anglais	18 337	174 000	1 900 325	11 100	*AllRecipes, BBCFood, RicardoCuisine*
Russe	34 232	292 300	2 567 400	4 385	*AllRecipes, Gotovim*
Espagnol	8 269	12 690	126 900	1 900	*AllRecipes,QueRicaVida*

TABLE 2.2 – Caractéristiques quantitatives des sous-corpus. Les recettes ont été principalement récoltées à partir des sites Web collaboratifs.

Exemple 2.3

Exemple de recette de cuisine (format uniformisé après extraction depuis le Web)

<recipe>
<id>2</id>
<title>Biscuits à la semoule de blé, par jessica</title>
<ingredients>
<ingredient>200 g semoule de blé</ingredient>
<ingredient>1 pincée de sel</ingredient>
<ingredient>100 g de sucre</ingredient>
<ingredient>200 g de beurre</ingredient>
<ingredient>200 g farine</ingredient>
<ingredient>1 pincée de levure chimique</ingredient>
<ingredient>1/2 cuillère à soupe de cardamome en poudre</ingredient>
</ingredients>
<nutrition>
<no>Sans viande</no>
<no>Sans œuf</no>
</nutrition>
<steps>
<step>Mélangez soigneusement tous les ingrédients du bout des doigts, puis ramener en boule. Façonnez ensuite des petites boules (de la taille d'une balle de golf) et aplatissez-les dans la paume de votre main.</step>
<step>Les disposer sur une plaque graissée. Les décorer en appuyant légèrement le dos d'une fourchette sur chaque. Enfournez à 180°C pendant 15 minutes.</step>
</steps>
<comments>
<line>Très bon avec de la semoule fine, on peut remplacer le beurre avec un peu d'huile d'olive et la cardamome par une autre épice ou de l'eau de rose.</line>
</comments>
</recipe>

Pour l'amorçage de la ressource en cours de construction, nous avons procédé comme suit :

1. chargement de vocabulaire acquis à partir des corpus dans notre ressource sous forme de uni-grammes. Cette étape correspond à l'acquisition des termes *tels qu'ils apparaissent dans le corpus* ;
2. construction des *liens forme/lemme* et les relations grammaticales (relations qui servent à modéliser le rattachement des marqueurs morpho-syntaxiques [28] à un terme donné ;
3. construction des termes multi-mot, à savoir, identification des n-grammes correspondant aux termes multi-mot grâce aux méthodes fondées sur un dictionnaire, méthodes statistiques, méthodes translingues ainsi qu'un ensemble de patrons lexico-syntaxiques et relié les nœuds correspondant aux termes multi-mot à leur composant à l'aide de la relation typée *r_phrase* ;
4. extraction des relations sémantiques depuis le corpus en utilisant un ensemble de patrons lexico-syntaxiques et lexico-sémantiques ;
5. intégration les relations entre les termes désormais présents dans le RLSM ainsi que leur relations sémantiques et les termes opposés de ces relations à partir des ressources existantes ;
6. amorçage du pivot interlingue à partir de l'édition anglaise de DBNary (Sérasset [2014]). Ainsi, le pivot est amorcé en tant que pivot naturel pour se transformer ensuite en pivot interlingue de façon incrémentale (au fur et à mesure que le $RLSM_{PI}$ est peuplé notamment par les mécanismes d'inférence ascendants).

Dans les sections qui suivent, nous allons détailler les méthodes que nous avons utilisées pour l'extraction des termes et des relations depuis le corpus, l'intégration des ressources existantes, la consolidation du réseau. Les méthodes d'extraction terminologique se décomposent en méthodes statistiques, indépendantes d'une langue donnée, et méthodes symboliques ancrées dans la langue car elles se basent sur des ensembles de patrons.

2.3.4 Extraction des termes

La segmentation des corpus correspond, dans le cadre de nos expériences de construction au découpage des corpus en phrases (instructions de cuisine) et au découpage du texte en n-grammes. Pour la production des n-grammes candidats, nous avons utilisé la fonction NGrammes suivante.

Dans le cadre du processus d'extraction des termes et des relations, nous avons expérimenté trois approches au pré-traitement des sous-corpus : outils de pré-

28. Les marqueurs morpho-syntaxiques sont représentés au sein du $RLSM_{PI}$ sous forme de termes interlingues. Leur étiquette est une chaîne de caractères spécifique, par exemple, **in :Noun**, **in :CaseInstrumental**, **in :CaseNominative-Pl**. À ces termes sont associées les relations typées : *r_pos*, *r_form*.

```
1  Fonction NGrammes(Chaîne, Séparateurs[]) :
2      // initialisation
3      liste_mots[]← ∅ ;
4      liste_mots[]←Segmenter(chaîne, séparateur[])
5      i=0 ;
6      longueur_chaine=longueur(liste_mots[])-1 ;
7      candidats[]← ∅ ; // longueur_chaîne teste tous les n-grammes allant de
           1 à la longueur de la chaîne fournie en entrée, la longueur
           maximale de n-gramme peut être imposée
8      while i ≤longueur_ chaîne do
9          j=i ; for n=0 ;n<longueur_ chaine+1 ;n++ do
10             for j++ ; j<=longueur_ chaine ; do
11                 candidats.Last()← liste_mots[j] ;
12                 j++ ;
13                 n++ ;
14         i++ ;
15     retourner candidats ;
```

traitement statistiques état de l'art, approche à dictionnaire ouvert et méthodes basées sur le graphe.

Outils de pré-traitement statistiques. Pour l'étiquetage en parties de discours et la détection des dépendances, nous avons utilisé un ensemble d'outils tels que TreeTagger [29] pour l'ensemble des langues, Stanford Parser [30] pour l'anglais et l'espagnol, MaltParser [31] pour le français, MyStem [32] et Russian Malt Parser [33] pour le russe.

Au cours du pré-traitement, nous avons constaté des erreurs d'étiquetage en parties de discours notamment pour le corpus anglais (langue avec relativement peu de marqueurs morpho-syntaxiques au niveau des formes rencontrées dans le corpus)

Exemple 2.4

Phrase sans complément d'objet direct (*Stanford Parser*) :

1 Drain _ NNP NNP _ 0 root _ _
3 peel _ JJ JJ _ 6 amod _ _
4 and _ CC CC _ 3 cc _ _

29. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
30. https://nlp.stanford.edu/software/lex-parser.shtml
31. http://alpage.inria.fr/statgram/frdep/fr_stat_dep_malt.html
32. https://tech.yandex.ru/mystem/
33. Informations détaillées sur les différentes ressources et outils disponibles pour le pré-traitement des textes en langue russe http://corpus.leeds.ac.uk/mocky/

5 finely _ RB RB _ 3 advmod _ _
6 mash _ NN NN _ 1 appos _ _
7 with _ IN IN _ 10 case _ _
8 the _ DT DT _ 10 det _ _
9 sour _ JJ JJ _ 10 amod _ _
10 cream _ NN NN _ 6 nmod _ _

Les erreurs d'étiquetage ne permettent pas d'extraire les verbes pertinents de cette phrase d'instruction sans complément d'objet direct.

Certains outils comme MyStem permettent d'obtenir de multiples hypothèses à propos d'un mot en activant une option spécifique.

Dans l'ensemble, nous nous sommes relativement peu servi des outils de pré-traitement statistique ce qui est dû à la difficulté de mettre en place une chaîne de traitement générique qui conviendrai pour tous les corpus et langues d'amorçage de même que pour les langues nouvelles qui pourraient être rajoutées dans le RLSM$_{\text{PI}}$. La performance de ces outils dépend fortement des corpus et dictionnaires qui servent à leur entraînement. La ressource que nous construisons étant elle-même placée dans le paradigme d'apprentissage, l'entraînement des outils dans le but de pré-traitement des corpus exogène nous a semblé superflu.

Approche dite « à dictionnaire ouvert ». Une autre approche que nous avons exploitée dans le cadre de détection des TMM et d'extraction depuis les corpus d'amorçage a été d'utiliser les ressources à base de n-grammes et des dictionnaires (notamment, les dictionnaires multilingues en ligne). Les termes présents à la fois dans les corpus d'amorçage et dans ces référentiels de connaissance ont obtenu le score global supérieur ce qui a permis de sélectionner les TMM pertinents. Le traitement a été fait de façon séquentielle en commençant par la taille de n-grammes la plus petite (bi-grammes). Ensuite, les TMM validés ont été détectés dans les listes de n-grammes candidats de taille supérieure. Une ligne supplémentaire a été ajoutée au fichier des candidats dans ce cas pour favoriser l'émergence de plusieurs options de découpage en TMM.

Définition 2.2

Le ***score global*** n-gramme est calculé de deux façons selon les référentiels de connaissance utilisés.

S'il s'agit des listes de n-grammes issus du corpus d'amorçage et d'un corpus externe de grande taille, nous produisons le $score_1$ qui est un score d'information mutuelle [a] locale normalisé par la somme d'information mutuelle locale et externe.

$$score_1 = \frac{I(X,Y)_{loc}}{I(X,Y)_{loc} + I(X,Y)_{ext}}$$

S'il s'agit d'un TMM présent dans un dictionnaire en ligne, selon la confiance accordée à ce dictionnaire exprimée sous forme de coefficient $\beta \in [0, \frac{1}{2}]$, le score $score_2$ est calculé :

$$score_2 = \beta \times \frac{I(X,Y)_{loc} + 1}{I(X,Y)_{loc}}$$

a. L'information mutuelle de deux variables aléatoires X et Y notée $I(X,Y)$ est une quantité mesurant l'indépendance statistique de ces variables. Dans le cadre de calcul de l'information mutuelle, on cherche à comparer l'apparition de X et Y séparément l'une de l'autre par rapport aux cas où ces variables apparaissent ensemble.

$$I(X,Y) = \frac{P(X,Y)}{P(X)P(Y)}$$

Dans le cas où on connaît la distribution jointe des variables (par exemple, distribution gaussienne, loi normale), il est possible d'affiner la méthode de calcul de l'information mutuelle en ajoutant un coefficient approprié à la formule de base.

L'approche à dictionnaire ouvert a été intéressante pour le russe qui dispose d'une ressource СинТагРус (SynTagRus) [34] qui possède un étiquetage morpho-syntaxique (arbre de dépendances) dérivé du modèle Sens-Texte (Jolkovsky and Mel'čuk [1967]) et propose également l'étiquetage sémantique basé sur les fonctions lexicales. Nous avons exploré partiellement les possibilités de cette ressource et des ressources associées (dont les listes de n-grammes), une exploration plus détaillée notamment en ce qui concerne l'étiquetage sémantique en fonctions lexicales et rapprochement avec certains types de relations présentes dans RezoJDM serait à venir.

L'inconvénient majeur de cette approche est la nature souvent générale des termes et TMM présents dans ces ressources. Ainsi les TMM relatifs au terme « gâteau »(*mopm*) n'appairassent qu'à partir de la fréquence 5 dans la liste des n-grammes extraits du SynTagRus. Nous observons ainsi un effet de *longue traîne*, soit présence d'un nombre de termes importants (dont termes spécialisés) avec des fréquences faibles.

Approche fondée sur le graphe. En parallèle avec l'approche statistique et celle « à dictionnaire ouvert », nous avons implémenté une approche fondée sur le graphe ($RLSM_{PI}$). Dans le cadre de cette approche, nous avons procédé comme suit :

1. chargement des données sur les catégories comme spécifié dans la sous-section 2.3.3 ;
2. utilisation successive des patrons de surface et des patrons lexico-sémantiques quand cela a été possible [35] pour extraire aussi bien des TMM que des relations sémantiques.

34. `http://www.ruscorpora.ru/index.html`

35. La mise en place d'un patron lexico-sémantique nécessite une présence des relations sémantiques et, en particulier, de relations taxonomiques.

Cette méthode permet d'effectuer un choix pondéré [36] entre plusieurs étiquettes morpho-syntaxiques.

En phase d'extraction, nous ne faisons pas de distinction trop "hâtive" entre les TMM qui correspondraient à de véritables « unités terminologiques » et des TMM qui correspondraient plutôt à une relation. Ces entités sont intégrées dans le $RLSM_{PI}$ et font l'objet de traitements ultérieurs.

Exemple 2.5

En termes de partie du discours, un terme qui est extrait dans le cadre de la construction du $RLSM_{PI}$ peut être aussi bien un nom qu'un verbe. Ainsi :

копченая колбаса ainsi que **вареная колбаса**, respectivement *saucisson bouilli* et *saucisson fumé* font partie des principaux types de saucisson présents dans la tradition charcutière russe. Ces entités correspondent à de véritables termes.

En revanche, **вареная курица**, *poulet bouilli* correspond plutôt à un état du poulet (et désigne également une personne apathique, ramollie).

Les entités extraites peuvent apparaître au pluriel : **адсорбционные свойства** (*propriétés absorbantes*).

Tous ces termes sont intégrés dans le $RLSM_{PI}$.

Nous avons répertorié systématiquement les différentes formes des termes. En effet, dans les langues dites flexionnelles telles que le russe, la forme d'un mot porte l'information morpho-syntaxique essentielle pour la distinction des relations prédicatives et des rôles thématiques. Par exemple, le terme `(в) тепле` qui signifie "(se trouvant) au chaud", dont la partie de discours est « nom » et qui est utilisé au singulier et dont le cas est "Prépositionnel" (cas qui implique la présence d'une préposition) porte l'information "se trouvant à l'intérieur". Le terme `(в) тепло` qui signifie "(qu'on met, qu'on est en train de mettre) au chaud" dont le cas est "Accusatif" porte l'information de "mouvement, processus" (mettre quelque chose dans un endroit chaud).

Une autre possibilité pour le calcul des TMM serait l'exploitation des plongements des mots et la statistique de co-occurrence des termes. Cependant, notre corpus est de taille assez réduite ce qui rend les statistiques de co-occurrence peu fiables. Par exemple, pour l'anglais, nous avons des candidats intéressants tels que (dans le cas de calcul des bi-grammes pour l'anglais) *stainless steel*, *basmati rice*, *paddle attachement* (ces termes apparaissent avec une similarité cosinus [37]

36. Choix entre plusieurs relations typées *r_pos* pondérées soit en fonction de leur force d'association si toutes les relations de ce type proviennent d'une seule et même ressource (le cas de la partition ***fr*** où toutes les informations morpho-syntaxiques ont été intégrées depuis RezoJDM) soit à partir des pondérations basées sur les origines des relations.

37. Pour deux vecteurs d'attributs non nuls A et B (obtenus, par exemple, à partir des

supérieure à 0,95) de même que d'autres co-occurrences malheureusement non pertinentes.

Dans le cadre de nos expériences, nous avons pu extraire des termes multi-mot qui se répartissent comme suit selon les langues.

Lang.	**EN**	**FR**	**RU**	**ES**
uni-grammes	4 494	3 534	4 302	2 089
bi-grammes	1 416	22 014	3 469	1510
tri-grammes	15 111	22 014	4 990	1 538
Total	21 021	11 400	6 485	3 166

TABLE 2.3 – Extraction des termes par les méthodes statistiques, symboliques et basées sur le graphe.

Le nombre des uni-grammes (tableau 2.3 reflète la taille du vocabulaire de base présent dans le corpus. Les nombres de bi-grammes tri-grammes correspondent à l'acquisition par des méthodes statistiques et à dictionnaire ouvert qui permettent de réduire le nombre des candidats.

Exemple 2.6

Exemple d'extraction statistique des TMM sur l'exemple du sous-corpus espagnol (calcul des *n-grammes*) :

al gusto, *a fuego bajo*, *vino blanco*, *queso crema* [a] : entités polylexicales à part entière, expressions figées et semi-figées, termes qui appartiennent à la terminologie de cuisine ;

hoja de laurel [b] : TMM qui "porte" les relations typées *r_ has_ part* et *r_ quantifier* ;

harina de trigo [c] : TMM qui "porte" la relation typée *r_ matter* ;

revuelve bien [d] : TMM qui correspond à la relation *r_ manner*.

a. Espagnol, resp. *selon goût, à feu réduit, vin blanc, fromage à tartiner.*
b. Espagnol, *feuille de laurier.*
c. Espagnol, *farine de blé.*
d. Espagnol, *remuez bien.*

Pour relier les formes récupérées aux formes canoniques (lemmes), nous avons utilisé les ressources suivantes pour les langues actuellement présentes au sein du RLSM$_{PI}$:

anglais : jeu de données FreeLing [38], une suite logicielle destinée au traitement automatique des langues ;

français : RezoJDM ;

fréquences des mots dans un corpus), la similarité cosinus correspond au produit scalaire des vecteurs divisé par la norme des vecteurs : $cos\theta = \frac{A \cdot B}{|A| \cdot |B|}$.

38. `http://nlp.lsi.upc.edu/freeling/index.php/node/1`

russe : extraction depuis un ensemble de ressources telles que Wiktionnary [39] et, à titre complémentaire, Glosbe [40], puis complétion manuelle des usages spécifiques au domaine ;

espagnol : FreeLing et SMM Lexicon qui fait partie de l'analyseur morphologique pour l'espagnol SMM, Mahlow and Piotrowski [2009].

En plus des méthodes décrites, nous nous sommes servie des structures présentes dans les listes d'ingrédients dans le cadre de la segmentation des textes et de la détection certaines entités polylexicales (EPL). Prendre en compte les EPL qui apparaissent dans les listes d'ingrédients (autrement dit, tenir compte de la structure des documents) permet de favoriser ces EPL pendant le calcul. En effet, il est plus probable que les n-grammes tels que *leche condensada.* [41] ou *leche entera* [42] qui apparaissent dans les listes d'ingrédients soient de véritables termes contrairement à *leche hirviendo* [43]. Cette méthode dépend toutefois de la façon dont les ingrédients sont représentés dans l'arbre `html` d'une recette.

Les termes sont créés comme détaillé dans l'algorithme 8 (annexe A). Une fois intégrés dans le $RLSM_{PI}$, les termes se présentent comme représenté dans l'exemple A.1 (annexe A).

2.3.5 Extraction des relations

L'extraction des relations sémantiques est au centre de toutes les méthodes de peuplement du $RLSM_{PI}$ car c'est une représentation sémantique riche et un alignement des termes et relations important qui sont recherchés. L'extraction des relations depuis le corpus est basée sur les patrons qui sont des patrons syntaxiques de surface ou des patrons lexico-sémantiques. Les patrons lexico-syntaxiques de surface ont permis d'extraire depuis les corpus les relations typées *r_ carac*, *r_ object*, *r_ location*, *r_ instrument*. Une liste succincte de marqueurs temporels et spatiaux ainsi que de mots outils a été constituée pour chaque langue.

Définition 2.3

Un patron lexico-sémantique est un patron syntaxique combiné avec un ensemble de contraintes sur les relations sémantiques que peuvent avoir les termes correspondant aux variables du patron.

`x le y avec un z`

$x \xrightarrow{r\_pos} Verbe$
$\& \; y \xrightarrow{r\_pos} Nom \; \& \; z \xrightarrow{r\_pos} Nom$

39. `https://ru.wiktionary.org/wiki`
40. `https://fr.glosbe.com/`
41. lait concentré
42. lait entier
43. esp. lait bouillant.

& $y \xrightarrow{r\_isa}$ *aliment* & $z==$ *"couteau"*

La règle ci-dessus permet de découvrir tous les aliments qui peuvent être coupés avec un couteau. Le patron lexico-sémantique permet de représenter une règle d'extraction dans laquelle se combinent les contraintes d'ordre syntaxique et les contraintes d'ordre sémantique.

type	**#relations**	**détail corpus, commentaire**
r_isa	67 894	Ensemble de corpus, patrons de surface, patrons à base d'arbre **html** (arborescence des catégories), inclusion lexicale.
r_hypo	688	La relation d'hyponymie est symétrique par rapport à la relation hyperonymie (typée r_isa), ces deux relations sont équilibrées de façon endogène.
r_has_part	662 737	Ensemble des corpus (listes d'ingrédients), patrons de surface pour différencier quantités et ingrédients.
r_matter	606	Ensemble des corpus (listes d'ingrédients), patrons de surface pour différencier quantités et ingrédients. La distinction entre parties-tout *partie* et parties-tout *substance* se fait ultérieurement dans le $RLSM_{PI}$ par des mécanismes de raisonnement en présence des génériques.
r_holo	224	Ensemble des corpus, puis symmetrisation endogène (la relation de holonymie est symétrique par rapport aux relations typées r_matter et r_has_part).
r_object	42 280	Corpus additionnels Yandex query logs (russe), Yahoo manner (anglais), ensemble des corpus d'amorçage. Patrons dont le plus simple est `V (JJ)? N`.
r_carac	8 300	Ensemble des corpus, patrons lexico-syntaxiques `(JJ)? N JJ`, `JJ (JJ)? N` etc.
r_location	2 086	Ensemble des corpus, patrons lexico-sémantiques avec marqueurs de lieu.
r_manner	3 250	Ensemble des corpus, patrons lexico-sémantiques spécifiques.
r_implication	430	Liste de paires prédicat-patient, analyse des ensembles prédicat-patient, validation manuelle.
r_consequence	116	Ensemble des corpus, patrons, validation manuelle.
r_instrument	58	Ensemble des corpus, patrons, validation manuelle.
r_pos	333 814	Ensemble des corpus, étiquetage automatique en parties de discours.

TABLE 2.4 – Extraction des relations depuis le corpus. Compte tenu de la spécificité du corpus, les relations typées *r_object*, *r_carac*, *r_has_part* prédominent parmi les relations extraites depuis les corpus d'apprentissage en utilisant les différentes méthodes que nous avons introduites. Les patrons aussi bien lexico-syntaxiques que lexico-sémantiques ont été largement exploités.

L'extraction des relations telle que nous l'avons utilisée dans le cadre de nos

travaux d'extraction correspond à l'*identification des relations à partir d'un ensemble de critères*. La *découverte* des relations sémantiques [44] (sans en définir les types) dans le cadre des méthodes non-supervisées comme celle proposée par Angeli et al. [2015] se situe au delà de nos préoccupations méthodologiques car implicitement nous cherchons à valoriser le processus même de construction du RLSM$_{PI}$ en tant que moteur d'apprentissage permanent. Par conséquent, nous faisons référence aux structures déjà présentes au sein de notre ressource pour acquérir de nouvelles connaissances de façon incrémentale.

Les relations extraites depuis le corpus d'amorçage grâce aux méthodes basées sur les patrons lexico-syntaxiques (environ 20 patrons) et lexico-sémantiques (environ 40 patrons) révèlent le caractère spécialisé du corpus utilisé. Le détail des extractions précisé dans le tableau 2.4 montre également l'utilisation des corpus autres que les corpus d'amorçage pour peupler certaines des relations sémantiques (notamment la relation *r_ object*). Ces extractions ont servi à enrichir les parties (sous-graphes) monolingues. Les différents sous-graphes ont été reliés via le pivot interlingue grâce à l'intégration des ressources multilingues.

Les nouvelles relations sont créées grâce à l'algorithme 9 (annexe A). une relation peut être recherchée

2.3.6 Intégration des ressources pré-existantes dans le RLSM$_{PI}$ en cours de construction

Dans le cadre du processus d'intégration, nous avons utilisé les ressources de connaissance variées. L'intégration se fait dans le but d'enrichir le RLSM$_{PI}$, notamment, grâce aux relations sémantiques présentes dans les ressources interopérables avec le RLSM$_{PI}$ sur le plan représentationnel ainsi que possiblement en termes de composants.

Les ressources qui ont été choisies sont pour être partiellement intégrées ont été RezoJDM, ConceptNet, WordNet, RWN (Russian WordNet, Loukachevitch [2016]), DBNary, Wiktionnary. Le mécanisme d'intégration est fondé sur les systèmes de production, soit, sur un ensemble de règles de mise en correspondance. Chaque règle comporte des *prémisses* qui portent sur les structures observées dans le cadre de la ressource à intégrer et produit une *conclusion* (une action de création des termes et des relations) au sein du RLSM$_{PI}$.

L'action de base du processus d'intégration est la création (intégration) d'une relation (soit terme de départ, relation orientée et typée, terme d'arrivée). Vis-à-vis de cet aspect relationnel du processus d'intégration, nous distinguons l'intégration des relations sémantiques et l'alignement via le pivot interlingue. Ainsi, de

44. Notamment, dans le cadre des méthodes d'extraction ouverte d'information (*open information extraction*, `https://nlp.stanford.edu/software/openie.html`). Cet ensemble de méthodes n'utilise pas de types de relations ni de contraintes sur ces relations prédéfinies.

façon générale, les ressources pouvant être intégrées dans le $RLSM_{PI}$ doivent comporter des relations sémantiques ou des relations de traduction.

Certaines particularités ont pu être observées au cours de l'intégration des différentes ressources.

RezoJDM. La partie du RezoJDM en rapport avec le domaine de la cuisine et de la nutrition a été intégrée dans le $RLSM_{PI}$. Le calcul qui a permis de sélectionner cette partie a été d'identifier les termes appartenant au domaine (relations typées *r_domain*, *r_meaning*) et leur voisinage. Étant donné que l'architecture RezoJDM est notre architecture de référence, il s'agit des relations qui ont pu être importées presque sans modification. Les ajustements ont concerné le nommage des types de relations (*r_patient* devient *r_object*, *r_lieu* devient *r_location* etc.). Les poids RezoJDM ont été intégrés tels quels car il s'agit de la force d'association entre les termes, type de pondération cohérent pour l'utilisation dans le cadre du $RLSM_{PI}$.

type	**#nombre relations**
r_isa	260 309
r_hypo	124
r_refinement	29 286
r_carac	60 044
r_incompatible	3 420
r_matter	26 249
r_has_part	139 184
r_domain	67 093
r_holo	22
r_location	36

TABLE 2.5 – Relations intégrées depuis RezoJDM. RezoJDM possède un ensemble large de types de relations. Ces types de relations permettent de représenter explicitement de nombreux aspects lexicaux (variantes de termes, relations lexicales telles que la synonymie) et sémantiques.

ConceptNet. Le modèle de ConceptNet est proche de celui de RezoJDM et, par conséquent, de celui de $RLSM_{PI}$. Il s'agit également d'un réseau initialement construit par peuplonomie. Contrairement à RezoJDM, ConceptNet n'est plus alimenté par la contribution directe de joueurs. En revanche, cette ressource a été reliée à DBPedia [45] ce qui a permis de la rendre multilingue. La sous-partie de ConceptNet qui nous a intéressée est le sous-graphe anglais qui contient une palette de relations sémantiques intéressante. La problématique qui a dû être résolue a principalement concerné le fait que les relations de ConceptNet ne sont pas orientées. Par conséquent, les règles de conversion ont inclus un mécanisme de détection de la direction de la relation [46].

45. `https://wiki.dbpedia.org/`

46. La direction d'une relation donnée est détectée en exploitant les relations hiérarchiques du réseau et en analysant les parties du discours concernées par les relations à intégrer. Par

type	#nombre relations
r_isa	2 552
r_hypo	563 131
r_carac	5 456
r_has_part	5 742

TABLE 2.6 – Relations intégrées depuis ConceptNet. Les types ciblés ont été les relations sémantiques les plus peuplées au sein du ConceptNet.

WordNet et RWN. En ce qui concerne son modèle, WordNet n'est pas directement interopérable avec le RLSM$_{\text{PI}}$ en cours de construction. Dans le cas de WordNet et des ressources qui en ont été dérivées (comme RWN), le regroupement par sens est fondé sur la similarité entre les sens et sur la notion de *synset*. Pour réorganiser les sens (ainsi que les relations sémantiques attachées à ces sens), il a été nécessaire de « déconstruire » les synsets dans une certaine mesure pour réorganiser les sens dans une hiérarchie des raffinements (où chaque terme polysémique s'inscrit dans une hiérarchie des raffinements). il s'agit d'un modèle « orthogonal » par rapport à celui de WordNet. Cet aspect sera développé dans la section 2.4.

type	#relations WordNet	#relations RWN	#total WN
r_isa	235 237	33 339	268 576
r_hypo	234 528	33 339	267 867
r_refinement	105 164	0	105 164
r_has_part	27 243	1 681	28 924
r_matter	3 394	204	3 598
r_holo	67 057	204	67 261
r_entailment	2 254	396	2 650
r_domain	50 668	1 794	52 462
r_causes	918	161	1 079
r_covers	0	7 474	7 474

TABLE 2.7 – Relations intégrées depuis WordNet et RWN. Toutes les relations sémantiques des termes présents dans le RLSM$_{\text{PI}}$ après amorçage ont été intégrées depuis WordNet et RWN. Cette intégration implique également l'intégration de la polysémie des termes et leur sens général.

En tant que ressource développée sur la base de WordNet, RWN a pu être partiellement aligné notamment via les gloses avec WordNet anglais.

exemple, la relation typée *r_carac* va d'un nom vers un adjectif. Dans le cas où les termes ont une même partie du discours, un filtrage logique par déduction ou par induction est appliqué. Pour un terme source (respectivement cible), le filtrage vérifie si ses hyperonymes ou ses hyponymes d'un terme donné ont une relation sortante (respectivement entrante) de type t (type issu de ConceptNet à tester). Si le termes à tester n'ont pas de relations hiérarchiques et leur partie du discours est inconnue, la direction est définie manuellement (rare).

DBNary et Wiktionnary. DBNary et Wiktionnary [47] ont principalement servi à amorcer le pivot du $RLSM_{PI}$. Dans un premier temps, ce pivot est créé en tant que pivot naturel (avec anglais comme langue du pivot). Au fur et à mesure que le $RLSM_{PI}$ sera peuplé, ce pivot se transformera en pivot interlingue de façon incrémentale.

type	**#nombre relations**
r_covers	403 966
r_pos	59 998

TABLE 2.8 – Relations intégrées depuis DBNary et Wiktionnary. Les types ciblés ont été les relations de traduction, quelques relations sémantiques ont également pu être intégrées (environ 60 triplets).

L'ontologie de référence MIAM fait objet du processus d'immersion dont les objectifs diffèrent de celles de l'intégration qui représente un des mécanismes de construction du $RLSM_{PI}$.

2.3.7 Augmentation

Augmentation de la ressource amorcée. Une fois que la ressource a été amorcée grâce à des extractions des termes et des relations à partir du corpus d'amorçage et les mécanismes d'intégration des ressources pré-existantes, le processus d'augmentation peut se mettre en place.

Dans le cadre de ce processus, les ressources telles que listes de termes, ressources terminologiques (telles que IATE), ressources termino-ontologiques et les ressources à base de corpus (par exemple, des listes de n-grammes) peuvent être utilisées. ***L'augmentation se fait principalement par ajout de nouvelles relations*** car on cherche à construire une ressource sémantiquement riche et à expliciter un maximum d'informations (on vise à assurer la connexité de la base). Par conséquent, il est nécessaire de spécifier les types de relations concernées par un processus d'augmentation donné.

La méthode appliquée se déroule selon les étapes suivantes :

1. préparer les termes candidats : découper le texte (les éléments d'une liste) fourni en entrée en n-grammes grâce à une fenêtre glissante de taille variable s telle que $1 \leq s \leq 4$ si l'entrée du processus ne constitue pas déjà une liste de n-grammes ;
2. choisir et spécifier la relation concernée par le processus [48] (ou des critères pour choisir un types de relation dans une liste) et spécifier sa valuation

47. DBNary est une extraction de Wiktionnary en format OntoLex, un certain nombre d'informations échappe à cette extraction. De ce fait, nous avons complété DBNary avec les extraction faites directement depuis Wiktionnary.

48. Le choix du type de la relation est un choix manuel car il s'agit des ressources ciblées de petite taille pour lesquelles il est possible de définir un type de relation ou un sous-ensemble de types de relation avec des critères de choix entre ces types.

en fournissant un *poids* (ex. poids RezoJDM qui peut être réutilisé) ou un *terme* qui servira à annoter [49] la relation ;

3. identifier les composants correspondants au sein du RLSM (termes dont les étiquettes correspondent aux candidats et relations que possèdent ces termes) ;
4. créer de nouvelles relations au sein du RLSM$_{PI}$.

Compte tenu de ce déroulement, les entrées du processus d'augmentation sont standardisées dans un format tabulaire (éventuellement, stockage dans un fichier). Dans le cadre monolingue, elles ont une forme `SOURCE; TYPE; VALUATION; CIBLE`. La valuation sous forme d'annotation implique que le terme qui permet d'annoter (contextualiser) une relation soit, de préférence, un terme interlingue (noté avec un préfixe « in : »). La langue utilisée est spécifiée à l'entrée du processus.

Exemple 2.7

Standardisation des entrées lors du processus de l'augmentation.

(cas monolingue)

```
SOURCE; TYPE; VALUATION; CIBLE
gâteau yaourt au chocolat; r_matter; 150, sucre
gâteau yaourt au chocolat; r_incompatible; in:possible; diabetes
GN Scarlet; r_carac; in:toxicity; very toxic
```

(cas translingue)

```
crema salmonado; apricot; es; en; in:color
```

(*Identification ou création du raffinement interlingue « in :apricot>color », puis création des relations typées* r_covers *du terme « in :apricot>color » vers « es :crema salmonado » et « en :apricot>color ». La glose « color » provient du caractère thématique de la liste (liste des couleurs).*

Dans le cadre translingue, il s'agit d'ajouter un lien translingue. Il est nécessaire de fournir une paire de termes avec leurs étiquettes de langue ainsi que éventuellement le terme interlingue désignant le contexte (lorsqu'il est connu, ce terme permet de rechercher un raffinement interlingue correspondant ou de créer un raffinement s'il n'existe pas). Par conséquent, la paire de termes fournie sera soit rattachée à un terme interlingue qui couvre déjà un des candidats

49. Ajouter une méta-information à une relation sous forme d'une structure en forme de graphe spécifique. Cette structure comprend une réification (terme qui sert à réifier la relation à annoter sous forme d'un terme) qui a trois relations sortantes : *r_ source* dont le terme cible est la source de la relation, *r_ annotation* dont le terme cible est le terme qui sert à annoter la relation et *r_ target* dont le terme cible est la cible de la relation. Le mécanisme d'annotation sera détaillé dans le chapitre 3.

(via la relation typée *r_ covers*) soit se dotera d'un terme interlingue couvrant nouvellement crée.

Compte tenu des contraintes liées au types de relations, s'il s'agit d'une extraction spécifique à partir d'un corpus de spécialité, cette extraction peut nécessiter des processus ciblés afin de capturer les types de relation que l'on souhaite augmenter et, éventuellement, leur valuation.

A titre d'exemple :

1. pour une relation typée *r_ carac*, il est possible de coupler des patrons de surface à un ensemble de contraintes sémantiques basées sur le $RLSM_{PI}$. En effet, le patron de surface `N JJ` peut renfermer différents types de caractéristiques : couleur (*rouge*), aspect (*brillant*), état (*découpé*) etc. Dans certains cas, il est possible de dégager ces différents types sous forme d'annotation de la relation *r_ carac* (inférer l'annotation en utilisant des règles spécifiques).
   (a) si nous avons *aspect* $\xrightarrow{r\_carac}$ *brillant* et *mélange* $\xrightarrow{r\_carac}$ *brillant*, nous pouvons valuer la relation et standardiser sa représentation à l'entrée du processus d'augmentation « `mélange; r_carac; in:aspect; brillant` ». En effet, « aspect » appartient à l'ensemble restreint de termes qui permettent de décrire un aliment à côté de « toucher », « bruit », « flaveur » ... ;
   (b) si nous avons *concombre* $\xrightarrow{r\_carac}$ *découpé* et *découpé* $\xrightarrow{r\_pos}$ *Verb :PPas*, nous pouvons standardiser sa représentation à l'entrée du processus d'augmentation comme « `concombre; r_carac; in:état; découpé` ». La présence de *Verb :PPas* indique que l'aliment a subi une action ou une transformation ;
2. pour une relation typée *r_ matter*, notamment dans le cas de composition nutritionnelle des aliments, il est possible de standardiser l'entrée du processus d'augmentation en passant par un terme aggloméré. Par exemple, si nous avons dans le cadre de la table CIQUAL, <orge $\xrightarrow{r\_carac}$ perlée> et <orge $\xrightarrow{r\_carac}$ entière> qui contiennent respectivement 17,3 et 9,1 protéines brutes, il est possible, en partant du format tabulaire de CIQUAL de standardiser cette information comme suit :

```
orge[r_carac]perlée; r_matter; in:17,3; protéine brute
orge[r_carac]entière; r_matter; in:9,1; protéine brute
orge; r_carac;; perlée
orge; r_carac;; entière
```

L'augmentation vise à enrichir les éléments déjà présents dans le $RLSM_{PI}$ de façon incrémentale. Les termes ajoutés dans le réseau par un processus d'augmentation que l'on pourrait appeler *aug-1* servent d'ancrage lors du processus d'augmentation *aug-2* au même titre que les termes issus de l'amorçage et d'autres processus antérieurs à *aug-2*.

L'augmentation permet de fournir les éléments nécessaires pour donner au RLSM$_{PI}$ une dimension plus rigoureuse du point de vue terminologique car c'est lors de ce processus que les informations de spécialité sont attachées aux termes sous forme de relations.

Si l'on se réfère à une méthode de construction ontologique telle que *Terminae* (Szulman [2012]), l'augmentation correspondrait à la phase de mise en place d'une termino-ontologie. Dans une certaine mesure, en utilisant la terminologie de l'auteur précité, l'augmentation permet de réaliser un *couplage entre les unités termino-ontologiques* (éventuellement récupérées au sein des ressources spécialisées) *et la langue générale*. Un tel couplage est bénéfique, en particulier, pour la recherche d'information. La différence consiste à proposer un typage de relations sémantiques plus riche lors de la modélisation.

type	**en**	**fr**	**ru**	**es**	**total**	*commentaire*
r_object	5 530	-	10 520	-	16 050	acquisition à partir des corpus de question de manière (*how to questions*) de Yandex pour le russe Völske et al. [2015] et Yahoo *Answers Manner Questions, version 2.0*
r_carac	52	65	334	58	509	Listes de caractéristiques des aliments (couleurs, formes, etc.) récoltées à travers le Web.
r_matter	1 217	28 000	-	910	30 127	listes de composition des aliments Oqali, Ciqual
r_covers	17 098	17 095	10 924	7 124	47 496	terminologies bilingues (base de données terminologiques multilingue de l'Union Européenne IATE)

TABLE 2.9 – Augmentation des relations typées *r_object, r_carac, r_matter, r_covers* à partir des ressources de spécialité.

La définition des ressources de spécialité est assez large car, selon les types de relations considérées, elle peut englober les ressources thématiques (questions de manière comme Yahoo [50], Oqali [51], Ciqual), les ressources terminologiques (IATE [52]) où les termes satisfont aux contraintes de canonicité. La canonicité implique que le terme a des relations sortantes et que dans le corpus il englobe les expressions de surface (formes, variantes, expressions).

50. `https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&guccounter=1`

51. Base de données de l'Observatoire de la Qualité de l'Alimentation`https://www.oqali.fr/Base-de-donnees-Oqali`

51. Table de composition des aliments réalisé par l'Observatoire des aliments de l'Agence nationale de sécurité sanitaire de l'alimentation`https://ciqual.anses.fr/`

52. `http://iate.europa.eu`

Dans le cadre du processus d'augmentation, nous nous sommes concentrée sur les relations listées dans le tableau 2.9.

Certains types de relations sont relativement difficiles à extraire depuis les textes. Il s'agit notamment des relations qui impliquent un rapport de subsomption/subordination telles que *r_ implication* (par exemple, *cuire à feu doux* $\xrightarrow{r\_implication}$ *chauffer*). Ce sont des relations qui s'expriment peu en termes de structures de surface. Une analyse plus approfondie telle que l'analyse des ensembles d'objet des actions désignées par les prédicats (verbes ou noms prédicatifs) peut être nécessaire. Par ailleurs, des listes alignées de termes peuvent être difficiles à trouver pour certaines paires de langues. Face à ce type de problème, nous utilisons les mécanismes translingues qui permettent d'inférer de nouvelles relations à partir des relations présentes dans le RLSM$_{\text{PI}}$.

En ce qui concerne les ressources terminologiques de spécialité, l'augmentation et l'intégration sont les deux processus qui permettent l'acquisition de ce type de connaissance. L'intégration concerne les ressources interopérables avec le RLSM$_{\text{PI}}$ en termes de modèle. L'augmentation permet l'acquisition depuis les ressources peu expressifs (listes, terminologies, taxonomies informelles).

2.4 Consolidation du RLSM$_{\text{PI}}$

En tant que méthode de construction, la consolidation comprend deux aspects. Il s'agit d'une part de *lier* les termes et les sous-graphes de termes présents dans les différentes partitions du RLSM$_{\text{PI}}$ via le pivot interlingue. D'autre part, il convient de *réduire le nombre de redondances au niveau du pivot interlingue* pour éviter l'"éclatement" des différents concepts d'ontologie potentiels.

Le formalisme que nous utilisons dans le présent travail pour modéliser et aligner les différents sens d'un terme au sein du RLSM$_{\text{PI}}$ afin de réduire l'impact de la polysémie des termes sur l'inférence des propriétés et concepts ontologiques comprend la notion de *raffinement d'un terme*. La construction ontologique s'appuie également sur la hiérarchie consistante et la richesse sémantique du RLSM$_{\text{PI}}$.

Dans la présente section, nous décrivons l'inférence des relations sémantiques, la consolidation des hiérarchies et l'inférence des concepts ainsi que l'inférence des raffinements.

2.4.1 Remontée - descente et inférence translingue des relations sémantiques

L'inférence de nouvelles relations sémantiques à partir des relations déjà présentes dans le RLSM$_{\text{PI}}$ est fondamentale pour le processus de construction ontologique utilisant ce type de ressource. Dans le cas d'une ressource mature qui possède un pivot interlingue stable (union raisonnée des bases monolingues),

il est possible de réduire le processus d'ajout des relations sémantiques à un mécanisme de remontée-descente. En revanche, un pivot interlingue en cours de construction nécessite les mécanismes d'inférence avec différentes formes de filtrage.

Principe. Il s'agit d'inférer les relations sémantiques présentes dans les différentes partitions du RLSM$_{\text{PI}}$ et obtenues par extraction depuis les textes ou par intégration des ressources structurées dans les partitions qui n'en contiennent pas.

Les relations exploitées dans le cadre d'inférence translingue pour identifier les termes dont les relations sémantiques génèrent les prémisses sont des relations typées *r_ covers* (par exemple, `in:onion` $\xrightarrow{r\_covers}$ `fr:ognon`) qui relient le terme interlingue aux termes qu'il couvre. Par conséquent, aussi bien dans la phase ascendante (langue → pivot) que dans la phase descendante (pivot → langue), nous pouvons supposer qu'il s'agit de termes équivalents. Or, un seul et même terme couvert peut avoir plusieurs termes interlingues couvrants qui peuvent correspondre à plusieurs sens (annexe A, fonction A.4). Le cas inverse est également fréquent (annexe A, fonction A.5). Ainsi, plutôt que de considérer la relation typée *r_ covers* comme une relation d'équivalence stricte, nous la considérons comme une variante translingue de synonymie potentiellement incomplète (ce qui est illustré par l'exemple 2.8).

Lorsque le terme lexicalisé a plusieurs termes interlingues couvrants, ce cas d'inférence est traité comme s'il s'agissait d'une inférence avec raffinements. Ce processus a été décrit par Zarrouk [2015] et consiste à vérifier la présence des relations sémantiques entre les raffinements du terme et l'extrémité opposée de la relation à inférer.

Déroulement. L'inférence translingue des relations sémantiques se déroule en deux temps. Dans une premier temps, les relations sont inférées dans le pivot interlingue. Puis, les relations sémantiques du pivot sont inférées dans les partitions (chaque partition concerne les données d'une langue). Ces deux phases ne sont pas successives, le moteur associé à chaque phase fonctionne de façon indépendante. *Dans le cadre monolingue*, on considère les demi-relations partagées par une paire de termes sachant qu'une demi relation comprend une relation typée, orientée et pondérée et un terme-voisin.

Dans le cadre multilingue et ascendant (langue → pivot) la relation à inférer est considérée comme une instance de règle d'inférence. On transforme ses termes source et cible en ensembles de termes qui peuvent contenir aussi bien des termes interlingues que lexicalisés. On recherche des "faisceaux" de relations existantes entre ces ensembles de termes. Autrement dit, dans le cadre de processus ascendant :

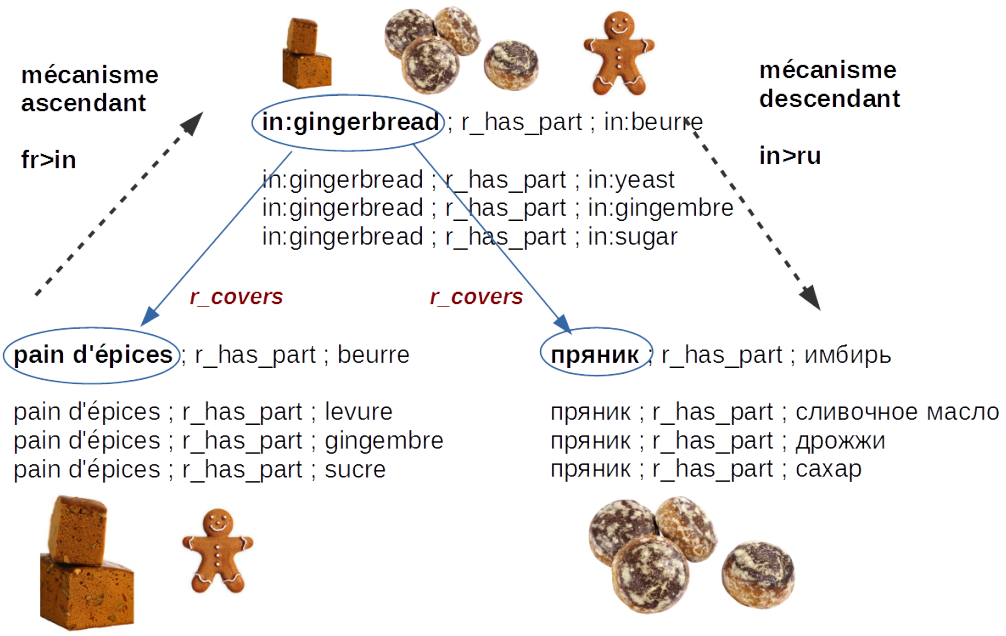
1. on récupère tous les termes interlingues et lexicalisés pour les deux termes considérés : un terme lexicalisé t_{lex} et un terme interlingue t_{in} tels que

$t_{in} \xrightarrow{r\_covers} t_{lex}$). Ainsi, on découvre dans quelle mesure le terme interlingue correspond au sens du terme lexicalisé. Si t_{lex} a plusieurs termes couvrants, seul un sous-ensemble des relations du t_{lex} pourra être proposé comme relations possibles de t_{in} ;

2. on explore le voisinage de l'intersection des ensembles de termes obtenus pour t_{lex} et pour t_{in} ;
3. si l'intersection entre les voisinages typés est suffisante (définie par un seuil), la relation est proposée pour être inférée entre les termes du pivot interlingue.

Exemple 2.8

L'exemple d'inférence des relations typées *r_has_part* à partir du français vers le pivot interlingue (ascendante) et à partir du pivot interlingue vers le russe (descendante) permet d'illustrer le fonctionnement d'inférence des relations sémantiques. Il montre également certaines problématiques liées à cette inférence dont notamment la présence de la synonymie incomplète propre à la relation typée *r_covers*. Il est sans doute intéressant de raffiner le terme interlingue `in:gingerbread` afin de distinguer ses deux sens : *gingerbread>biscuit* et *gingerbread>cake*. Par ailleurs, si l'on souhaite expliquer le produit qui correspond à "prianik" (пряник) , on serait amenée de le définir comme « une sorte de pain d'épice » car, en effet, en termes de texture, « prianik » se situe entre la texture d'un biscuit et celle d'un cake. L'inférence des relations sémantiques pour ce terme russe suivrait alors le mouvement descendant depuis `in:gingerbread` non raffiné (générique).



Dans le cadre multilingue et descendant, le processus d'inférence s'appuie de nouveau principalement sur le filtrage logique car de multiples termes couverts

pour un seul terme couvrant sont possibles. Par conséquent, une procédure similaire à celle décrite pour l'inférence ascendante est appliquée. Il s'agit de vérifier la présence des relations sémantiques entre le terme-source lexicalisé couvert par le terme interlingue source et les termes couverts par le terme interlingue cible.

Filtrage. Les inférences basées sur les exemples sont génératrices d'un nombre important de relations-candidates qui nécessitent une procédure de filtrage fiable afin de ne pas introduire de bruit dans la ressource en cours de construction.

Dans tous les cas de filtrage, nous avons appliqué un **pré-filtrage par parties du discours**. En effet, la plupart des relations sémantiques considérées dans le cadre de nos expériences relient les termes dont la partie du discours est "nom" [53]. Pour d'autres relations (par exemple, *r_object*, *r_carac*), il existe des contraintes d'ordre morpho-syntaxique qui sont considérées avant les autres types de filtrage.

Les procédures de filtrage se divisent en processus de filtrage appliqués en amont et en aval du pivot interlingue. *En amont du pivot interlingue*, il s'agit de proposer les relations issues des processus d'acquisition exogènes et endogènes. Par conséquent, elles peuvent être analysées du point de vue statistique en considérant leur *nombre*, leur *poids* et leur *origine*.

Le ***poids*** noté w correspond à la *force d'association* et s'applique aux relations issues des ressources construites par peuplonomie et disposant de ce type d'information comme, par exemple, RezoJDM. Un poids par défaut ($w = 25$) est également attribué à toutes les relations créées dans le cadre du RLSM$_{\text{PI}}$. A l'instar de la structure de RezoJDM, le poids peut être aussi bien un entier positif qu'un entier négatif. En effet, les relations avec un poids négatif sont des relations fausses (des vrais négatifs) qui sont maintenues dans la ressource pour modéliser l'erreur ou l'exception dans le but de s'en prémunir mais aussi pour donner plus de cohérence à la ressource. Ainsi, dans RezoJDM et dans RLSM$_{\text{PI}}$, pour un terme polysémique, les relations spécifiques à un raffinement peuvent avoir un poids négatif pour les autres raffinements.

Exemple 2.9

Le terme frégate possède deux raffinements *frégate>bateau* et *frégate>oiseau*.
Le terme *frégate>oiseau* possède la relation

$$\textit{frégate>oiseau} \xrightarrow[w>0]{r\_agent-1} \textit{voler}$$

Cette relation aura un poids négatif pour le raffinement frégate>bâteau :

53. L'étiquette de cette partie du discours est *in :Noun*, car les parties du discours ainsi que d'autres caractéristiques morpho-syntaxiques telles que le *cas*, le *nombre*, le *genre*, la dichotomie *animé/inanimé*, la dichotomie *perfectif/imperfectif*) des termes font partie du graphe interligue.

$$frégate > bâteau \xrightarrow[w<0]{r\_agent-1} voler$$

Dans le cadre du RLSM$_{\text{PI}}$, une stratégie de pondération qui prend en compte les origines des relations (extraction depuis le corpus, intégration depuis les ressources extérieures, inférence) a été mise en place. Cette stratégie est détaillée dans la section 4.1. Elle s'appuie sur la production et l'utilisation des ensembles d'indices de confiance accordée aux différentes origines des relations. La fonction de filtrage (définition 2.4.) exploite également les indices de confiance associés aux différentes origines des relations.

L'***origine*** correspond à la chaîne de caractères qui identifie la ressource ou le processus qui a fourni la relation lors d'augmentation, d'intégration ou d'inférence. Une relation peut avoir plusieurs origines. Un indice de confiance est associé à chaque origine (décision prise au cas par cas sauf pour l'inférence où la précision associée à l'itération la plus récente sert d'indice de confiance).

Au niveau du pivot interlingue et en aval de celui ci, il est possible de considérer le troisième paramètre statistique clé, le *nombre de relations partagées par les deux termes* soit le nombre des relations qui lient ces deux termes à un ensemble de termes aussi bien interlingues que lexicalisés. Dans son ensemble, la fonction de filtrage statistique peut être définie comme suit.

Définition 2.4: fonction de filtrage.

Fonction de filtrage $f(r)$ pour une relation r. Le but de cette fonction est d'éliminer du calcul des relations avec un poids négatif ou nul ainsi que des paires de termes qui ne partagent pas de demi-relations vers les ensembles de termes pré-définis. La fonction s'applique aux relations et aux paires de termes qui disposent d'un ensemble d'indices de confiance.

$$w \in \mathbb{Z},\ |\psi| \geq 1$$

w - poids de la relation.
ϕ - nombre de relations vers un un ensemble de termes partagées par les deux termes.
$\psi = \{i_1, i_2, ...i_n\}$ - ensemble des indices de confiance exprimés sous forme de nombres rationnels non nuls et inférieurs à 1 ($i_j \in [1, ...n] \wedge i_j \in [0, 1]$) accordés aux différentes sources intégrées d'où r peut provenir ainsi que les différentes méthodes d'acquisition (dont inférence).

$$f(r) = \phi \times \frac{w}{Max(\psi) \times log(|\psi|)}$$

$$f(r) \leq 0 \iff w \leq 0$$

Par exemple, on pourrait vouloir discriminer la relation telle que

$$courge > légume \xrightarrow[w=-50]{r\_has\_part} nez$$

Dans ce cas le poids négatif produirait un score de filtrage négatif. Au niveau du pivot interlingue et en aval de celui-ci, le filtrage logique et le filtrage par consensus peuvent être appliqués.

1. *filtrage logique* qui est basé sur la présence des relations sémantiques ;
2. *filtrage statistique* dans lequel on infère les relations d'un poids suffisant (ou dont l'origine est fiable) qui proviennent des termes qui partagent un nombre minimum suffisant de demi-relations sortantes (y compris les relations typées *r_ covers*) ;
3. *filtrage par consensus* dans lequel on prend en considération la convergence qui existe entre les différentes partitions du RLSM$_{PI}$.

La figure 2.6 détaille le fonctionnement du filtrage logique. Dans le cadre du RLSM$_{PI}$ les relations présentes au niveau des graphes de langue et du pivot interlingue peuvent être considérées. À terme, le filtrage est amené être fait majoritairement sur le pivot interlingue.

Le filtrage par consensus prend en considération la convergence qui existe entre les différentes partitions du RLSM$_{PI}$. Si la relation existe entre les termes reliés via le pivot interlingue et dans plusieurs graphes de langue, elle est prise en compte lors du calcul des relations à inférer.

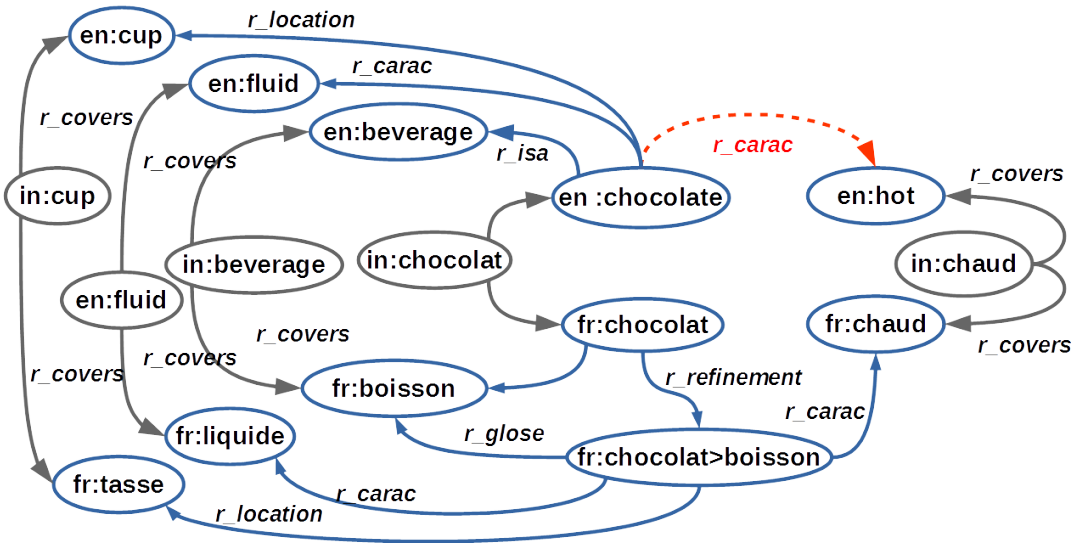


FIGURE 2.6 – Exemple simplifié de filtrage logique concernant l'inférence de la relation *chocolate* $\xrightarrow{r\_carac}$ *hot* dans le graphe anglais basée sur l'existence de relations sortantes partagées via pivot interlingue entre le terme français *chocolat>boisson* et le terme anglais *chocolate*.

Pour plus de lisibilité, dans la figure 2.6, nous n'avons pas montré les relations sémantiques présentes au niveau du pivot interlingue. La particularité du processus de filtrage proposé dans le cadre du RLSM$_{PI}$ réside dans le fait que, à l'étape actuelle, le ***test de présence des relations partagées peut être aussi bien monolingue (dans les limites d'une partition lexicalisée donnée) que translingue (vérification de la présence des relations entre les termes***

correspondants dans les autres partitions du $RLSM_{PI}$ ou dans le pivot interlingue).

Expérimentations. Les expérimentations ont été conduites sur l'ensemble de relations sémantiques et sur l'ensemble des langues du $RLSM_{PI}$.

Inférence ascendante. Le chiffrage des expérimentations d'inférence ascendante ci-dessous s'appuie sur des informations telles que le nombre de relations dans la partition d'origine, le nombre de relations candidates, le nombre de relations acceptées, la productivité de l'algorithme et le taux de relations acceptées. Les résultats sont présentés par langue sans tenir compte des intersections entre les ensembles de relations produits. Les champs peuvent être détaillés comme suit :

- #**orig** donne le nombre de *relations typées dans la partition d'origine* ;
- #**cand** correspond au *nombre de candidats* produits dans la partition d'arrivée (le pivot interlingue pour la méthode ascendante) ;
- **prod** est la *productivité du processus* pour une paire langue/pivot donnée (le rapport entre le nombre de candidats produits et le nombre de candidats d'origine) ;
- #**acc** correspond au *nombre de relations acceptées* (après validation automatique) ;
- **%acc** est l'expression de la *proportion des relations acceptées par rapport au nombre de candidats proposés* ;
- **rang** est une mesure d'évaluation introduite pour *situer le résultat obtenu par rapport à la productivité idéale qui se situe autour de 100%*. En effet, chaque relation est supposée produire idéalement une seule relation candidate dans la partition cible de la ressource. Lorsque l'algorithme produit trop de candidats, cela laisse soupçonner une mauvaise gestion de la polysémie (raffinements absents, trop de termes couverts pour un seul terme couvrant). Lorsque la productivité est inférieure à 1, cela indique une couverture insuffisante du pivot. Par conséquent, les partitions de la ressource dont la productivité s'écarte le moins de 1 (100%) ont un meilleur rang (la meilleure performance correspond au rang 1). Le rang est attribué par type de relation ;
- **pr** correspond à la *précision* constatée par validation manuelle sur un échantillon de 1 000 relations acceptées (proportion des relations justes parmi les relations acceptés). La précision obtenue permet de proposer l'indice de confiance accordée aux inférences produites lors de l'itération en cours pour de futures itérations qui utiliseraient les relations inférées.

La (pré)validation automatique est basée sur une série de règles à la fois de surface (listes de termes à exclure, règles dont les prémisses sont basées sur

la présence des relations sémantiques et grammaticales de base, par exemple, $x \xrightarrow{r\_carac} edible$ ou $x \xrightarrow{r\_pos>animacy}$: $Inanimate$ [54].

type	**#orig**	**#cand**	**prod**	**#acc**	**%acc**	**rang**	**pr**
r_isa	148 409	24 379	16%	13 886	**57%**	2	72%
r_has_part	178 286	4855	3%	3978	**37%**	2	59%
r_matter	16 419	4 547	28%	3728	**82%**	1	87%
r_object	14 655	54 517	372%	10 592	**19%**	1	94%
r_carac	32 585	58 809	180%	7 057	**12%**	2	75%
r_location	24 994	18 216	73%	10 715	**59%**	1	89%
r_instrument	3 543	4 074	115%	3 544	**87%**	1	77%
r_manner	1 873	1423	55%	1 109	**78%**	2	77%
r_incompatible	1710	1 386	81%	1150	**83%**	1	92%
r_telic	1295	738	57%	724	**83%**	1	67%
total	423 769	171 521	-	56 428	-	-	-
moyenne	-	-	99.6%	-	**68%**	-	80%

TABLE 2.10 – Inférence ascendante des relations sémantiques **fr→pivot**.

La partition **fr** de la ressource est la partition la plus riche en termes du nombre de types de relations et du nombre de relations des types clés pour le domaine de spécialité que nous étudions. L'inférence de certaines relations taxonomique et méronymiques affiche une productivité assez faible ce qui est dû aux intersections entre les ressources déjà intégrées dans le $RLSM_{PI}$ dont notamment WordNet ainsi que à la nature non encore interlingue du pivot amorcé grâce à l'intégration de l'édition anglaise de DBNary.

L'expérience continue sur les types de relation spécifiques typées *r_consequence*, *r_cause* ... Elle est régulièrement réitérée sur l'ensemble des relations. Les résultats listés sont amenés à évoluer constamment.

type	**#orig**	**#cand**	**prod**	**#acc**	**%acc**	**rang**	**pr**
r_hypo	314 452	50 118	16%	20 534	**41%**	1	93%
r_has_part	39 086	10 628	27%	3978	**37%**	1	78%
r_matter	1 709	500	29%	3728	**82%**	1	94%
r_object	7 088	9 190	129%	7 566	**91%**	2	89%
r_carac	5 474	2 436	44%	2 094	**86%**	1	94%
r_manner	1 751	5 938	339%	1603	**27%**	1	89%
r_entailment	1 127	2 638	234%	1 339	**51%**	1	78%
total	370 687	78 810	21	-	40 8420	-	-
moyenne	-	-	116%	-	**59%**	-	87%

TABLE 2.11 – Inférence ascendante des relations sémantiques **en→pivot**.

L'inférence depuis le sous-graphe **en** (tableau 2.11 reflète l'intégration massive des relations taxonomiques depuis les ressources structurées issues de la construction experte. la productivité des inférences des relations typée *r_hypo* est faible.

L'inférence depuis le sous-graphe russe a été utile pour le type de relation *r_object* (*r_patient* dans RezoJDM).

54. Ici "r_pos>animacy" indique une relation typée "pos" et annotée "in :animacy".

type	**#orig**	**#cand**	**prod**	**#acc**	**%acc**	**rang**	**pr**
r_object	14 458	53 638	371%	10 455	**72%**	3	59%

TABLE 2.12 – Inférence ascendante des relations sémantiques **ru→pivot**.

Les inférences ascendantes depuis le sous-graphe **ru** (tableau 2.12) présente la couverture insuffisante de ce sous-graphe par le pivot interlingue. Des patrons spécifiques à l'identification et l'inférence ascendante des relations telles que *r_instrument*, *r_location* à partir des relations typées r_pos (du fait de la richesse morpho-syntaxique de la langue russe) semblent être nécessaires.

L'inférence ascendante depuis le sous-graphe **es** n'est pas représentative en l'état de peuplement de celui-ci.

Inférence descendante L'expérience d'inférence descendante a concerné les partitions les moins peuplées, **es** et **ru**.

type	l	**#avant_inf**	**#inf**	**#après_inf**	**évolution**
r_isa	ru	46 827	7 036	53 863	+14 %
	es	36 807	268 040	304 847	+828 %
r_has_part	ru	65 772	3 682	69 454	+5 %
	es	10 166	56 883	67 049	+559 %
r_matter	ru	5190	4230	9 420	+81 %
	es	4013	7 351	7 764	183 %
r_manner	ru	1 265	1 655	2 920	+131 %
	es	1 753	9 507	11 260	+542 %
r_location	ru	640	621	1 261	+97 %
	es	90	567	657	+630 %
par langue	ru	119 694	17 224	136 918	+14 %
	es	52 739	342 348	395 087	+649 %
Totaux (moyenne)	-	172 433	359 572	532 005	***+208 %***

TABLE 2.13 – Inférence descendante des relations sémantiques.

Dans le tableau 2.13, nous faisons état du nombre de relationsdans la partition lexicalisée avant inférence descendante (**avant_inf**), par type. Nous détaillons le nombre de nouvelles relations obtenues par inférence (**inf**) et **évolution** (apport du mécanisme d'inférence). En fonction de la disponibilités des ressources externes pour telle ou telle langue et pouvant être intégrés dans le RLSM$_{\text{PI}}$, l'impact positif d'inférence peut être variable (colonne **évolution** du tableau 2.13).

Observations. En fonction des types de relations, de spécificité des langues et des sources de données, nous pouvons faire les observations suivantes.

1. **Monotonie et richesse des relations sémantiques obtenues par inférence.** Dans le cadre de nos expériences d'inférence des relations sémantiques, nous introduisons la notion de *monotonie*. Il s'agit d'un score d'observation M (score qui n'est pas utilisé dans le cadre de calculs sur le RLSM_{PI}) qui permet de se rendre compte de la richesse des relations sémantiques fournies par telle ou telle partition de la ressource et de la qualité d'alignement entre la partition source et cible. L'observation de départ concerne le fait qu'il existe souvent une asymétrie entre le nombre de termes sources et le nombre de termes cibles dans un ensemble de relations typées. Par exemple, ce dernier peut contenir beaucoup de relations typées *r_ object* dont le terme source est *couper* pour lequel beaucoup de termes cibles peuvent être observés. Il est possible de calculer le ratio entre la cardinalité de l'ensemble des termes sources et cibles.

   Par exemple, si dans le cadre d'un ensemble de relations typées *r_ object*, nous avons un ensemble de termes source $S, |S| = 1231$ et un ensemble de terme cible C, $|C| = 2469$ issus de la partition ***ru***, nous pouvons calculer $M_{in} = \frac{|C|}{|S|}$, $M_{in} = 0,5$ et $M_{out} = \frac{|S|}{|C|}$, $M_{out} = 2$ soit $M_{glob} = \frac{M_{out}}{M_{in}} = 4$. M reflète l'importance des relations sortantes par rapport à celle des relations entrantes eu niveau de cette partition. Les observations sont similaires quant à la monotonie du même type de relation dans les autres partitions lexicalisées.

   Lorsque l'on analyse le nombre de termes sources et cible de la relation typée *r_object* après le processus d'inférence ascendante (ensemble source S tel que $|S| = 787$, ensemble cible C tel que $|C| = 1019$), le ratio $M = M_{in\ pivot}/M_{out\ pivot}$ est de 1,67. Il est possible de définir ΔM qui correspond à la perte d'information sémantique lors du processus d'inférence qui est due à des problèmes d'alignement mais aussi à la granularité des ressources. Cette perte (qui correspond à un $\Delta M = 4 - 1,67 = 2,33$ dans notre exemple) représente un indicateur global. Il est ainsi possible de détecter automatiquement via le calcul du ratio de monotonie les parties du RLSM_{PI} qui nécessitent d'être peuplées ou mieux alignées.

2. **Valeur ajoutée des intersections entre les ensembles de relations provenant des différentes partitions du RLSM_{PI}.** Les intersections indiquent la présence des relations sémantiques partagées par plusieurs langues présentes dans le cadre d'un modèle. Nos expériences ont montré que les relations ainsi partagées affichaient une précision supérieure.

L'idée derrière la démarche d'inférence translingue (inférence ascendante et descendante sont conçues comme deux processus séparés) est d'enrichir des partitions des ressources peu peuplées grâce aux partitions riches en relations sémantiques.

L'enjeu principal de cette démarche concerne la présence des raffinements dans la ressource et la couverture du pivot interlingue (tableau 2.14).

label	fr	en	ru	es
#termes	189 961	247 635	92 655	45 236
#couverts	61 834	80 021	5 946	18 016
#silences (non symétrisés)	128 127	167 614	86 709	6
#polysémiques	23 472	12 281	1 497	18 007
#raffinés	8 251	26 932	1 497	0

TABLE 2.14 – État du RLSM avant consolidation

Le tableau 2.14 permet de se rendre compte des particularités suivantes des ressources multilingues :

— inégalité entre les sous-graphes de la ressource liées aux alignements existant entre les différentes langues de la ressources ;

— inégalité liée à l'acquisition des données multilingues ;

— inégalité liée au pivot interlingue (couverture du pivot lui-même), silences dans le sous-graphe *ru*.

Le tableau 2.14 reflète les inégalités du RLSM$_{\text{PI}}$ avant les processus de consolidation endogènes. Ce déséquilibre influence considérablement le processus d'inférence car, comme nous l'avons remarqué, il s'agit du schéma d'inférence avec raffinements. La section suivante est consacrée à la problématique des raffinements de sens et de leur inférence sur la base des raffinements glosés.

2.4.2 Inférence des raffinements et alignement

Le sens est l'usage. La relation de *raffinement* est une relation spécifique qui permet de modéliser la distinction de sens au sein d'un réseau lexico-sémantique. Il s'agit d'une notion fondamentale de la sémantique relationnelle. Le raffinement peut être *morphologique* ou *sémantique*. Dans le cas de raffinement morphologique, il s'agit de distinction entre les différents sens d'une même vocable due à son appartenance à plusieurs catégories grammaticales à la fois.

Exemple 2.10: raffinement.

Pour le terme *swallow*

— deux raffinements morphologiques *swallow>verbe* qui signifie "avaler" et *swallow>nom* ;

— *swallow>nom* possède deux raffinements sémantiques *swallow>gulp* qui signifie "gorgée" et *swallow>bird* qui signifie "hirondelle".

Dans le cas de raffinement sémantique, il s'agit de distinction des différents sens des termes (figure 2.7).

Définition 2.5

La relation de raffinement peut être exprimée en tant que fonction lexicale [a] dont le domaine de définition est le terme t à raffiner et l'image est l'ensemble des raffinements *raff(t)*. La « fonction » de raffinement peut être exprimée en termes des usages nommés par un ensemble de sens S.

$$S = \{s_1,\ s_2, ..., s_i\}$$

$$raff : t \rightarrow t > g_i,\ s_i \in S$$

S correspond à l'ensemble des sens de t. Chaque sens s correspond à une clique ou quasi-clique de t.

a. Proposée dans le cadre de la théorie Sens-Texte, la ***fonction lexicale*** fait uniquement partie de la théorie linguistique. L'argument de la fonction lexicale est un item lexical I, la valeur de l'application de la fonction lexicale f_{lex} à l'item I est un ensemble d'items lexicaux $\{I_1, I_2, ...I_n\}$. Chaque item de la valeur de la fonction lexicale $f_{lex}(I)$ remplit à peu près le même rôle vis-à-vis l'item de *I.* Ainsi, il s'agit d'une fonction multivoque et approximative.

Globalement, sur le plan de l'architecture des ressources langagières, le modèle de représentation du sens par raffinement coexiste avec le modèle de représentation du sens basée sur les synsets (par exemple, WordNet, Fellbaum [1998]). Contrairement aux architectures de ressource basées sur les synsets, le modèle par raffinement repose sur le contraste entre les termes plutôt que sur leur similarité. La définition de sens par contraste est plus appropriée dans le cadre des méthodes contributives car elle peut être partielle et s'affiner au fil des contributions.

Il est important de ne pas confondre deux problématiques distinctes :

— **la distinction des différents sens** (ce qui correspond dans le contexte d'exploration d'un réseau lexico-sémantique à l'identification des *cliques maximales* ou dans un contexte de jeu d'acquisition lexicale à la contribution des joueurs) ;

— **le choix de la glose pour nommer le sens** intuité ou identifié grâce aux techniques de parcours de graphe ou à des méthodes contributives.

Dans le cas d'une ressource langagière qui possède déjà des relations de raffinement, il est possible de se baser sur les sens nommés existants dans l'une des partitions de la ressource pour inférer les sens dans les partitions qui ne contiennent pas ce type de relation. La recherche des cliques vient appuyer cette démarche mais il est important de noter que le calcul des cliques est fiable dans le cadre des ressources stables qui contiennent un nombre important de relations sémantiques.

Lors de la construction des ressources langagières par peuplonomie, les gloses qui permettent de nommer les différents sens sont choisies par les joueurs ou les

contributeurs. L'intégration de ces ressources dans le $RLSM_{PI}$ peut permettre d'intégrer les usages nommés (raffinements glosés) qui en font partie. En outre, les mécanismes de consolidation peuvent permettre d'inférer les usages nommés dans les langues qui n'en disposent pas lorsque le contraste en termes de sens existe et lorsque la glose (nom de l'usage) permet de découvrir les cliques ou quasi-cliques correspondantes dans la partition ciblée par le processus d'inférence.

Glose en tant que nom de l'usage et *raffinement glosé*. Le terme *glose* apparaît fréquemment dans le cadre des approches à base de définitions du dictionnaire (ex. algorithme *Lesk* (Lesk [1986]), la glose est un texte court à usage humain et peut, à ce titre, présenter des ambiguïtés. Comme remarqué par Lafourcade [2011] à propos de l'utilisation d'un réseau lexico-sémantique pour le calcul des vecteurs d'idées, "si le corpus d'apprentissage est un réseau lexical, on s'affranchit d'une bonne partie des ambiguïtés présentes dans les définitions". Cet auteur poursuit, à propos de la même thématique, "à partir d'un réseau lexical, il est possible d'avoir des vecteurs conceptuels étiquetés c'est-à-dire, qui couvrent une facette particulière (plus seulement les idées associées ou la thématique, mais également les agents ou patients typiques, etc.)".

Dans la littérature, la distinction des sens d'un terme est souvent associée à la notion de *domaine sémantique*. Selon la remarque de Gliozzo [2006], la notion de *domaine sémantique* a lié les approches linguistiques structuralistes (les mots ont un sens lorsqu'ils appartiennent à un champ sémantique particulier) et ceux basées sur la vision de Wittgenstein [2009] (les mots ont un sens lorsqu'il existe un jeu linguistique dans lequel ils peuvent être formulés, leur sens est leur usage). Aujourd'hui, l'approche à la définition de sens par rapport à un contexte large, une aire commune de discussion telle que *Économie*, *Politique*, *Cuisine* est largement adoptée en TAL. Elle a permis un nombre important de modélisations et implémentations informatiques concrètes basées sur l'analyse de la cooccurrence des mots dans des corpus. Parmi les exemples de ces implémentations, nous pouvons citer le projet autour des domaines WordNet (WordNet Domains [55]).

L'approche qui repose sur les domaines sémantiques a l'inconvénient de se baser sur une liste de domaines d'intérêt. Construire manuellement une référence multilingue qui définit une liste de domaines de référence serait une tâche ardue lorsqu'il s'agit d'un domaine de spécialité car il ne serait guère possible de se contenter des domaines généralistes.

L'approche proposée ici pour tenter d'utiliser les raffinements présents dans une partition donnée pour inférer (partiellement) des raffinements dans d'autre partitions dont le pivot interlingue, s'appuie sur l'option de raffinement glosé.

55. `http://wndomains.fbk.eu/`

Définition 2.6

Un **raffinement glosé (RG)** est un terme du RLSM$_{PI}$ (ou de tout autre RLS qui suit le modèle défini par Lafourcade [2011] pour RezoJDM) qui modélise un des sens d'un terme polysémique donné.

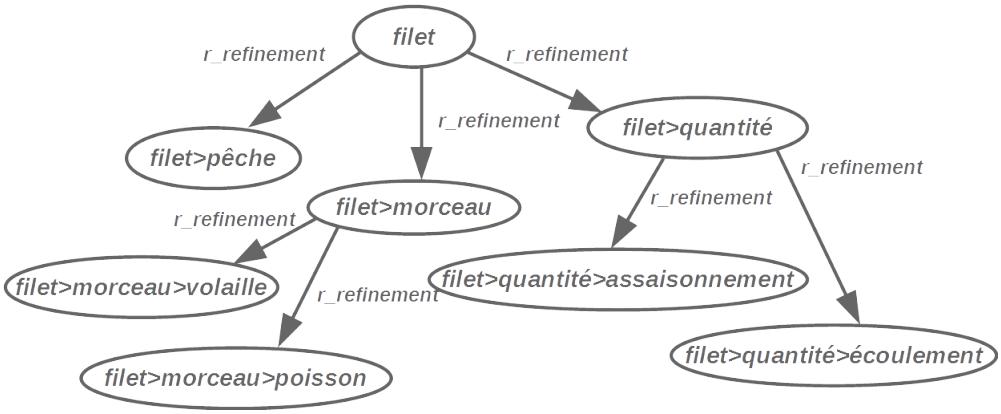
Un **RG** dispose obligatoirement de la relation entrante typée *r_refinement* (*r_raff* dans RezoJDM) qui le relie au terme polysémique à raffiner et d'une relation sortante typée *r_glose* (*r_meaning* dans RezoJDM) qui le relie au terme qui permet de nommer l'usage du terme à raffiner, la glose.

Un **RG** permet d'*identifier la clique* ou quasi-clique donnée et de la *nommer* via la concaténation de l'étiquette du terme à raffiner et celle de la glose.

Ainsi, pour le terme *baguette*, nous avons le sens « pain » par opposition aux autres sens (« encadrement », « bâton », « baguette magique »). Le raffinement glosé correspondant à ce sens est *baguette>pain*. Ainsi, nous avons la structure suivante dans notre ressource :

$$baguette \xrightarrow{r\_refinement} baguette{>}pain \xrightarrow{r\_glose} pain$$

Comme tout raffinement, un **raffinement glosé** peut être raffiné uniquement ou raffiné et glosé à son tour. Il peut faire partie d'un *arbre de raffinement glosés* (ARG) soit *arbre d'usages nommés* (exemple pour le terme *filet*).



Dans le cadre du présent manuscrit, les synonymes de *raffinement glosé* sont *raffinement nommé*, *sens nommé*, *sens d'usage nommé*.

La sémantique d'un raffinement glosé (RG). Lorsque l'on considère les différentes gloses présentes dans notre ressource de référence, RezoJDM, on se rend rapidement compte que, dans le cadre des raffinements glosés, la nature de la relation entre le terme à raffiner et la glose choisie de façon contributive ne se limite pas à la relation de rattachement à un domaine sémantique particulier.

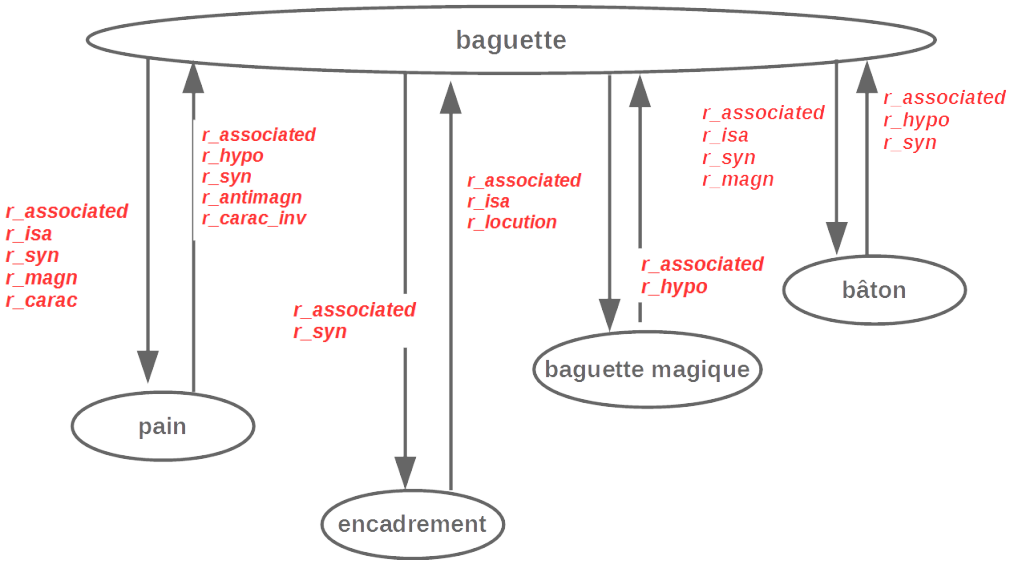


FIGURE 2.7 – Une partie du sous-graphe des usages nommés du terme « baguette ».

La figure 2.7 détaille les relations sémantiques qui existent entre le terme et les gloses qui définissent ses différents usages nommés. Outre les relations d'association (typées *r_associated*), se dégagent les relations de synonymie, hyperonymie (et hyponymie), magnification (amplifcation).

Ainsi, le sous-graphe des usages nommés d'un terme peut prendre la forme représentée sur la figure 2.7. Par conséquent, dans une certaine mesure, l'inférence translingue des raffinements limitée aux raffinements glosés peut être étudiée du point de vue de la sémantique du lien terme à raffiner - glose.

Une analyse exploratoire sur 2 224 paires terme-glose issues de notre ressource de référence, RezoJDM a permis de découvrir la présence des types de relation et leur distribution suivante : *r_associated* (39%), *r_syn* (18%), *r_isa* (9%), *r_hypo* (7%), *r_domain* (6%), *r_has_part* (3%), *r_holo* (3%), autres types [56] (15%).

Datant des écrits de de Saussure and Engler [1990] (« constellations associatives »), l'idée de réunir les vocables selon les associations des locuteurs d'une langue a été introduite par Bally [1965] (« champ associatif »). La relation typée *r_associated* a pour cible un ensemble d'entités qui peut se séparer en plusieurs rangs cognitifs. Il s'agit d'une relation sous-spécifiée. Cet aspect rend l'exploitation exclusive de cette relation insuffisante pour le choix automatique d'une glose dans le cadre d'inférence translingue réduite aux raffinements glosés.

Les relations de *synonymie* et *locution* sont des relations lexicales, très probablement spécifiques à une langue donnée. Par conséquent, les RG où le choix de la glose a été motivé par ces types de relations n'ont pas été exploités. À titre d'exemple, *carafe/container* est bien un raffinement partagé par le français et

56. Types tels que *r_consequence, r_color, r_lieu, r_locution, r_carac, r_sentiment* ...

l'anglais tandis que *carafe/tête* est basé sur la synonymie partielle et n'existe qu'en français.

Les relations d'*hyperonymie* et *hyponymie* font partie des relations sémantiques et, à ce titre, ne dépendent pas d'une langue donnée. Elles ne sont pas symétriques et le contraste artificiel en termes de granularité peut survenir [57]. Il est nécessaire de vérifier l'existence des relations sémantiques réelles ou inférables entre le terme et les candidats gloses dans la langue cible. La *méronymie* est une relation qui se rapproche de celle de *hyperonymie* et peut être exploitée dans le cadre d'inférence translingue des usages nommés.

Dans le cadre de $RLSM_{PI}$, l'inférence de nouveaux raffinements réduite aux RG peut se faire en choisissant une des stratégies suivantes :

1. ***inférence des RG ascendante et descendante à partir des RG déjà présents*** dans la ressource à partir des relations sémantiques observées entre le terme à raffiner et la glose ;
2. ***inférence descendante des raffinements non glosés et glosés en absence de RG et en présence de polysémie*** identifiée grâce aux multiples termes couvrants interlingues pour un même terme lexicalisé.

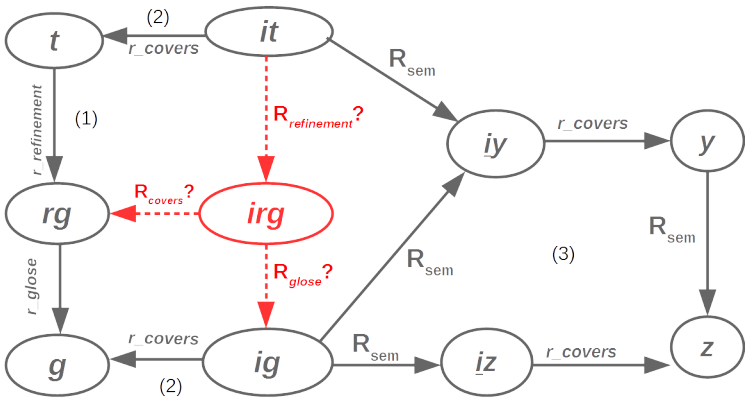
Inférence translingue des relations de raffinement. L'inférence des raffinements réduite aux RG se déroule en deux temps. Premièrement, les relations de raffinement présentes dans une ou plusieurs des langues sont inférées au niveau interlingue par un mécanisme ascendant. Deuxièmement, le schéma d'inférence descendant est appliqué afin d'inférer les sens depuis le sous-graphe interlingue vers les sous graphes des langues ayant peu ou pas de raffinements.

Dans le cadre du ***schéma ascendant*** d'inférence des raffinements réduite aux RG (langue vers le pivot interlingue), nous proposons des différents sens pour raffiner les termes interlingues. Tous les sens présents dans les partitions lexicalisées sous forme de RG se retrouvent dans le pivot interlingue.

Définition 2.7

Le ***schéma de raffinement de RG ascendant*** à partir d'un RG existant.

57. À titre d'exemple, si l'on souhaite proposer par transfert les raffinements du terme anglais *cow* que l'on sait polysémique à partir des raffinements disponibles en français basés sur l'hyperonymie tels que *cow/animal*, *cow/card game*, *cow/fish*, *cow/meat*, on se rend compte que le candidat-raffinement *cow/meat* serait superflu car en anglais il apparaîtrait comme $beef$, *viande* ne participe pas à la création de sens de *cow*. Le candidat-raffinement *cow/card game* serait faux car en anglais le terme désignant le jeu de vache pratiqué dans l'Ouest de la France est *aluette* et de ce fait *card game* ne peut pas raffiner *cow*.



Pour

- un terme raffiné t, son raffinement glosé rg, la glose g ;
- le terme couvrant de t, it ;
- le terme couvrant de g, ig.

S'il existe aussi bien une relation sémantique ou un chemin typé entre it et ig, un RG interlingue irg peut être proposé.

Nous considérons les prémisses suivantes en gardant les contraintes de poids des relations :

- RG dans une langue choisie pour l'inférence (1) ;
- au moins un terme couvrant pour le terme à raffiner et un terme couvrant pour la glose dans le pivot interlingue (2) ;
- la preuve de l'existence d'un lien (relation ou chemin) entre le terme à raffiner interlingue et la glose interlingue au sein même du pivot ou dans une des partitions lexicalisées (3) ;

$$\forall\ t \xrightarrow{r\_refinement} rg \xrightarrow{r\_glose} g\ (1)$$
$$\wedge\ it \xrightarrow{r\_covers} t\ \wedge\ ig \xrightarrow{r\_covers} g\ (2)$$
$$\wedge\ it \xrightarrow{R} ig\ (3)$$
$$\Rightarrow it \xrightarrow{r\_refinement} irg \xrightarrow{r\_glose} ig$$

L'inférence déclenchée si les prémisses sont réunies, implique la proposition de trois relations : $irg \xrightarrow{r\_covers} rg$, $it \xrightarrow{r\_refinement} irg$ et $irg \xrightarrow{r\_glose} ig$.

L'inférence ascendante des raffinements réduite aux RG permet de transférer les distinctions de sens vers le pivot interlingue. Suite à ce processus qui a été mis en œuvre à partir des partitions **en** (anglais) et **fr** (français), a permis d'atteindre le taux de raffinement du pivot interlingue de 30%.

cible	fr>int	en>int	intersection	total
int	8 558	33 930	254	31 752

TABLE 2.15 – RG interlingues acquis en appliquant le schéma ascendant. Au départ du processus, le pivot interlingue ne contenait pas de raffinement de sens. Par conséquent, le taux de productivité est de 100%.

Exemple 2.11

La partition ***fr*** a permis d'inférer les RG suivants :

```
in:world,in:world/humankind,in:humankind
in:world,in:world/subject,in:subject
in:world,in:world/land,in:land


in:oil,in:oil/painting,in:painting
in:oil,in:oil/substance,in:substance


in:milk,in:milk/cosmetic,in:cosmetic
in:milk,in:milk/food product,in:food product
```

tandis que les proposition telles que :

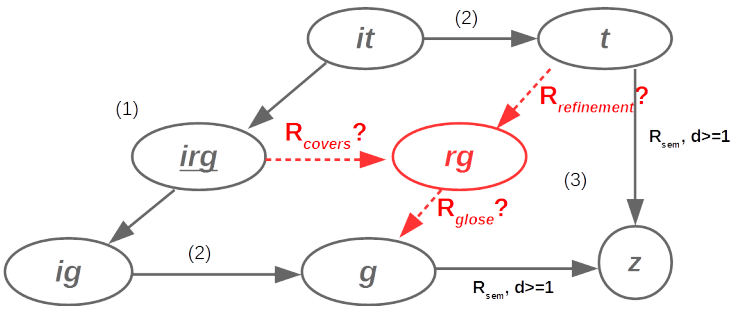
```
in:calculation,in:calculation/medicine,in:medicine
```

ont pu être rejetées.

Lorsque plusieurs termes couvrants sont reliés au même terme lexicalisé, des redondances peuvent survenir. Un mécanisme dédié se charge de la réduction des redondances dans le pivot interlingue.

Le ***schéma descendant*** à partir d'un chemin de raffinement s'applique de façon similaire au schéma ascendant décrit ci-dessus.

Définition 2.8



Dans le cadre du schéma descendant, nous disposons au départ des prémisses suivantes :

— RG interlingue dans le pivot interlingue (1) ;

— au moins un terme couvert pour le terme interlingue *it* et un terme couvert pour la glose interlingue *ig* dans la partition de la langue choisie comme cible (2) ;

— un lien (relation ou chemin) entre t et g au sein de la langue ciblée par l'inférence descendante (3). Le ***schéma descendant*** à partir d'un chemin de raffinement interlingue peut être résumé comme suit :

$$\forall\ it \xrightarrow{r\_refinement} irg \xrightarrow{r\_glose} ig\ (1)$$

$$\wedge\ it \xrightarrow{r\_covers} t\ \wedge\ ig \xrightarrow{r\_covers} g\ (2)$$

$$\wedge\ t\ \xrightarrow{R} g\ (3)$$

$$\Rightarrow t \xrightarrow{r\_refinement} rg \xrightarrow{r\_glose} g$$

source	int>es	int>ru
int	1 800	890

TABLE 2.16 – RG acquis en appliquant le schéma descendant en espagnol et en russe. Le succès de ce type d'inférence est fortement dépendant de la couverture du pivot interlingue notamment dans l'hypothèse qu'il s'agirait des langues dites « peu dotées ».

Lorsqu'il n'existe pas de RG interlingue pour un terme lexicalisé donné, mais que ce terme est susceptible d'être polysémique car il possède plusieurs termes couvrant dans le pivot, le ***schéma de raffinement descendant en absence de RG*** peut être appliqué. Ce schéma considère les différents termes couvrants comme termes potentiellement liés la glose g qui doit être découverte. Il ne s'agit plus de reconstruire le chemin de raffinement à partir de ses extrémités, mais de découvrir la glose g la plus appropriée pour désigner les sens suggérés par les termes couvrants multiplies. L'intuition ici est double :

1. les termes couvrants indiquent la glose g ;
2. g se trouve parmi les nœuds sémantiquement liés à t.

Le processus d'inférence se décompose en deux processus distincts et indépendants.

Initialement, l'inférence est dite « non glosée » : les sens sont numérotés ou provisoirement étiquetés en exploitant les étiquettes des nœuds interligues. Par exemple, pour le terme russe `белок` (« blanc »(nom) mais aussi *protéine*), le résultat de ce processus initial est représenté sur la figure 2.8.

Ensuite, il est question de valider la distinction proposée, à savoir, regrouper les sens redondants, puis de proposer les gloses candidates appropriées dans la langue en cours de traitement. L'étape finale du processus serait le choix de la

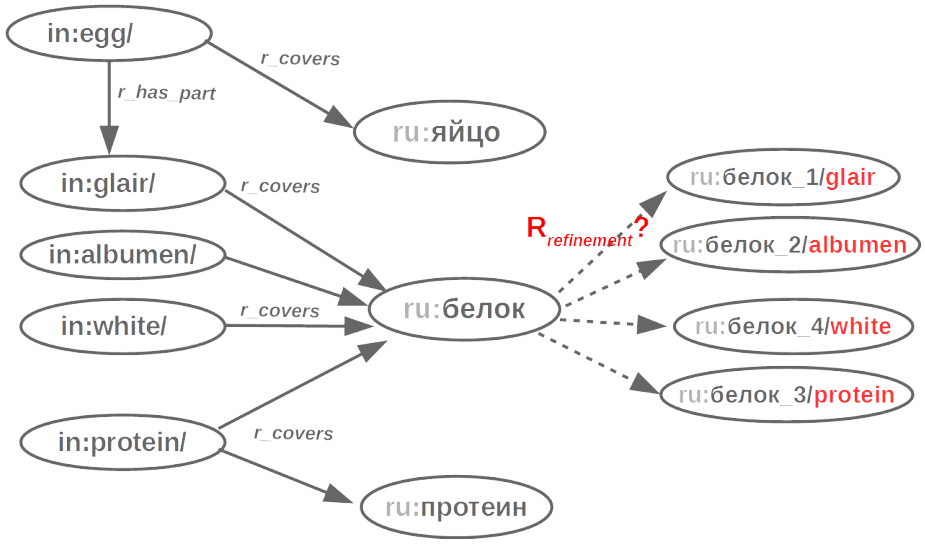


FIGURE 2.8 – Inférence de raffinement en absence de RG interligue. Raffinement non glosé.

glose appropriée dans la langue en cours de traitement. Ce choix est guidé par l'analyse de la sémantique des termes. Le schéma 2.8 fait apparaître le terme interlingue `in:egg` ainsi que la relation de méronymie (partie-tout) qu'il a vers le terme `in:glair`. Cette relation de méronymie permet d'inférer une glose candidate `яйцо` . Suite au regroupement des sens redondants grâce au filtrage statistique et logique concernant des relations sémantiques partagés par les termes potentiellement redondants, nous obtenons la distinction représentée sur la figure 2.9.

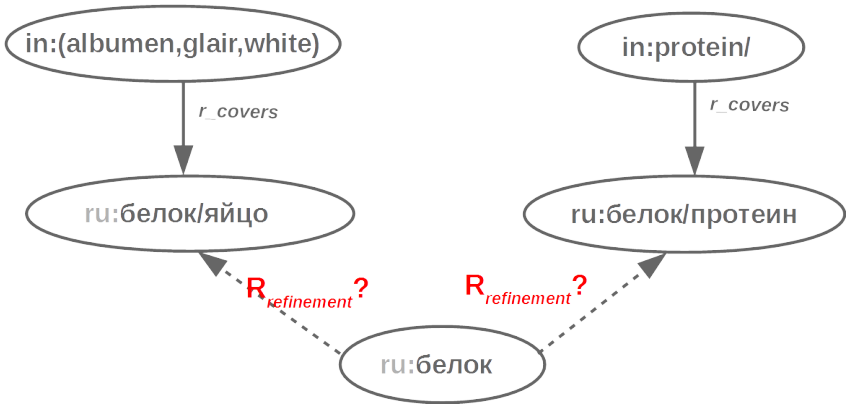


FIGURE 2.9 – Raffinement en absence de raffinement glosé (RG). Choix des gloses.

L'inférence en absence de RG interlingue recherche intuitivement les cliques qui correspondraient à des RG indiqués par les multiples termes couvrants d'un seul et même terme lexicalisé. Puis, les extrémités des relations ayant le plus de poids dans ces cliques sont proposées en tant que gloses.

2.5 État du RLSM$_{\mathbf{PI}}$

L'état du réseau peut être exprimé en termes du nombre de termes par langue, nombre de termes qui constituent le pivot, nombre de relations par type.

Vue a construction récente, le RLSM$_{\text{PI}}$ n'est pas encore une ressource stable. De nouvelles relations continuent à être ajoutées par des méthodes exogènes et endogènes. Cependant il est possible d'affiner les observations en considérant également :

— le nombre de termes *« couverts »* (ayant au moins un nœud couvrant dans le sous-graphe interlingue) sachant que la présence des nœuds couvrants pour un terme n'est pas synonyme d'alignement ;

— le nombre de termes *polysémiques* (ayant plusieurs termes couvrants) ;

— le nombre de termes non couverts dont la polysémie est inconnue (silences) ;

— le nombre de termes raffinés.

Les **termes** du RLSM$_{\text{PI}}$ se répartissent comme indique dans le tableau 2.17.

in	en	ru	fr	es	total
108 472	247 679	92 657	191 145	45 436	685 389

TABLE 2.17 – Répartition des termes au sein du RLSM$_{\text{PI}}$.

Les **relations** du RLSM$_{\text{PI}}$se répartissent comme suit :

type de relation	**#n**
r_isa	325 934
r_hypo	403 233
r_has_part	419 786
r_matter	19 213
r_holo	33 931
r_object	50 860
r_covers	327 829
r_pos	200 709
r_similar	39 000
r_refinement	92 730
r_location	28 609
r_instrument	3 633
r_agentive_implication	1 026
r_entailment	1 327
r_property	49
r_domain	59 851
r_incompatible	1 710
total	2 009 430

TABLE 2.18 – Relations présentes dans le RLSM$_{\text{PI}}$ à l'heure où nous écrivons.

La construction du RLSM_{PI} par des méthodes endogènes semble être un bon compromis en termes d'économie de temps et de ressources. Avec ce type de méthodes, le réseau se construit de manière incrémentale. Les problématiques de validation humaine des inférences produites peuvent être atténuées en proposant un système de pondération basé sur l'évaluation sur un échantillon constitué d'un ensemble de relations résultant d'un cycle d'inférence (moteur, version de l'algorithme d'inférence). Le score de cette évaluation fera partie du score de poids global de la relation (détaillé dans le chapitre 4) et impactera les mécanismes d'inférence ultérieurs.

2.6 Discussion

Le point qui reste à discuter à propos des méthodes que nous avons proposées pour la construction du RLSM_{PI} concernent la perte d'information sémantique lors des la mise en œuvre des différents mécanismes que nous avons introduits.

Dans le cadre exogène (cas d'*Augmentation*), cette perte d'information est plutôt un manque d'information et de contraste (possiblement le cas des ressources bilingues) inhérent aux ressources externes utilisées.

Dans le cadre multilingue et endogène, la perte d'information est due à la couverture possiblement insuffisante du pivot interlingue et au nombre encore insuffisant des relations sémantiques. Dans un cas comme dans l'autre, lors du passage d'une langue à l'autre, les inférences ne peuvent pas se faire car les prémisses ne sont pas retrouvées dans la ressource.

Un autre point important est le **coût des différents mécanismes en termes du temps de calcul**. Le coût des mécanismes de filtrage logique est détaillé dans le chapitre 5. Quant au coût d'inférence des raffinements réduite à RG, il est globalement moins élevé. En effet, compte tenu des observations sur la sémantique de la relation *terme-glose*, nous considérons un ensemble très réduit de types de relations sémantiques (relations taxonomiques, méronymie). Ce parcours typé fait baisser le facteur de branchement des termes et permet de réduire la complexité globale du processus.

Enfin, le fait de n'avoir que des raffinements glosés au sein du pivot interlingue peut susciter des questionnements. Cependant, au fur et à mesure que le pivot interlingue sera peuplé de relations sémantiques et qu'il évoluera vers un véritable pivot interlingue, les mécanismes de calcul des cliques interlingues vont devenir de plus en plus fiables afin de permettre le calcul des raffinements non glosés.

2.7 Conclusion du chapitre

Dans le présent chapitre, nous avons détaillé notre méthode de construction d'une ressource langagière sous forme de réseau lexico-sémantique multilingue avec pivot interlingue. Cette méthode permet un ancrage dans les textes issus des corpus de spécialité. Elle intègre partiellement les ressources de connaissance et des ressources langagières existantes ce qui permet d'assurer une interopérabilité du modèle par référence (par exemple, interopérabilité avec RezoJDM) et une interopérabilité de contenu. Par conséquent, il devient possible d'automatiser le processus de mise à jour de la ressource à partir de nouvelles versions des ressources ayant fait objet de processus d'intégration.

Nous avons également détaillé le processus d'augmentation à partir des ressources de spécialité (généralement, des ressources peu structurées) abondantes dans le contexte industriel (mémoires de traduction, listes de termes, petits corpus de spécialité).

Enfin, il semble, compte tenu de nos expériences, que les processus d'inférence translingue puissent représenter une solution intéressante quant à l'alignement des différentes partitions de la ressource notamment par sens.

Contributions du chapitre 2

Parmi les contributions du chapitre :

1. proposition d'une **méthode de construction d'une ressource langagière multilingue avec pivot interlingue**
   (a) définition de l'architecture de la ressource ; définition des méthodes utiles pour l'amorçage de la ressource à partir d'un corpus multilingue de textes de spécialité ;
   (b) définition des principales approches à la construction de cette ressource une fois qu'elle a été amorcée : intégration augmentation, consolidation, alignement.
2. **application des processus d'inférence endogène** définis notamment par Zarrouk [2015] (dont en particulier l'inférence par raffinement) **à la consolidation d'un réseau lexico-sémantique multilingue**.

Chapitre 3

Exploitation du réseau lexico-sémantique multilingue pour la construction termino-ontologique

Le présent chapitre porte sur les mécanismes d'enrichissement et de construction d'ontologie grâce à un réseau lexico-sémantique multilingue ($RLSM_{PI}$). Une fois le réseau construit, des projections de modèle peuvent être faites sur celui-ci afin d'en extraire des ressources de médiation qui peuvent servir notamment à accompagner la construction termino-ontologique et à réduire l'effort humain nécessaire pour la mener à bien. En effet, lorsque l'on dispose d'une ressource de référence, on peut l'immerger dans le $RLSM_{PI}$ pour ensuite inférer les éléments et structures correspondantes. L'ensemble de ces structures conformes à un modèle (par exemple, un modèle termino-ontologique) forme une ressource médiatrice qui peut être extraite depuis le $RLSM_{PI}$ pour ensuite appuyer la construction terminologique ou ontologique conduite par des experts humains.

Termes et notations utilisés dans le chapitre 3

ontologie : spécification formelle d'une conceptualisation partagée.

folksonomie : ressource informelle rattachée à une ontologie (ressource formelle) afin de fournir un « vivier » de termes et de structures sémantiques issus de l'acquisition des connaissances et non intégrables dans l'ontologie (pour des raisons de redondance, par exemple).

classe : ensemble d'individus ayant les mêmes caractéristiques.

propriété : relation d'ontologie. Cette relation peut exister entre une source (appelé *domaine*) et une cible (*co-domaine*). Si le domaine et le co-domaine de la propriété sont des classes (concepts), il s'agit d'une *Object Property* (propriété à valeur objet). Si le domaine est un individu tandis que le co-domaine est un littéral, il s'agit de *Data Property* (propriété à valeur donnée).

immersion : expression d'une ressource de connaissance spécifique en termes de structures du $\mathrm{RLSM_{PI}}$.

projection : identification d'un sous-ensemble de structures conformes à un modèle spécifique donné au sein d'une ressource de connaissance plus générale (par exemple, RLSM_{PI}).

Les ressources termino-ontologiques actuelles sont de ressources de plus en plus riches en données. Un projet spécifique, le projet NeOn [1] détaillé dans le chapitre 2, a été mis en œuvre pour définir un cadre méthodologique de construction ontologique permettant d'intégrer les connaissances variées dans le processus de construction, de construire des ontologies modulaires et de mettre les ontologies en réseau. Ce cadre méthodologique comprend notamment les phases d'acquisition et de modélisation. Le RLSM_{PI} est utilisé pour permettre de proposer une démarche de construction ontologique pour assister ces deux étapes de construction. Une telle démarche s'inscrit dans la perspective d'élicitation (mise en valeur) des connaissances nécessaires à la construction ontologique. Extraire des structures ontologiques à partir des ressources non ontologiques (telles que le RLSM_{PI}) constitue une alternative à la construction manuelle de telles structures notamment dans le contexte des ontologies multilingues et des ontologies noyau.

Enrichir ou extraire une ontologie de référence depuis une ressource langagière est une démarche qui nécessite la mise en place d'algorithmes qui se servent des structures sémantiques disponibles dans le cadre d'une telle ressource pour produire des structures ontologiques. Il ne peut pas y avoir de polysémie dans une ontologie tandis que la valeur ajoutée d'une ressource langagière est dans la manière dont elle modélise l'ambiguïté d'une langue naturelle. Comme il a été justement remarqué par Tchechmedjiev [2016], « la sémantique en langue naturelle n'est pas vraiment consistante, ni transitive ». Or, la construction ontologique repose sur des principes axiomatiques étant donné qu'elle est guidée par une sémantique formelle.

Les mécanismes qui seront décrits dans le présent chapitre sont basés sur des règles. Les règles d'immersion ainsi que les règles d'extraction sont des *règles de mise en correspondance* (*mapping*), à savoir des règles d'inférence exogènes tandis que les règles d'enrichissement sont des règles d'inférence des relations et d'annotations endogènes qui suivent plusieurs mécanismes décrits par Zarrouk [2015] et se fondent principalement sur l'abduction (inférence à partir des exemples de structures similaires).

Problématiques et organisation du chapitre - L'enrichissement et l'extraction d'ontologie depuis le RLSM_{PI} se base sur l'utilisation d'une ontologie de référence et se déroule en trois étapes qui se reflètent dans l'organisation du présent chapitre :

1. Encodage de l'ontologie de référence en termes du RLSM_{PI} (*immersion*) ;

1. `http://neon-project.org/nw/About_NeOn.html`

2. Consolidation des hiérarchies du $RLSM_{PI}$ par immersion de la hiérarchie des classes de l'ontologie de référence (ontologie MIAM) dans le $RLSM_{PI}$;
3. Calcul des structures conceptuelles approchant les relations d'ontologie de référence dans $RLSM_{PI}$ et extraction des structures ontologiques sous forme d'une ressource médiatrice.

À titre préliminaire, nous allons synthétiser notre méthode en spécifiant notamment le concept de l'*élément remarquable* tel qu'il peut être entendu dans le cadre de l'exploitation d'un réseau lexico-sémantique.

Nous allons détailler, *dans un premier temps*, le processus d'immersion de l'ontologie de référence (ontologie à enrichir) dans le $RLSM_{PI}$.

Dans un second temps, il s'agira de présenter la façon dont les structures hiérarchiques présentes dans le $RLSM_{PI}$ peuvent être consolidées pour que le $RLSM_{PI}$ puisse être utilisé dans le cadre de construction ontologique.

Dans un troisième temps, nous allons décrire le mécanisme exploitant un $RLSM_{PI}$ pour la construction ontologique. Ce mécanisme utilise les relations sémantiques présentes au sein d'un $RLSM_{PI}$ pour calculer des structures pouvant être utilisées dans le cadre de constitution d'une folksonomie adossée à une ontologie modélisée par les experts d'un domaine donné ou, s'il s'agit d'une ontologie en cours de modélisation, des classes et propriétés potentielles. Ce calcul peut s'appliquer aussi bien à la langue de départ d'ontologie ou d'ébauche d'ontologie qu'à une autre langue faisant partie du $RLSM_{PI}$.

3.1 Outils existants de construction d'ontologie

La problématique de construction ontologique outillée par le TAL à partir des textes en langue naturelle est explorée depuis plus de 20 ans. Parmi les outils de recherche des candidats termes se sont distingués SYNTEX-UPERY (Bourigault [2002]), YaTeA (Aubin and Hamon [2006]), BIOTEX (Lossio-Ventura et al. [2014]). Des architectures de développement des ontologies telles que *Terminae* (Szulman [2012]) utilisent ce type de outils. L'outil Archonte (Charlet et al. [2006]) qui définit les étapes de normalisation et de formalisation s'inscrit dans la même démarche ascendante que les outils pré-cités. Parmi les outils et architectures dédiés au développement des ontologies à partir des textes, se distinguent :

— les outils différenciés (outils qui prennent en compte la différence entre le terme et le concept d'ontologie). Dans le cadre de ce type d'outils, les unités terminologiques extraites des textes et organisées sous forme de réseau via un ensemble de relations hiérarchiques (hyperonymie) et d'équivalence (synonymie) servent à guider l'ontologue dans la construction ontologique. Ainsi, la construction d'ontologie passe par une structure intermédiaire,

une termino-ontologie (par exemple, *Terminae*, Szulman [2012], permet de mettre en oeuvre ce type de démarche) ;

— les outils non différenciés qui n'introduisent pas de distinction entre terme et concept et se basent sur des mesures statistiques (fréquences, *tf-idf*) pour proposer des candidats-concepts soit par analyse des concepts formels (ACF, ex. Mondary [2011]) soit par des méthodes basées sur la connaissance (ex. *TextToOnto*[2]).

Comme remarqué par Mondary [2011], TextToOnto n'est adossé à aucune méthode de construction ontologique. De plus, les mesures statistiques prises comme appui tendent à favoriser le phénomène dit « *long tail* » où les mots peu fréquents ont peu d'impact sur les concepts pouvant être calculés par ces méthode (notamment les *hapax*[3]). Dans le cadre des approches historiques de construction ontologique basées sur le corpus, le terme *élément remarquable* a été introduit. Ce terme désigne à la fois les termes, expressions fréquentes et apparaissant dans un corpus donné et les connaissances tacites contenues dans les textes. Ces connaissances tacites sont principalement les relations sémantiques (notamment les relations de subsomption et des relations spécialisées) dont les indices peuvent apparaître dans un corpus donné. Quant aux éléments explicites, ils peuvent indiquer la présence d'un concept. L'inconvénient de ce type de définition d'*élément remarquable* est qu'elle repose sur les indices répertoriés pour une langue donnée notamment lorsqu'il s'agit d'utiliser des ensembles de patrons lexico-syntaxiques pour leur détection. Les critères quantitatifs sont privilégiés et il est difficile de qualifier ces éléments du point de vue sémantique de manière simple et portable entre les langues.

3.2 Synthèse de la méthode proposée

La méthode proposée est construite autour de la notion de projection d'un modèle donné sur le RLSM_{PI} afin d'en extraire une ressource médiatrice destinée à être utilisée par un expert humain. La qualité de cette ressource reflète l'adéquation du RLSM_{PI} par rapport au contexte applicatif du modèle projeté.

La méthode diffère des méthodes traditionnellement utilisées pour la construction d'ontologie par sa définition de ce qu'est un ***élément remarquable***. Dans le cadre de notre approche, nous proposons la définition de l'élément remarquable en le considérant en tant que terme ou relation du RLSM_{PI} muni des caractéristiques particulières.

Définition 3.1

2. `https://sourceforge.net/p/texttoonto/wiki/Home/`

3. Fait de langue (mot, expression, construction) dont il n'existe qu'une seule occurrence dans un corpus donné.(Larousse, `https://www.larousse.fr/dictionnaires/francais/hapax/39017`.)

> Un ***élément remarquable*** est un terme, une relation ou une structure sémantique qualifiée et qualifiante.
>
> — *qualifié* se réfère à une possibilité de décrire cet élément de manière discrète (en termes d'éléments discrets présents au sein du $RLSM_{PI}$). S'il s'agit d'un terme, il doit posséder un nombre de relations entrantes important (avoir un rôle conceptuel). S'il s'agit d'une relation, elle doit être contextualisée via un mécanisme d'annotation des relations. S'il s'agit d'une structure, elle doit être repérée un nombre de fois suffisant dans le réseau.
>
> — *qualifiant* se réfère à la possibilité d'utiliser l'élément remarquable dans le cadre d'inférence endogène. S'il s'agit d'un terme, il doit avoir dans son voisinage des hyperonymes, des hyponymes et/ou des synonymes. Il doit également être couvert par le pivot et avoir des termes correspondants dans d'autres langues du $RLSM_{PI}$ ainsi qu'au niveau interlingue. S'il s'agit d'une relation, elle doit être non unique (il doit exister d'autres relations du même type réelles ou inférables dans le réseau). S'il s'agit d'une structure, ses termes et relations doivent être qualifiants.

Cette définition d'élément remarquable établie sur des critères sémantiques permet de décrire notre approche comme une approche non différenciée et basée sur la connaissance. En effet, aucune distinction n'est faite entre le terme et le concept au sein du $RLSM_{PI}$. La conceptualisation s'appuie sur l'identification et le calcul des éléments remarquables à partir des éléments présents au sein du $RLSM_{PI}$.

Lorsque l'on dispose d'une ontologie de référence déjà construite, il s'agit dans le cadre de nos expérimentations de l'« immerger » dans le $RLSM_{PI}$ pour l'enrichir notamment grâce à la proposition automatique d'une folksonomie (ce processus sera décrit dans la section 3.3). Les éléments de l'ontologie de référence sont ainsi modélisés en tant que termes, relations et structures du $RLSM_{PI}$. L'ontologie de référence apparaît alors comme une source de connaissance et comme un modèle qui permet une projection.

Si l'on choisit de représenter les structures sémantiques qui correspondent potentiellement à des structures ontologiques d'une ontologie de référence sous forme de règles (prémisses et conclusion), l'élément remarquable serait une instance valide de cette règle (soit structure sémantique qui vérifie les prémisses et permet de fournir la conclusion [4]).

4. Il est possible d'obtenir une règle en se basant sur de multiples occurrences de celle-ci.

3.3 Immersion

Le mécanisme d'immersion de l'ontologie de référence dans le RLSM_{PI} s'appuie sur un ensemble de règles qui servent à définir les correspondances. De telles règles ont été définies manuellement dans le cadre de nos expérimentations. Cependant, une génération partiellement ou entièrement automatique des règles de mise en correspondance peut être possible notamment dans le cadre des ontologies qui utilisent des vocabulaires standard [5]. Il prend en entrée l'ontologie de référence et l'ensemble de règles et fournit à la sortie une action : inférence des termes et des relations dans le RLSM_{PI}.

Les règles de mise en correspondance mobilisent les notions de *classe d'ontologie* et de *terme de réseau lexico-sémantique* et ont la forme telle qu'elle est décrite dans la définition 3.2.

Définition 3.2

Si x et y sont respectivement domaine et co-domaine d'une propriété à valeur objet p de l'ontologie de référence et y est sous-classe de C, alors x a une relation R avec y et y a une relation *is-a* avec C dans le réseau lexico-sémantique d'immersion (RLSM_{PI}).

$$O \cap RLSM_{PI} = \varnothing$$

$$x, y \in O \wedge x, y \in RLSM_{PI}$$

$$\forall x, R_p(x, y) \wedge subClassOf(y, C)$$

$$\Rightarrow R_t(x, y) \wedge isa(y, C)$$

Accessoirement, la relation créée au sein du RLSM_{PI} peut être annotée.

$$a \in PI \wedge annotation(a, R) \wedge covers(a, H) \wedge isa(y, H)$$

Les prémisses et la conclusion de la règle d'immersion portent sur les ensembles d'éléments disjoints qui appartiennent respectivement à l'ontologie à enrichir et au RLSM_{PI}.

Exemple 3.1

Illustration du processus de mise en correspondance.

$$poulet\ basquaise \xrightarrow{aPourProduitInitial} poulet$$

5. C'est-à-dire, les ontologies qui utilisent uniquement des propriétés déjà définies dans des vocabulaires pré-existants tels que RDFS, FOAF, SKOS et dont la sémantique est accessible dans un format qui peut être lu par une machine.

$$\wedge\ poulet \xrightarrow{subClassOf} viande\ type\ poulet$$

$$\Longrightarrow poulet\ basquaise \xrightarrow{r\_has\_partl} poulet$$

$$\wedge\ poulet \xrightarrow{r\_isa} viande\ type\ poulet$$

$$\wedge\ poulet\ basquaise[r\_has\_part]poulet \xrightarrow{r\_annotation} in : raw\ product$$

La relation ainsi obtenue est directement créée via la fonction décrite dans l'annexe A (fonction 9.

Le prémisses de cette règle de mapping reposent sur la contextualisation des relations présentés dans un réseau lexico-sémantique. La possibilité de contextualiser la relation est liée à la présence dans le voisinage des termes (objets lexicaux) du RLSM_{PI} qui modélisent les structures suivantes :

— *structure d'appartenance* : ensemble d'hyperonymes des termes source et cible de la relation ainsi qu'éventuellement le terme qui sert à annoter la relation ;

— *structure de contextualisation* : ensemble de relations sémantiques des termes source et cible de la relation ;

— *structure de liaison* peut également être distinguée pour désigner un ensemble de types de relations pouvant relier le terme source et cible de la relation. La structure de liaison peut être réduite à une simple relation sémantique (par exemple, *pétrir* $\xrightarrow{r\_object}$ *pâte* $\wedge$ *pétrir* $\xrightarrow{r\_isa}$ *technique de base* $\wedge$ *pâte* $\xrightarrow{r\_isa}$ préparation). Elle peut également concerner plusieurs types de relation (ce qui est souvent le cas des relations de méronymie partie-tout notée *r_has_part* par opposition à substance notée *r_matter)*[6].

Ces différentes structures sont détaillées sur la figure 3.1.

En pratique, il s'agit des sous-ensembles de l'ensemble des relations sémantiques d'un terme donné.

Si l'on s'intéresse à la modélisation dans le but de construction d'une ressource terminologique, il serait également possible de distinguer une *structure de lexicalisation* (terme donné, ses variantes, ses synonymes stricts).

Quel que soit le type de structure, les lexicalisations des termes peuvent dépasser les limites d'une seule langue. Les termes qui se trouvent dans le voisinage d'un terme donné peuvent se trouver dans le sous-graphe interlingue et être dépourvus de lexicalisation.

Au sein de l'ontologie de référence, la représentation des classes d'ontologie est principalement une **représentation « implicite »**. Conformément au docu-

6. Par exemple, *tarte aux pommes* $\xrightarrow{r\_has\_part}$ *pomme* par opposition à *tarte aux pommes* $\xrightarrow{r\_matter}$ *sucre.*

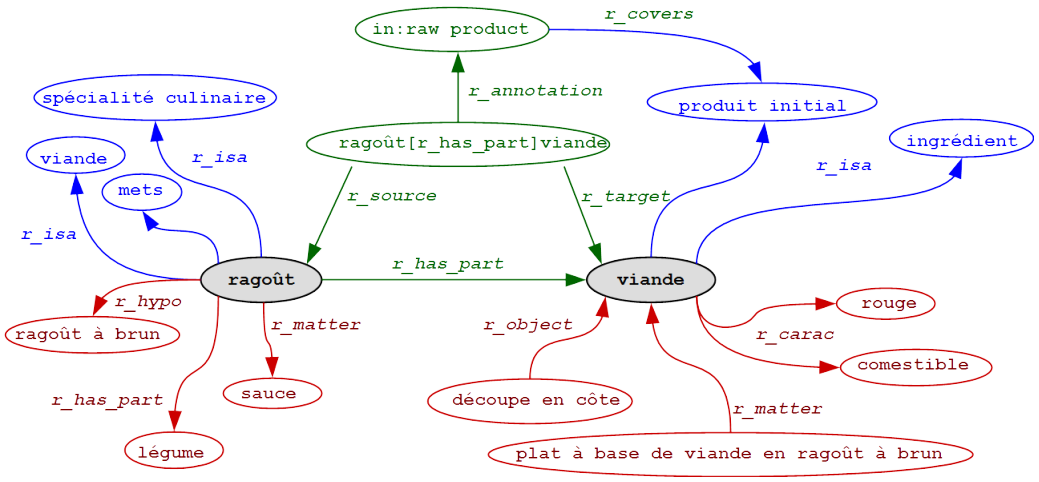


FIGURE 3.1 – Structures d'appartenance (en bleu), de contextualisation (en rouge), de liaison (en vert)

.

ment ISO 1087-1 [7], le concept est une « unité de connaissance créée par la combinaison unique des caractères ». Le concept peut être ***désigné*** par un terme, mais aussi ***défini*** :

— par un énoncé descriptif (intention) ;

— par extension (énumération de l'ensemble de ses individus) ;

— par compréhension (énumération de l'ensemble de ses super-classes).

À titre d'exemple, la classe dénotée *agrume* dans le cadre de l'ontologie de référence est modélisée sous forme d'une structure telle qu'elle est représentée sur la figure 3.2.

```
<owl:Class rdf:about="&aliment;Agrume">
<rdfs:label xml:lang="fr">agrume</rdfs:label>
<rdfs:subClassOf rdf:resource="&aliment;Fruit"/>
<miam:aPourCodeLangual>B1139</miam:aPourCodeLangual>
<rdfs:comment xml:lang="fr">Les agrumes sont les fruits
des végétaux des genres Citrus, Fortunella,
 Microcitrus, Eremocitrus et Poncirus...
[Wikipédia 2012]</rdfs:comment>
<rdfs:isDefinedBy rdf:resource="http://miam.org/ontology/aliment"/>
</owl:Class>
```

FIGURE 3.2 – Exemple de la classe d'ontologie de référence *agrume* (module *Aliment*).

Lors du parcours d'une ontologie, l'élément clé est le nœud (classe d'ontologie) auquel toutes les informations sont attachées. Par ailleurs, la hiérarchie explicitement représentée comporte une partie « cachée » qu'il est possible d'obtenir

7. https://edisciplinas.usp.br/pluginfile.php/312608/mod_resource/content/1/ISO_1087-1_2000_PDF_version_\%28en_fr\%29_CPDF.pdf

par raisonnement. Par exemple, il est possible d'avoir la classe *blanquette de veau* `aPourComposant` *ingrédient* et il est possible d'obtenir la liste des ingrédients (individus qui modélisent les produits quantifiés par exemple *1 œuf*, *200 grammes de oignon* etc.) en explorant les individus typés qui étendent la classe *ingrédient* et les classes filles de la classe *aliment* en fonction des propriétés et des restrictions fournies par l'ontologie.

Par opposition, un réseau lexico-sémantique fournit une **représentation « explicite »** des relations. Celles-ci sont au centre du parcours de ce type de ressource de connaissance. Lorsqu'une ontologie de référence est immergée dans le RLSM, ses étiquettes qui, dans le cadre de construction ontologique, participent à la dénotation formelle, deviennent nœuds du réseau et sont traités comme de simples termes (exemple 3.2). Ces classes apparaissent alors sous une forme explicite.

Exemple 3.2

`agrume-r_pos->in:Noun`

relations interlingues :

```
in:citrus fruit/noun/-r_covers->agrume
in:citrus/noun/-r_covers->agrume in:citrus-r_covers->agrume
```

relations hiérarchiques :

```
agrume-r_isa->arbre fruitier
agrume-r_isa->fruit
agrume-r_isa->ingrédient de cuisine
agrume-r_isa->produit végétal
agrume-r_isa->plante
agrume-r_isa->produit initial
agrume-r_isa->aliment
agrume-r_isa->concept aliment
citron/fruit/-r_isa->agrume
lime/citron vert/-r_isa->agrume
limette-r_isa->agrume
citron-r_isa->agrume
cédrat-r_isa->agrume
etc.
```

relations partie-tout :

```
agrume-r_has_part->peau(fruit)
agrume-r_has_part->chair(pulpe)
agrume-r_location->presse-agrume
etc.
```

caractéristique, lieu :

```
agrume-r_carac->comestible
agrume-r_carac->acide
etc.

agrume[r_matter]arôme d'agrume-r_source->agrume
agrume[r_matter]arôme d'agrume-r_target->arôme d'agrume
agrume[r_matter]arôme d'agrume-r_annotation->in:aroma
consommer-r_object->agrume
éplucher(peler)-r_object->agrume
presser-r_object->agrume
```

Propriétés de l'ontologie de référence encodées dans le format RLSM :

```
agrume-r_matter->arôme d'agrume agrume-r_matter->arôme fruité
confiture de tomates vertes aux agrumes-r_matter->agrume
```

Relations de raffinement :

```
agrume-r_refinement->agrume(fruit)
agrume-r_refinement->agrume(produit)
agrume-r_refinement->agrume(arbre)
```

Les termes issus de l'intégration de la hiérarchie de l'ontologie de référence bénéficient de la sémantique du RLSM_{PI} :

— ***par coïncidence***. Les labels peuvent coïncider avec un terme qui existe tel quel en langue naturelle. Dans ce cas, le nœud correspondant du RLSM se trouve dans l'intersection entre le lexique d'une langue naturelle et l'ensemble des labels d'ontologie. La taille de l'intersection entre les termes issus de MIAM et les termes faisant partie du sous-graphe français du RLSM est de 3 930 (l'ensemble des étiquettes de MIAM compte 8 065 étiquettes tandis que le sous-graphe français du RLSM_{PI} compte 268 017 termes au moment de l'expérience) ;

— ***par composition***. Des labels tels que *unité mesure capacité* ou *filet de volaille type dinde* bien que absents du lexique d'une langue naturelle peuvent subir une analyse sémantique compositionnelle qui explorerait les termes simples (y compris les entités polylexicales) qui les composent afin d'expliciter la sémantique qui peut leur être associée. Environ 4 135 termes issus de l'ontologie de référence ont été intégrés dans le sous-graphe français du RLSM_{PI} grâce à la décomposition en sous-chaînes et à l'exploration des relations sémantiques des composants. Pour les étiquettes d'ontologie ne comportant pas de connecteurs ou marqueurs syntaxiques, RLSM_{PI} est un support d'analyse approprié car il permet (si nécessaire) de restituer les dépendances entre les termes via une liste de règles et, par dessus tout, de **connecter ces termes dans le réseau**.

Exemple 3.3

Par exemple, pour le terme issu de l'ontologie de référence *unité mesure capacité*, l'analyse compositionnelle dans le but d'inférer des relations sémantiques se déroule comme suit :

Pour le terme *unité* nous retrouvons dans le RLSM_{PI} les relations suivantes :

unité $\xrightarrow{r\_refinement}$ {*unité (unité de mesure), unité (élément), unité (arithmétique), unité (militaire)*} ;

unité de mesure $\xrightarrow{r\_hypo}$ {*pied, cuillère, calorie,..., tasse*} ;

Pour le terme *mesure*, nous avons :

mesure $\xrightarrow{r\_hypo}$ {*tasse, cuillère, cuillère à soupe, canette, bar, pied, mètre*} ;

mesure $\xrightarrow{r\_agentive\_implication}$ {*contrôler, mesurer, doser, sonder, peser*} ;

ainsi que dans le graphe anglais :

measure $\xrightarrow{r\_hypo}$ {*volume, measuring cup, metrical foot, scale, quantification*} ;

Cette énumération partielle des relations laisse entrevoir une proximité entre les termes *unité(unité de mesure)* et *mesure*. Calculée d'après indice de Jaccard, elle est estimée à 0,44 ce qui correspond empiriquement à une proximité sémantique assez importante (au cours de nos expériences, le seuil de fiabilité est situé autour de 0,30 pour les termes dont le degré (entrant et sortant) d est supérieur à 100.

Pour le terme *capacité* :

capacité $\xrightarrow{r\_isa}$ *qualité*

mais aussi dans le graphe anglais :

capacity $\xrightarrow{r\_isa}$ *indefinite quantity* $\xrightarrow{r\_isa}$ *quantity*

Nous gardons l'intersection entre le voisinage des termes *unité* et *mesure* et le voisinage du terme *capacité* connecté à l'intersection *unité/mesure* comme ensemble de relations à inférer pour le terme *unité mesure capacité*.

Propriétés à valeur objet(*Object Properties*). Les propriétés à valeur objet sont des relations transversales de l'ontologie de référence. Chaque propriété porte sur un ensemble de départ (domaine) et un ensemble d'arrivée (codomaine). Sur le plan logique, il s'agit de *rôles à deux variables*.

Basé sur la logique de description, le modèle de l'ontologie de référence exploite le standard OWL de représentation de connaissances. Il s'agit d'un langage de représentation puissant qui reste décidable [8] car ses axiomes et ses constructeurs sont restreints.

8. Autrement dit, on peut la démontrer ou démontrer sa négation.

Algorithme 1 : Connexion des étiquettes d'ontologie dans un réseau lexico-sémantique

input : $RLSM_{PI}$, étiquette_ontologie, langue_source, langue_cible
output : Relations[]

```
// initialisation
Relations[]=∅ ;
Terme t ;
demi-relations[]← ∅ ;
CréerTerme (étiquette_ontologie) ;
for chaîne ∈ Ngramme (étiquette_ontologie) do
    t←RechercherTerme (chaîne) ;
    demi-relations.Insérer (RechercherRelations (t)) ;
Relations[]←Fusionner (demi-relations) ;
demi-relations← demi_relations ∨ Relations[] ;
for relation_restante ∈ demi-relations do
    for relation_ajoutée ∈ Relations[] do
        if SemantiquementLié (relation_restante,relation_ajoutée) then
            Relations[] ← Relations[] ∪ relation_restante ;
retourner Relations[]
```

L'ontologie à enrichir, MIAM [9] contient en tout 93 propriétés à valeur objet organisées sous forme d'une hiérarchie de propriétés. L'ontologie compte 21 565 instances des propriétés de ce type. Après l'immersion de l'ontologie de référence dans le $RLSM_{PI}$, il est possible de considérer, de manière quelque peu approximative, que nous disposons du même nombre de règles et d'instances de règles qui peuvent servir pour l'inférence des propriétés d'ontologie. Une approche naïve consisterait à mettre en place une simple inférence translingue par transfert, voir algorithme 2.

Cependant, il s'agit d'une technique peu productive et sujette aux erreurs dues à des problèmes d'alignement. En effet, l'immersion au sein de $RLSM_{PI}$ « transforme » de nombreux concepts formels d'ontologie en termes polysémiques. Des techniques de désambiguïsation ainsi que des approches d'alignement par sens en amont sont alors nécessaires. De plus, construite selon une méthode descendante par une communauté d'experts du domaine, l'ontologie de référence comporte un nombre variable d'instances par propriété.

Les instances peuvent également être absentes pour une propriété définie. Une approche naïve ne ferait que « reconduire » ce déséquilibre. Le tableau 3.1 montre les résultats de cette approche pour 11 propriétés clés de l'ontologie à enrichir.

9. http://www-limics.smbh.univ-paris13.fr/ontoMIAM/

Algorithme 2 : Algorithme naïf d'inférence des structures de type « propriété à valeur objet (*Object Property*) »

input : RLSM_{PI}, nom_de_propriété, patron,langue_source, langue_cible
output : Inférences[]

```
// initialisation
Inférences[]=∅ ;
type_relation← patron.type ;
for r ∈ RechercherRelations (langue_source,type,patron.annotation) do
    candidat_source ← ObtenirTraduction (r.source, POS) ;
    candidat_cible ← ObtenirTraduction (r.cible, POS) ;
    if SémantiquementLié (r.source,r.cible) then
        Inférences[]← Inférences[] ∪ {(candidat_source,
          nom_de_propriété,candidat_cible)} ;
Retourner Inférences[]
```

mod	propriété	#inst	$\mathbf{I}_{en}$	$\boldsymbol{Pr}_{en}$	$\boldsymbol{V}_{en}$	$\mathbf{I}_{es}$	$\boldsymbol{Pr}_{es}$	$\boldsymbol{V}_{es}$
A	aPourProduitInitial	1 844	79	4,23%	60%	52	2,79%	40%
A	aPourProduitConstituant	98	2	2,04%	10%	-	-	-
A	aPourProduitAromatisant	41	9	21,95%	10%	-	-	-
A	aPourEtatPhysique	226	18	7,96%	25%	12	5,31%	90%
P	aPourEtatTransformé	289	183	63,32%	90%	167	57,79%	50%
A	aPourTypeLait	390	16	4,5%	2%	12	3,08%	3%
P	aPourComposantPrincipal	16	1	6,25%	98%	-	-	-
S	aPourConsistance	499	53	10,62%	15%	72	14,43%	-
S	aPourAspectSurface	176	90	51,14%	0	187	106,25%	-0
A	aPourLabel	114	2	1,75%	68%	-	-	-
A	aPourSaison	31	10	32,26%	4%	9	29,03%	99%

TABLE 3.1 – Résultats approche naïve (propriétés clés du module *Aliment* (A), *Préparation* (P), *Sensoriel* (S). Le tableau donne le détail des résultats en termes du nombre d'instances des propriétés (#inst), du nombre d'inférence produites pour une langue I_{lang} (nous avons exploré l'anglais et l'espagnol), de la productivité du moteur d'inférences naïf pour une langue donnée Pr et du taux de validité des inférences produites V (suite à l'évaluation manuelle des candidats).

Compte tenu de ces résultats préliminaires, la consolidation du RLSM_{PI} est devenue centrale pour notre démarche. Les structures que nous avons définies au début de la présente section nécessitent une ressource stable et harmonisée autant que possible notamment en ce qui concerne l'alignement par sens via le pivot interlingue.

Propriétés à valeur donnée (*Data Properties*). Les propriétés à valeur donnée (exemple 3.4) représentent un intérêt assez limité dans le cadre de nos expériences.

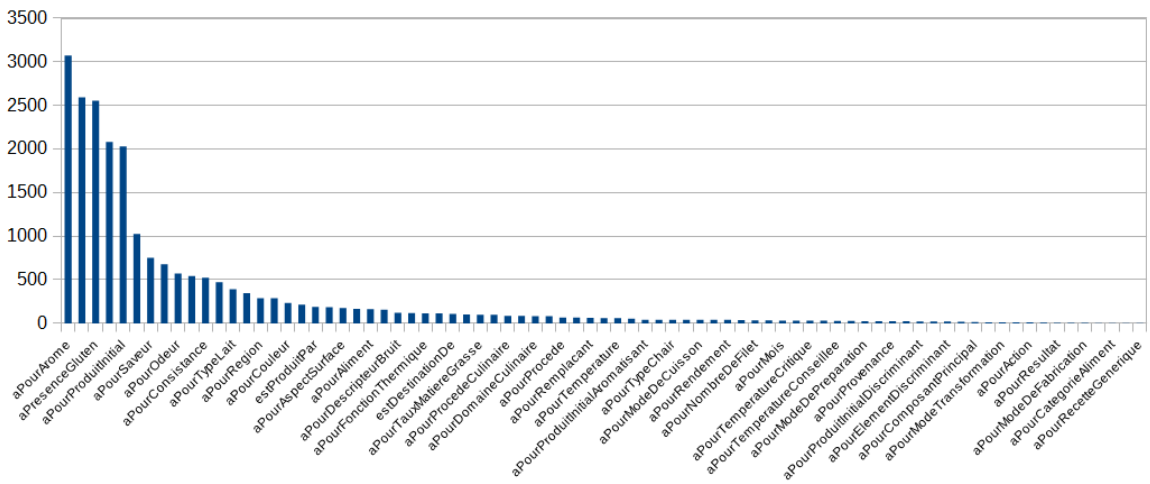
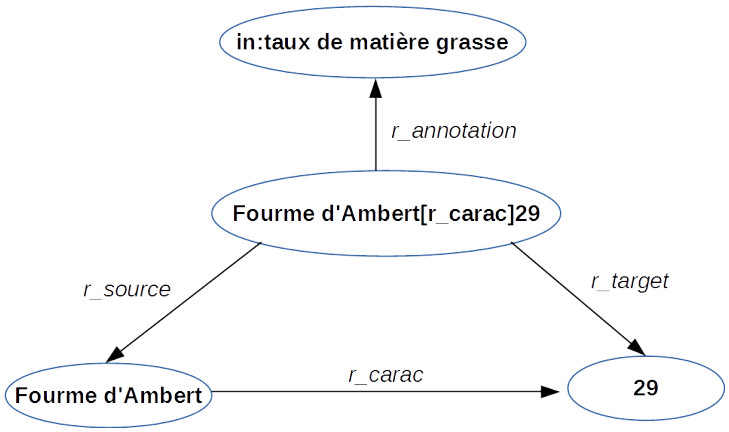


FIGURE 3.3 – Distribution du nombre de relations (instances de propriété) par propriété d'ontologie MIAM tous modules confondus. La composition (dont la présence de allergènes et les caractéristiques sensorielles des aliments sont les groupes de propriétés le plus renseignés par les experts lors de la constitution de l'ontologie).

Exemple 3.4

Par exemple, la propriété à valeur donnée `aPourTauxMatiereGrasse`, peut être représentée comme suit pour le terme *fourme d'Ambert*.



« *Fourme d'Ambert[r_ carac]29* » est l'étiquette (chaîne de caractères) de la réification sous forme de terme de la relation à annoter.

Certaines de ces propriétés notamment celles où le type des valeurs du co-domaine est « booléen » s'apparentent a une simple relation sémantique. Par exemple, la propriété `aPresenceGluten` serait encodée $x \xrightarrow{r\_matter} gluten$. D'autres propriétés à valeur donnée qui concernent les types tels que « integer » sont intégrées sous forme de relations annotées.

L'encodage des classes et des relations d'ontologie donne lieu aux problématiques suivantes :

— gestion des structures d'appartenance : place des principaux concepts [10] d'ontologie au sein du $RLSM_{PI}$ (sous-graphe interlingue) ;

— contextualisation : définition de la structure de contextualisation ;

— représentation des qualités et des quantités (quantifieurs et ingrédients quantifiés).

La réponse à ces problématiques peut être issue d'une *décision d'implémentation*. Elle peut également s'imposer *par émergence*, à savoir, à partir de l'analyse des données et via les processus qui utilisent le $RLSM_{PI}$.

3.4 Découverte des éléments remarquables par inférence

3.4.1 Principe de l'abduction.

La définition générale de l'abduction peut être formulée comme *raisonnement par lequel on restreint le nombre d'hypothèses pouvant expliquer un phénomène observé*. Appliqué à un réseau lexico-sémantique tel que le $RLSM_{PI}$ tout comme dans le cas de RezoJDM (Lafourcade [2007]), le mécanisme d'***inférence par abduction*** se base sur « le partage de certaines relations sortantes entre les termes » [11]. Ainsi, au moyen d'un ensemble de contraintes d'ordre statistique (le nombre de relations sortantes en commun, la moyenne des poids des relations sortantes) les relations détenues par un terme sont proposées à des termes jugés similaires.

Dans le cadre de nos expériences concernant l'exploitation du $RLSM_{PI}$ pour la construction d'ontologie, notre but est de trouver des termes, des relations et des structures susceptibles d'être des éléments remarquables (conformes à la définition 3.1) pouvant enrichir une ontologie donnée. Ce contexte de découverte des éléments remarquables nous place dans la recherche des éléments qui correspondent à des concepts et à des relations d'ontologie. Par conséquent, à l'intérieur de l'ensemble de leur relations sortantes, leur structures d'appartenance et de contextualisation (figure 3.1) doivent être *présentes* et *similaires* avec celles des termes qui représentent les classes d'ontologie MIAM. Autrement dit, un terme-candidat doit posséder des relations hiérarchiques (typées *r_isa* et *r_hypo*) et partager des relations sémantiques sortantes typées *r_has_part*, *r_matter*, *r_location*, *r_carac*, *r_object-1* [12] etc. avec un terme dont l'étiquette correspond à la lexicalisation d'un concept MIAM.

10. Parmi les principaux concepts de l'ontologie de référence MIAM, on retrouve les classes Aliment, Produit, AlimentQualifié, EtatProduit etc.

11. Zarrouk [2015], p. 62.

12. Il s'agit d'une relation qui permet de représenter une relation qu'entretient le patient d'une action avec l'agent qui produit cette action.

La découverte des éléments remarquables de type « classe d'ontologie » et « propriété d'ontologie » diffère du point de vue de la sélection des termes à comparer. Lorsqu'il s'agit de découvrir des éléments de type « classe », il faut comparer les termes voisins dans une chaîne hiérarchique. L'un des termes à comparer doit correspondre à une classe de l'ontologie MIAM, immergée dans le RLSM_{PI}. Lorsqu'il s'agit de découvrir les éléments de type « propriété », l'un des éléments à comparer doit correspondre à une structure (souvent, relation annotée) qui représente une propriété réellement existante dans l'ontologie MIAM et immergée au sein du RLSM_{PI}.

Dans un contexte de construction ontologique classique, les relations hiérarchiques servent à modéliser la hiérarchie des classes. Une classe correspond à un ensemble d'individus possédant les mêmes caractéristiques. Elle peut être définie par énoncé, par extension, par compréhension ou désignée par un terme. Dans le cadre de l'exploitation d'un RLSM pour la construction termino-ontologique, la définition par énoncé descriptif s'apparente à la définition par l'ensemble des relations sémantiques sortantes, la définition par extension - à la définition par l'ensemble des hyponymes, la définition par compréhension - à la définition par l'ensemble des hyperonymes et leur sémantique.

Du point de vue de la lexicalisation, il s'agit principalement de l'acquisition ou de modélisation des éléments remarquables qui correspondent à des classes et individus désignés par un terme du lexique. Par ailleurs, il est possible de rechercher des structures autour d'un élément inférable non lexicalisé dans une langue donnée et acquérir ainsi les définitions potentielles des concepts potentiels par compréhension (via l'ensemble de hyperonymes), extension (via un ensemble de hyponymes) ou par description (via l'ensemble des relations sémantiques hors taxonomie). Un élément « non lexicalisé » indique une différence de granularité entre les langues. Un mécanisme d'inférence interlingue par abduction peut être mis en place pour les termes issus de l'immersion de l'ontologie MIAM.

Exemple 3.5

Dans la langue russe, il n'existe pas de terme spécialisé qui permettrait de dénoter le concept « pâte brisée ».

Ce concept est représenté comme suit dans l'ontologie MIAM :

```
<owl:Class rdf:about="&preparation;PateBrisee">
<rdfs:label xml:lang="fr">pâte brisée</rdfs:label>
<rdfs:subClassOf rdf:resource="&preparation;PateAFoncerEtBrisee"/>
<rdfs:isDefinedBy rdf:resource="http://miam.org/ontology/preparation"/>
</owl:Class>
```

Cette classe est disjointe avec la classe `PateAFoncer`, `PateARissoles`.

Au sein du RLSM_{PI}, dans la partition *fr* le terme *pâte brisée* possède un ensemble de relations sémantiques suivant (vue partielle avec redondances caractéristiques du RLSM_{PI}) : pâte brisée $\xrightarrow{r\_isa}$ {*pâte, préparation culinaire, ingrédient de recette de cuisine, ingrédient de cuisine*}

pâte brisée $\xrightarrow{r\_has\_part}$ {*œuf(alimentation), jaune d'œuf*}

pâte brisée $\xrightarrow{r\_matter}$ {*gluten, farine, matière grasse végétale, beurre, eau froide, sel*}

{*dérouler, consommer, fraiser*} $\xrightarrow{r\_object}$ *pâte brisée*

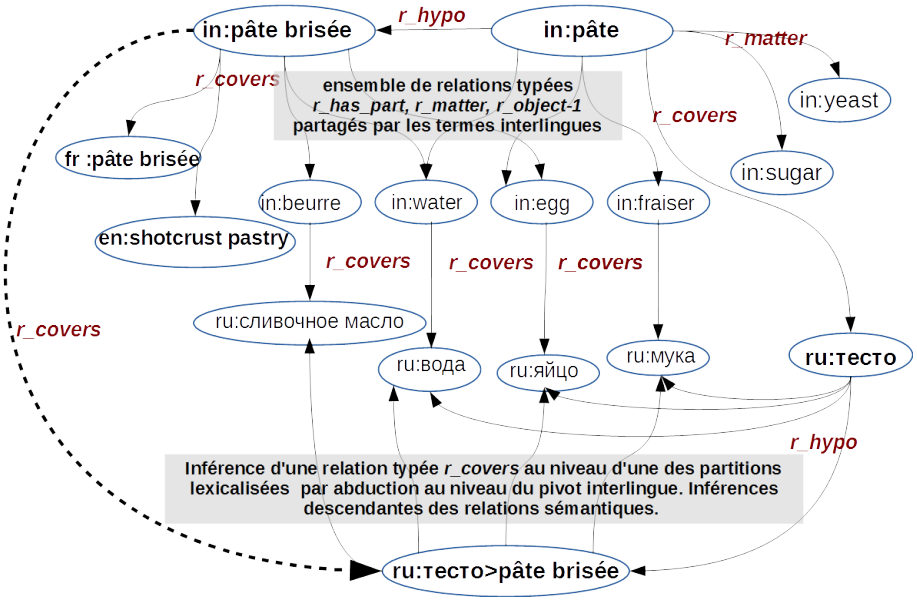
{*tartelette, tarte tatin*} $\xrightarrow{r\_has\_part}$ *pâte brisée*

pâte brisée $\xrightarrow{r\_domain}$ *cuisine*

pâte brisée $\xrightarrow{r\_location}$ *moule à tarte*

{régime hypocalorique, régime sans œuf} $\xrightarrow{r\_incompatible}$ *pâte brisée*

On souhaite exprimer le concept de la *pâte brisée* en russe car on constate une nombre de recettes qui partagent ingrédients et processus qui correspondent à la préparation de la pâte brisée. Des relations typées *r_ covers* peuvent être inférées par abduction interlingue. Les relations sémantiques pertinentes sont ensuite inférées.



3.5 Découverte des éléments de type « classe » et « individu »

Le principe de l'abduction peut être exploité pour la découverte des éléments remarquables de type « classe d'ontologie » (concept d'ontologie). Dans un premier temps, il s'agit d'identifier les chaînes hiérarchiques sémantiquement valides au sein du RLSM_{PI}. Cette étape inclut le filtrage statistique puis logique visant

à désambiguïser les termes qui constituent la chaîne et s'assurer que les termes voisins dans une chaîne donnée sont sémantiquement liés.

Hiérarchie des classes de l'ontologie de référence (MIAM) Dans le cadre du RLSM_{PI}, les relations hiérarchiques sont modélisées sous forme de relations typées *r_isa* ainsi que sous forme de relations typées *r_hypo*. Compte tenu des techniques utilisées pour le peuplement (Chapitre 2) qui reflètent la variété de méthodes pouvant être mises en oeuvre dans un contexte industriel de gestion des contenus multilingues, *la nature de ces relations hiérarchiques n'est pas homogène.*

Ainsi, dans le cadre de la sélection des termes candidats pour l'abduction, il est indispensable de sélectionner un sous-ensemble de chaînes hiérarchiques. Autrement dit, considérer les axiomes de l'ontologie de référence (définition 3.3 et la sémantique des termes RLSM_{PI} issus de MIAM et sélectionner toutes les chaînes hiérarchiques qui remontent vers les concepts MIAM et dans lesquelles la consolidation des hiérarchies correspond à une « mise en conformité » des relations hiérarchiques présentes au sein du RLSM_{PI} avec les axiomes de l'ontologie de référence désormais immergée dans le réseau. Cette mise en conformité prend la forme d'ajout de méta-information aux relations typée *r_isa* et *r_hypo* afin de pouvoir les identifier.

Définition 3.3

Dans le cadre de l'ontologie MIAM, l'**axiome général** est de forme suivante :

```
<rdf:Description>
<rdf:type rdf:resource="&owl;AllDisjointClasses"/>
<owl:members rdf:parseType="Collection">
<rdf:Description rdf:about="&aliment;AB"/>
<rdf:Description rdf:about="&aliment;AppelationOrigine"/>
<rdf:Description rdf:about="&aliment;IGP"/>
<rdf:Description rdf:about="&aliment;LabelRegional"/>
<rdf:Description rdf:about="&aliment;LabelRouge"/>
<rdf:Description rdf:about="&aliment;STG"/>
<rdf:Description rdf:about="&aliment;SpecialiteOrigine"/>
</owl:members>
</rdf:Description>
```

Les axiomes concernent la disjonction entre les classes de MIAM. C'est-à-dire, dans notre exemple, une instance d'aliment ne peut pas appartenir à plusieurs classes disjointes à la fois. Cette disjonction garantit la consistance de l'ontologie.

Si l'on souhaite « traduire » les axiomes générales de MIAM en termes de RLSM_{PI}, il est possible de considérer les étiquettes des classes listées dans les

axiomes afin de pouvoir identifier les différents critères (caractéristiques) sémantiques selon lesquels ces disjonctions auraient pu être faites.

Après avoir analysé un sous-ensemble des axiomes manuellement, nous avons pu faire ressortir les catégories suivantes et faire le rapprochement entre les *axiomes* MIAM et les *types de relations* RLSM_{PI} :

— catégories basées sur l'appartenance, par exemple, appartenance à un label (agriculture bio, indication géographique protégée etc.) : *r_ has_ part* ;

— catégories basées sur la transformation *r_ carac* (aliment découpé) ;

— catégories basées sur la composition : *r_ matter* (aliment *à base de poisson*) ;

— catégories basées sur le type/appartenance à une catégorie : *r_ hypo* (volaille type dinde).

Ce rapprochement permet d'identifier les types de relations à considérer dans le cadre d'inférence par abduction.

Découverte des structures hiérarchiques par abduction. Pour pouvoir découvrir par abduction des structures hiérarchiques similaires à celles faisant partie de hiérarchie de référence (hiérarchie MIAM), nous procédons comme suit :

1. nous explorons des ensembles de chaînes hiérarchiques dans lesquelles apparaît un terme t et qui doivent vérifier les propriétés suivantes :
   (a) absence de cycle (la chaîne hiérarchique est un *chemin*) ;
   (b) poids de chaîne suffisant (poids normalisé des relations strictement positif, supérieur ou égal à un seuil) ;
   (c) cohérence sémantique (présence des relations sémantiques entre les termes appartenant au voisinage des deux termes voisins dans une chaîne hiérarchique donnée).
2. nous identifions des triplets candidats où une de extrémités correspond à une classe ou individu issu de l'ontologie de référence et acquis par immersion et l'autre extrémité est un élément remarquable - candidat qui pourrait enrichir l'ontologie ou sa folksonomie.

Ainsi, chaque terme-candidat apparaît inclus dans une chaîne hiérarchique valide.

Le calcul sur les hiérarchies du RLSM est entravé par deux problématiques :

1. un ou plusieurs termes faisant partie d'une chaîne hiérarchique peuvent être polysémiques et de ce fait entraîner la concaténation des sous-chaînes hiérarchiques de façon erronée. Il est important de garder à l'esprit que, dans un contexte industriel la distinction de sens se met en place progressivement et peut être partiellement ou totalement absente ;

2. une fois immergés dans le RLSM_{PI}, les éléments (classes, individus) de l'ontologie de référence peuvent être peu connectés du fait de leur spécificité ce qui rend difficile leur définition en tant que éléments remarquables qualifiants.

Par conséquent, une validation de la chaîne hiérarchique est nécessaire quelle que soit la place de ce terme dans la chaîne.

La validation d'une chaîne donnée se déroule comme suit :

1. Le poids de la chaîne est calculé grâce à la fonction `PoidsDeChaine` qui prend en entrée tous les poids harmonisés des arcs dans une chaîne. Les chaînes dont une seule relation aurait un poids négatif ou insuffisant (inférieur à 50) sont éliminées. Pour les chaînes restantes, les poids sont harmonisés et la moyenne est calculée.

   Exemple : Pour le terme pain nous avons obtenu ***16 011 chaînes filtrées par poids****.*

2. La proximité sémantique est calculée comme indice de Jaccard simple (sans prise en compte des poids des relations), le seuil admis est fixé assez bas.

   Exemple : Pour le terme pain nous avons obtenu ***6 404 chaînes filtrées par proximité sémantique****.*

3. L'absence de cycle est vérifiée grâce à la fonction `PasDeCycle` qui renvoie un booléen qui correspond à la présence ou absence de cycles dans la chaîne.

4. La chaîne est évaluée en termes de présence des relations sémantiques (relation, chemin, sous-graphe) et lexicales (variante lexicale) entre ses composants voisins. Le cas général de validation d'une chaîne hiérarchique est présenté sur le schéma 3.4.

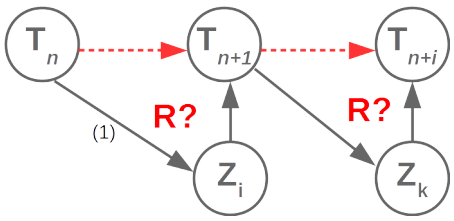


FIGURE 3.4 – Cas général de validation d'une chaîne hiérarchique.T_i sont des termes qui constituent la chaîne à valider. R correspond à une relation sémantique (non lexicale, non morpho-syntaxique,non ontologique) qui existerait entre les termes voisins de T_i dans la chaîne hiérarchique.

Le calcul a été effectué sur l'ensemble des termes correspondant aux 1 322 concepts concepts haut niveau de MIAM qui appartiennent au module *Aliment*. Au départ, nous avons obtenu 132 213 chaînes. Après filtrage par poids de la chaîne, cet ensemble a été réduit à 53 749 chaînes (40% de chaînes retenues par filtrage statistique). De nombreuses redondances existent à l'intérieur de cet ensemble car une chaîne peut contenir d'autres chaînes plus courtes. Par ailleurs, une chaîne hiérarchique peut n'être que partiellement validée. Le filtrage logique

a permis d'aboutir à un ensemble de chaînes validées qui contient 9 600 chaînes (18% de chaînes retenues par rapport à l'ensemble de chaînes pré-validées statistiquement et 7% par rapport à l'ensemble des chaînes non filtrées).

L'analyse et la validation des chaînes hiérarchiques constitue la partie la plus importante de l'algorithme de découverte des éléments remarquables de type « classe », « individu », « relation de subsomption ». Il s'agit d'une partie coûteuse en termes de ressources. La complexité de l'algorithme dépend de l'importance du concept en train d'être traité et de la longueur des chaînes à filtrer. Le degré typé $r\_isa \ \ d_{isa}$ le plus élevé étant de 5 264 (pour le terme *aliment*) et la longueur maximale l des chaînes hiérarchiques obtenues étant égale à 9, la complexité dans le pire des cas serait $O(d_{isa}^{l}$ soit $O(5\ 264^{9}) = 3,103436942 \times 10^{33}$.

Exemple 3.6

Exemples de chaînes hiérarchiques après filtrage sémantique :
pain de campagne→*pain*→***ingrédient de cuisine***→*aliment*
baguette complète→*pain complet*→*pain*→***ingrédient de recette de cuisine***→*aliment*
galantine→pâté
angélique→confiserie→bonbon

Le poids des chaînes se base sur le calcul du score de poids global qui sera détaillé dans le chapitre 4. Ici , il est intéressant d'analyser une corrélation entre la distribution des longueurs des chaînes et de leur poids.

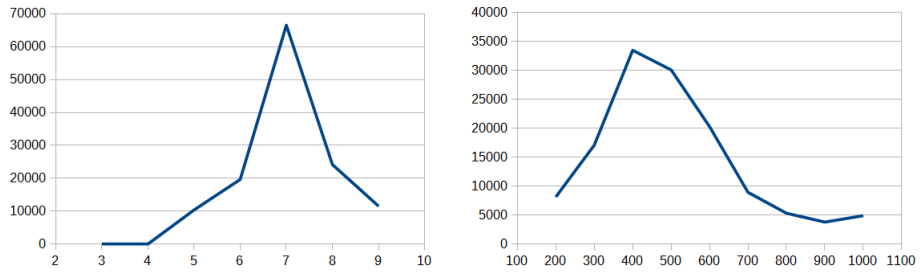


FIGURE 3.5 – Courbe de distribution des longueurs des chaînes (à gauche) comparés avec la distribution des poids des chaînes avant le filtrage (à droite). Même pour une ressource non stabilisée, l'utilisation des poids pour le calcul des chaînes hiérarchiques est pertinente.

En ce qui concerne les termes du RLSM_{PI} de type « classe d'ontologie » détectés par abduction, il s'agit des termes utiles pour la constitution des folksonomies car par ce procédé on récupère principalement les individus pouvant être intégrés dans l'ontologie de référence. Toutefois, nous constatons la répartition suivante entre les catégories des éléments acquis d'après les observations faites sur 500 candidats.

	appartenance	**transformation**	**composition**	**catégorie**
candidats	10	270	95	110
%candidats	2%	54%	19%	22%

TABLE 3.2 – Répartition des éléments remarquables candidats.

Exemple 3.7

Exemples produits :

Suggestion des individus des classes de l'ontologie de référence :

baguette de campagne `subClassOf` pain de campagne
angélique `subClassOf` confiserie
truffe>chocolat `subClassOf` chocolat
pomme douce amère `subClassOf` pomme
pomme à cidre `subClassOf` pomme
sucre de pomme `subClassOf` confiserie

Suggestion des « classes » et rapports hiérarchiques de l'ontologie de référence :

viennoiserie `subClassOf` pâtisserie
sucre vanillé `subClassOf` sucre en poudre

Le processus de découverte d'éléments remarquables de type « classe », « individu » ou « relation de subsomption » d'ontologie de référence a pu être quantifié comme suit :

— #**candidats** : nombre de relations candidates (triplets potentiels d'ontologie) ;

— #**valides** : nombre de relations valides (cet ensemble peut contenir des triplets

— **%valides** : taux de relations valides parmi l'ensemble des relations proposées ;

— #**nouveaux** : nombre de relations nouvelles parmi l'ensemble des relations valides proposées ;

— **%nouveaux** : taux de relations nouvelles parmi l'ensemble des relations valides proposées.

#candidats	#valides	%valides	#nouveaux	%nouveaux
11 520	11 289	98%	4 741	42%

TABLE 3.3 – Résultats de découverte par abduction des éléments remarquables de types « classe », « individu », « relation de subsomption » d'ontologie (expérience limitée au module *Aliment* et au sous-graphe français).

Algorithme 3 : Découverte des éléments de type « classe », « individu », « relation de subsomption » d'ontologie de référence MIAM.

input : terme, langue_source, langue_cible
output : Terme[]

```
// l'algorithme permet de proposer à la fois des classes candidates en
   langue cible (langue de l'ontologie de départ) et dans d'autres
   langues, langue_source == ou != langue_cible
// initialisation
Terme[] ← ∅ ; // liste de termes (classes candidates) pour la langue_cible
Candidats[] ← ∅ ; // liste de triplets candidats
foreach c ∈ ObtenirChaînesHiérarchiques (terme) do
    ValiderChaîne (c) ;
    Candidats[] ← ObtenirCandidats () ;
// utilisation du processus par règles pour l'inférence des classes
   d'ontologie
foreach candidat ∈ Candidats[] do
    // validation par abduction (en exploitant les exemples proches)
       dans la langue source, s'il existe suffisamment d'exemples
    if Match (candidat) then
        // recherche des structures proches dans la langue cible
```

3.6 Découverte des éléments remarquables de type « propriété d'ontologie »

Hiérarchie des propriétés de l'ontologie MIAM. Chaque module de l'ontologie de référence possède sa propre hiérarchie de propriétés. Lors de l'immersion dans le RLSM_{PI}, les propriétés ont été exprimées en termes de relations sémantiques contextualisées via l'ajout d'une annotation (comme détaillé dans la section 3.3). Le choix du type ou d'un ensemble de types de relations permet de distinguer les cas de figure suivants :

- *Object Properties* construites autour de la composition : `aPourProduitInitial`, `aPourProduitImplicite` ;
- *Object Properties* basées sur les relations spatio-temporelles : `aPourMoisPrimeur`, `aPourRegion` ;
- *Object Properties* explicitant les caractéristiques : `aPourEtat`, `aPourArome` ;
- *Object Properties* procédurales : `aPourMethodeDeConservation` ;

— *Object Properties* qui peuvent être définies par un sous-graphe spécifique [13] : `aPourAlimentAmi`, `aPourQualificateur`.

Découverte des éléments remarquables de type « Object Property »
Compte tenu de sa structure basée sur un ensemble de relations typées, des annotations et des prend la forme d'une règle. Fournie via un fichier statique de règles, générée automatiquement ou construite à la volée par un expert, la règle est donnée en entrée à l'algorithme de découverte des éléments remarquables de type « ObjectProperty ».

L'algorithme se déroule en deux temps. Dans un premier temps, la validité de la règle pour la langue de départ est vérifiée. L'objet `CorrespondanceRègle` est instancié dans la langue source à partir de la règle de conversion fournie en entrée. Cette règle à une forme générale suivante :

```
property=aPourEtatPhysique
source=?s
reltype=r_carac
target=?o
target_isa=état physique,état
annotation=int:physical state/
source_isa=aliment,nourriture,préparation culinaire,mets
target_isa=[]
source_features=OUTGOING/r_pos/int:Noun/
target_features=OUTGOING/r_pos/int:Adj/
```

La fonction `CorrespondanceRègle` recherche un maximum de sous-graphes qui satisfont les critères de la règle.

— `property` correspond au nom de la propriété d'ontologie de référence ;
— `source` correspond à l'ensemble de termes qui constitue le *domaine* (ensemble de termes source) de la propriété ;
— `reltype` est le type de relation choisi pour encoder la propriété d'ontologie de référence dans le RLSM_{PI} ;
— `annotation` est le terme $\in$ PI utilisé pour contextualiser l'ensemble des relations de type `reltype` ;
— `target` correspond à l'ensemble de termes qui constitue le *co-domaine* (ensemble de termes cible) de la propriété ;
— `source_isa` correspond à l'ensemble d'hyperonymes pour l'ensemble source (définit la structure d'appartenance du terme source) ;
— `target_isa` correspond à l'ensemble d'hyperonymes pour l'ensemble cible(définit la structure d'appartenance du terme cible) ;

13. Nous entendons par sous-graphe, un sous-ensemble de termes du RLSM_{PI} connectés par des relations sémantiques.

— `source_features` correspond à l'ensemble de relations entrantes et/ou sortantes qui caractérisent les termes de l'ensemble source (définit la structure de contextualisation du terme source) ;

— `target_features` correspond à l'ensemble de relations entrantes et/ou sortantes qui caractérisent les termes de l'ensemble source (définit la structure de contextualisation du terme cible) ;

Si la règle à permis de détecter suffisamment de structures dans la langue de départ (au minimum 2 structures), elle est considérée comme valide et permet de générer un objet qualifiant. Grâce à cet objet, dans un second temps, des structures candidates sont détectées au sein du RLSM.

L'algorithme peut être résumé comme suit.

Algorithme 4 : Découverte des éléments de type « Object Property » à partir d'une règle.

input : RLSM_{PI}, langue_source, langue_cible, règle
output : $\text{Triplets}[]_{\text{propriété}}$

```
// initialisation
EnsemblePropriétés ← ∅ ;
CorrespondanceRègle ←false ;
EnsemblePropriétés ← AppliquerRègle (règle) ;
if CorrespondanceRègle (EnsemblePropriétés, langue_source)==true then
    Triplets[]propriété ← Inférer (CorrespondanceRègle, langue_cible) ;
retourner Triplets[]propriété
```

Dans le cadre de nos expériences, nous avons poursuivi les objectifs suivants :

1. construire des règles adéquates pour la découverte des structures de type Object Property ;
2. faciliter peuplement de certaines propriétés de l'ontologie MIAM ;
3. tester le modèle d'inférence des éléments remarquables translingue en allant au-delà de la méthode basée sur la simple traduction des étiquettes des termes.

Le traitement réservé aux aux éléments relatifs à *Object Property* (OP) et ceux relatifs à *Data Property* (DP) de MIAM laisse apparaître les cas de figure suivants :

— relation sémantique éventuellement annotée (cas des propriétés telles que `aPresenceLactose`, `aPresenceGluten`, `aTeneurLipide` ;

— patron spécifique composé de relation annotée dont les extrémités sont des termes enrichis (termes et leurs hyperonymes ainsi que éventuellement des relations sémantiques). Dans ce cas, on pose une règle suivante pour un terme source S, un terme cible C, un terme annotation A qui appartient à un ensemble de termes d'annotation utilisés dans MIAM *Annotations* (*arôme*, *produit initial*,*ingrédient principal* etc.), un type de relation *type* qui appartient à un ensemble de types prédéfinis *Types*, le terme - réification de la relation annotée *Reif* :

« si un terme C a un hyperonyme A qui appartient à l'ensemble de termes *Annotations* et une relation *Rel* entrante de type t qui appartient à l'ensemble *Types*, alors il est possible d'annoter *Rel* avec A et la structure résultante est une élément remarquable candidat » :

$$\exists\ S \xrightarrow{Rel_t} C \wedge \exists\ C \xrightarrow{r\_isa} A$$
$$\Rightarrow Reif \xrightarrow{r\_source} S \wedge R \xrightarrow{r\_target} C \wedge Reif \xrightarrow{r\_annotation} A$$

Une définition de S en termes de ses relations sémantiques est alors nécessaire. L'ensemble *Annotations* inclut les termes qui correspondent au noms des propriétés MIAM mais aussi les synonymes et éventuellement hyponymes et les termes couvrants de ces termes. Des contraintes particulières ont été proposées compte tenu des spécificités des différentes *Object Properties* de MIAM.

— structure plus complexe pour les propriétés procédurales, les propriétés telles que `aPourRemplacant`, `aPourAlimentAmi`

Pour les propriétés pour lesquelles il est possible d'obtenir es éléments remarquables sur la base des relations sémantiques, nous avons obtenu le résultat présentés dans le tableau 3.4 dans l'état actuel du RLSM_{PI}. Notre démarche

#DP	**#triplets MIAM**	**#élém RLSM_{PI}**	**filtrage**	**%aug**
aTeneurLipide	0	4 741	3 271	-
aPrésenceLactose	2 593	530	408	+16%
aPrésenceGluten	289	820	762	+263%

TABLE 3.4 – Découverte des éléments remarquables de type *Object Property* et *Data Property* sur la base des relations sémantiques simples. La colonne **%aug** correspond à l'augmentation potentielle du nombre de triplets d'ontologie.

permet d'acquérir les données structurées pour peupler les propriété spécifiées mais non peuplées et dont la structure s'apparente à celle d'une relation sémantique ainsi que de fournir des structures sémantiques pour lesquelles il est possible d'obtenir des éléments d'explication de l'inférence produite.

Dans le cadre des éléments de type « Object Property » dont la découverte s'appuie les règles plus complexes, notre méthode a permis de fournir les résultats tels que :

La colonne **filt** de la table 3.5 correspond au nombre d'éléments après filtrage (statistique, logique, validation, manuelle sommaire par un non expert), **%aug** correspond à l'augmentation que représentent les éléments remarquables acquis par rapport aux structures initialement présents dans MIAM. La différence des résultat entre les langues s'expliquent par le peuplement du RLSM_{PI} au moment de l'expérience ainsi que par les particularités culturelles. Par exemple, la culture culinaire russe comporte un très grand nombre de préparation sucrées (pâtisseries) ce qui génère un nombre important d'éléments candidats. L'intérêt de la méthode concerne l'acquisition massive de éléments permettant d'alimenter une

m	**prop**	**#trip**	**en**	**fr**	**es**	**ru**	***filt***	***%aug***
A	aPourProduitInitial	2 031	292	1 208	203	2 245	3 039	+149%
A	aPourEtatPhysique	543	30	29	10	53	85	+16%
A	aPourForme	39	77	78	5	37	132	+338%
A	aPourLabel	114	15	11	3	1	29	26%
A	aPourMethodeDeConservation	115	94	101	13	156	309	+269%
A	aPourMois	116	117	221	23	28	116	+288%
A	aPourRegion	289	98	71	2	57	216	+75%
A	aPourPays	27	192	285	91	113	483	+1 791%
A	aPourtypeDeLait	390	24	136	7	3	85	+22%
A	aPourProduitConstituant	98	256	302	143	103	570	+582%
A	aPourProduitInitialAromatisant	41	94	147	12	567	259	+633%
A	aPourRemplacant	64	312	356	128	267	733	+1 146%
P	aPourTypeDeCuisson	23	155	124	80	285	686	+2 986%
P	aPourDomaineCulinaire	82	112	92	120	1 313	1 276	+1 557%
P	aPourDecoupe	82	82	78	56	77	272	+332%
S	aPourSaveur	752	51	78	47	98	232	+31%
S	aPourDescripteurBruit	119	67	80	10	6	159	+134%
S	aPourCouleur	233	192	451	59	423	911	+391%
S	aPourAspectSurface	176	40	35	12	52	101	+58%
S	aPourSensationToucher	54	84	77	21	12	155	+287%
-	Total	5388	2384	3960	937	4953	9531	+177%

TABLE 3.5 – Résultats approche par règles (*ces résultats sont susceptibles d'évoluer en fonction de l'évolution de la ressource*). **m** correspond au nom du module (*Aliment* (A), *Préparation* (P), *Sensoriel* (S)), **prop** - au nom de la propriété, #**trip** - au nombre de triplets qui correspondent à une propriété données dans MIAM, **en**, **fr**, **es**, **ru** sont des contributions des différentes partitions en termes d'éléments de type *Object Property*.

folksonomie adossée à une ontologie de spécialité. Ces résultats peuvent être ventilés en termes de méthode utilisée (types de relations sémantiques, calcul de similarité etc.).

À l'heure actuelle de nos expériences et compte tenu du peuplement de notre ressource en termes de structures sémantiques qui nécessitent une contextualisation fine, nous ne sommes pas en mesure de produire une ressource médiatrice totalement structurée. Pour l'heure, cette ressource rassemble les différentes structures identifiées sans les structurer sous forme d'une ontologie-candidate. Le travail à venir consisterait à combiner les résultats respectifs des expériences sur les hiérarchies et sur les éléments de type *Object Property* afin de procéder à une éventuelle validation automatique via un outil spécialisé ou par la tâche (étiquetage d'un corpus).

Exemple 3.8

A titre d'exemple, l'information retournée par l'algorithme par règles a la forme suivante :

M : *жаркое* aPourProduitDiscriminant *подливка*

(ragoût aPourProduitDiscriminat la sauce)

M : *жаркое* aPourProduitDiscriminant *соус*

R : *ragoût* aPourProduitInitial *vegetable(produce)*

M correspond au cas de figure où l'inférence a été obtenue par correspondance, en exploitant les liens par pivot interlingue;

R indique que le mécanisme d'inférence a utilisé les raffinements des termes et a exploré les raffinements glosés;

Autres possibilités sont :

HM : hiérarchie des termes décrits par la règle ;
HR : hiérarchie et raffinements ;
M-no-rel : les termes n'ont pas de relation directe de type indiqué mais correspondent aux contraintes fournies par la règle (définition par l'ensemble de hyperonymes ou hyponymes, éléments de contexte typique sous forme de relations sémantique.

3.7 Discussion

Les structures extraites peuvent être rassemblées au sein d'une ressource médiatrice qui constitue la projection d'un modèle donné (ici, le modèle ontologique MIAM) sur le RLSM_{PI}. Une autre possibilité serait de viser la constitution d'une ressource médiatrice terminologique (glossaire monolingue ou multilingue).

Exemple 3.9

Dans l'environnement de gestion des terminologies multilingues *Libellex* (Brown De Colstoun et al. [2011]) développé par Lingua et Machina, entreprise-partenaire du présent projet de thèse, l'échange se fait via le format TBX [a] customisé et basé sur le concept des termes *vedettes* et *variantes*.

Dans le cadre de ce format, les entrées terminologiques (termEntry) peuvent regrouper plusieurs ensembles de termes (langSet) et les éléments descriptifs associés éventuellement en plusieurs langues qui peuvent contenir plusieurs termes enrichis (ntigs) qui contiennent à leur tour les unités terminologiques (term) et les informations associées à ces termes (termNote).

Une entrée regroupe ainsi le terme et ses variantes, permet de relier le terme à ses traductions ainsi que d'associer à un terme donné des relations sémantiques. Représentation du terme *mango juice* en utilisant le format TBX.

```
<termEntry>
<langSet id="en.langSet_1037" xml:lang="en">
<ntig id="en.ntig_1037">
<termGrp>
```

```
<term>mango juice</term>
<termNote type="termType">entryTerm</termNote>
<termNote type="partOfSpeech">Noun</termNote>
</termGrp>
</ntig>
<descripGrp>
<descrip target="en.langSet_332021" type="lexicoSemanticRelation">
partMeronym
</descrip>
</descripGrp>
<descripGrp>
<descrip type="crosslingualRelation" xml:lang="fr">
translationCandidate
</descrip>
<xref target="0612test1_tgtTbx.tbx" type="externalCrossReference">
fr.langSet_1039
</xref>
</descripGrp>
</langSet>
</termEntry>
```

Dans cet exemple, l'on retrouve à la fois les relations de traductions mais aussi la relation *partie-tout* pour le terme représenté. Le format TBX est principalement conçu pour la manipulation et le partage des glossaires (dont l'expressivité est limitée). Ainsi, pour la représentation des relations sémantiques il est nécessaire de redéfinir le format afin de pouvoir intégrer un typage de relations plus large.

a. `http://www.tbxinfo.net/tbx-about/`

Conclusion du chapitre

Dans le présent chapitre nous avons présenté les méthodes d'exploitation du RLSM_{PI} dans le but d'acquisition des termes et structures termino-ontologiques à partir d'une ressource lexico-sémantique structurée sous forme de graphe. Nous avons adapté la notion d'*élément remarquable* utilisée dans le cadre d'acquisition des traces de concepts à partir des textes au contexte d'acquisition de ces éléments à partir d'une ressource structurée. Nous avons montré comment le mécanisme d'inférence par abduction décrit par Zarrouk [2015] peut être utilisé pour la découverte des éléments remarquables. Dans le cadre de la sélection des termes pour l'abduction, nous avons détaillé deux stratégies :

— exploitation des relations taxonomiques (typées *r_isa* et *r_hypo*) relatives aux termes du RLSM_{PI} qui modélisent les concepts de l'ontologie à enrichir pour découvrir les termes et les structures qui correspondent aux concepts et individus d'ontologie ;

— exploitation des relations sémantiques éventuellement contextualisées via l'ajout d'une méta-information (annotation) pour découvrir les structures qui correspondent aux propriétés d'ontologie.

Contributions du chapitre 3

— définition du mécanisme d'immersion d'une ressource termino-ontologique dans une ressource lexico-sémantique (non ontologique) ;

— redéfinition de la notion de l'élément remarquable (adaptation au contexte de l'exploitation d'une ressource lexico-sémantique structurée sous forme de graphe) ;

— proposition d'utilisation du mécanisme d'inférence par abduction pour la découverte des éléments remarquables de type « concept », « individu », « Object Property ».

Chapitre 4

Évaluation de la ressource multilingue

L'amélioration continue d'une ressource notamment via les processus qui l'exploitent implique que cette ressource se trouve en perpétuelle évolution et nécessite du temps pour être stabilisée. L'évaluation de celle-ci est, elle aussi, effectuée de façon continue et s'apparente au diagnostic qui participe à son amélioration.

Comme tout type d'évaluation, l'évaluation d'une ressource sous forme de graphe telle que le réseau lexico-sémantique multilingue avec pivot interlingue ($RLSM_{PI}$) peut être quantitative (statistique notamment) ou qualitative (évaluation par la tâche). Lors de l'évaluation quantitative de ressources langagières classiques (dictionnaires, terminologies, ressources sous forme de graphe monolingues, ressources multilingues construites à partir des données ouvertes dont Wikipedia), les critères mis en avant sont le nombre d'entrées, d'informations sur chaque entrée, la variété de ces informations, le nombre de langues représentées. Dans le cadre d'évaluation d'un $RLSM_{PI}$, d'autres critères quantitatifs plus affinés peuvent être proposés. Certains critères d'évaluation quantitatifs du $RLSM_{PI}$ (nombre de termes raffinés) servent aussi de critères qualitatifs (granularité de la ressource, qualité d'alignement entre les partitions).

Termes et notations utilisés dans le chapitre 4

partition : (*RLS*) Sous-graphe d'un réseau lexico-sémantique multilingue.

relation : 1.*(RLS)* Relation sémantique faisant partie du $RLSM_{PI}$, par exemple, r_has_part. 2. *(Ontologie)* Relation hiérarchique ou thématique d'ordre ontologique.

Sur la plan quantitatif, nous introduisons les critères suivants :

1. **nombre de termes** notamment par type de terme (item lexical, termes qui modélisent les caractéristiques morpho-syntaxiques). La ventilation par type de terme permet de mettre en évidence l'efficience (qui peut être

entendue comme conception du modèle qui permet une *utilisation optimale de ses composants*) du modèle de représentation des connaissances ;

2. **distribution des termes en fonction de leur degré et de leur poids** qui dépendent du nombre de relations entrantes et sortantes d'un terme donné et des pondérations de celles-ci. Ce critère permet de se rendre compte de la qualité de peuplement d'une partition donnée ;
3. **nombre de termes raffinés comparé au nombre de termes identifiés comme polysémiques**. Il s'agit de la granularité de la ressource dont, en particulier, la représentation de sens. Quelle corrélation entre la distribution des termes lexicalisés [1] ayant plusieurs termes couvrants au niveau du pivot interlingue et les termes raffinés ?
4. **distribution des relations en fonction de leur type**. Un type de relation peu peuplé peut révéler les difficultés à extraire depuis le corpus, l'indisponibilité dans les ressources lexicales pouvant être intégrées, le type de relation mal défini ;
5. **distribution des relations en fonction de leur origine** (corpus, intégration, mécanismes d'inférences endogènes) ;
6. **couverture du pivot interlingue** soit la proportion des termes couverts par le pivot par rapport au nombre total des termes correspondant aux formes canoniques (par langue).

4.1 Évaluation quantitative

Nombre de termes. RLSM_{PI} contient **821 781 nœuds**. Parmi ces nœuds, se distinguent :

— les *items lexicaux* qui correspondent à un terme (mot du lexique) ou à un raffinement d'un terme ;

— les *items catégoriels* qui modélisent une catégorie grammaticale, des traits morpho-syntaxiques ;

— les *items interlingues* (termes interlingues qui correspondent à une ou plusieurs lexicalisations) ;

— les *items relationnels* qui correspondent à des réifications des relations sémantiques sous forme de nœuds (notamment dans le cas d'annotation).

Ces différents types de termes sont répartis comme suit :

	lexicaux	**interlingues**	**relationnels**	**catégoriels**	**total**
#nœuds	685 389	107 472	28 847	73	821 781
%nœuds	83%	13%	3,5%	0,008%	100%

TABLE 4.1 – Caractéristiques quantitatives du RLSM_{PI} : types de nœuds.

1. appartenant à une des partitions lexicalisées du RLSM_{PI}

Distribution des termes. Dans la plupart des ressources disposant d'un mécanisme de pondération dont notre ressource de référence, RezoJDM, le poids des termes a un caractère fréquentiel. Autrement dit, il s'apparente à la fréquence des termes dans un corpus. Ainsi, dans RezoJDM, le poids est lié à l'activité sur le réseau, au nombre de fois où un terme donné est utilisé par les joueurs pour répondre à la consigne.

Contrairement à notre ressource de référence RezoJDM, les poids présents dans le RLSM_{PI} ne sont pas obtenus grâce aux contributeurs. Malgré l'intégration des ressources de connaissance existantes guidée par le corpus, le choix n'a pas non plus été fait en faveur de l'utilisation de la fréquence des termes observée dans les corpus. Ainsi, le poids des termes du RLSM_{PI} est uniquement déterminé par les relations entrantes, mais aussi sortantes d'un terme. En effet, il s'agit d'un poids à caractère conceptuel et non fréquentiel car, compte tenu des objectifs applicatifs du RLSM_{PI} dont, en particulier, la construction ontologique pour un domaine de spécialité. Nous cherchons à connaître l'importance d'un terme donné pour le réseau lexico-sémantique de spécialité donné et non pour la langue en général.

Mis à part les relations issues de RezoJDM pour lesquelles nous avons fait le choix d'assimiler le poids d'origine à l'importance de la relation et d'intégrer ce poids tel quel, toutes les relations positives (sémantiquement vraies) sont créées avec un poids par défaut fixé à 25 ($w = 25$) et un ensemble d'indices qui correspondent à la confiance accordée à la ressource ou processus ayant fourni la relation ($\psi = \{i_1, i_2, ..., i_n\}$). Lorsque la relation est proposée par plusieurs ressources ou processus, l'ensemble d'indices de confiance augmente en intégrant les indices correspondants à ceux-là. Pour RLSM_{PI} en tant que ressource récente, à part les relations intégrées depuis RezoJDM, on observe une monotonie importante concernant les poids des relations. Pour faire évoluer ces poids, nous considérons l'*origine* des relations [2] à laquelle est associé un indice de confiance. Une relation possède ainsi un *score de poids global*.

Définition 4.1

Score de poids global d'une relation (SPG). Un ***score de poids global*** a été introduit pour prendre en compte à la fois le poids w (force d'association de la relation) lorsqu'il est disponible mais aussi la cardinalité $|\psi|$ et le maximum $Max(\psi)$ de l'ensemble des indices de confiance accordés à la relation en particulier dans le cas de pondération "par défaut" (force d'association inconnue). Le score de poids global ω est calculé comme suit :

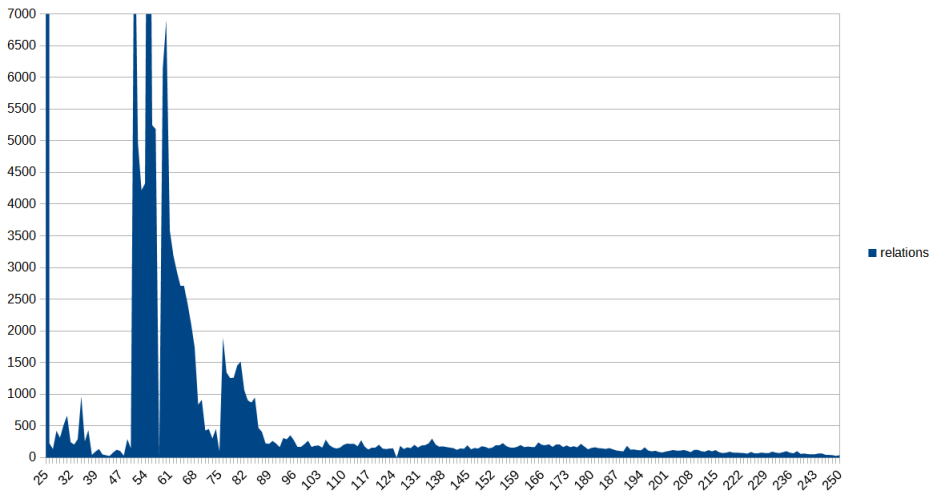
$$\omega = w \times \frac{|\psi|}{1 - log(Max(\psi))}$$

Pour le poids w fixé à 30 et un ensemble d'indices de confiance $\psi = \{0.2, 0.75\}$,

2. Ressources externes utilisées lors du processus d'intégration, mécanismes d'inférence.

> le SPG serait $\omega = 53$. Le même score pour le même poids par défaut et un maximum de confiance faible mais avec une cardinalité de l'ensemble des indices plus grande telle que $\psi = \{0.2, 0.15, 0.1\}$, on obtiendrait $\omega = 52$.
>
> Ainsi, si une relation apparaît un nombre de fois important dans des ressources avec peu de confiance, cela impacte le score de poids global globalement de la même façon que l'apparition de la relation dans peu de ressources ayant une confiance importante. Le maximum de confiance élevé associé à un nombre élevé de ressources augmente considérablement le score. Dans tous les cas, l'augmentation du poids est prévue par notre mode de calcul.
>
> Pour les poids négatifs issus notamment de RezoJDM ou inférés, étant donné que notre ressource ne comporte pas d'indices de confiance négatifs, le score de poids global vient accentuer le poids négatif déjà présent en augmentant la valeur absolue de ce poids.

La figure 4.1 donne la distribution des relations en fonction de leur poids « brut » tandis que la figure figure 4.2 montre la distributions des relations qui se fonde sur le SPG. Sur la figure 4.1, un pic autour du poids par défaut se distingue clairement. L'utilisation de SPG permet de nuancer la pondération des relations. Sur la figure 4.1, on remarque que de nombreuses relations ont un poids similaire



(a) distribution poids-#relations

FIGURE 4.1 – Distribution des relations du RLSM_{PI} en fonction de leur poids (échantillon concernant un poids entre 0 et 250).

voire identique ce qui rend difficile la discrimination d'une relation donnée en se servant du critère de poids. Les scores de poids des termes permettent de nuancer la pondération et de créer les contrastes nécessaires entre les termes issus de différentes ressource intégrées dans le RLSM_{PI}. Certaines relations qui étaient exclues des processus d'inférence du fait de leur poids peuvent désormais être utilisées. L'utilisation des poids dans le cadre des filtrages et des inférences a tout son sens lorsque l'on dispose d'une stratégie de pondération adaptée. Le calcul du poids des termes utilise le score SPG. Les termes qui ont plus de relations entrantes (définissent les autres termes) et éventuellement beaucoup de

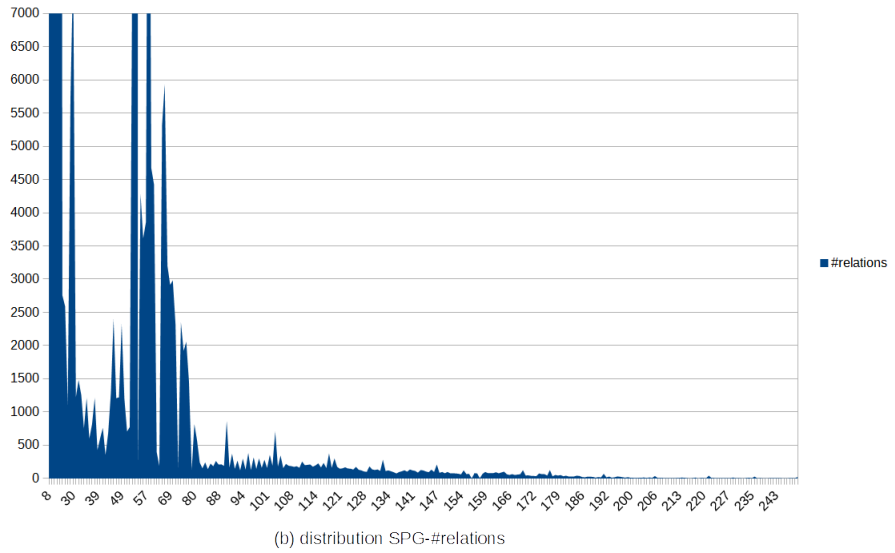


FIGURE 4.2 – Distribution des relations du RLSM$_{PI}$ en fonction de leur SPG (échantillon concernant un SPG entre 0 et 250).

relations sortantes (sémantique riche) ont un poids plus important.

Définition 4.2

Poids d'un terme.

Le poids d'un terme est la somme des poids de l'ensemble de ses relations entrantes et sortantes.

Si l'on souhaite connaître l'importance d'ordre conceptuel du terme, on considérera uniquement ses relations entrantes. Au contraire, si l'on s'intéresse à sa richesse sémantique, on calculera le poids du terme uniquement sur la base de ses relations sortantes.

Un terme t possédant un ensemble de relations pondérées [a] $R = \{r_1, r_2, ...r_n\}$ et un score de poids global ω associé à chaque relation a le poids *w(t)* :

$$w(t) = w_{in}(t) + w_{out}(t)$$

$$w_{in}(t) = log(|R_{in}(t)|) \times \sum_{i=1}^{R_{in}(t)} w_i$$

$$w_{out}(t) = log(|R_{out}(t)|) \times \sum_{i=1}^{R_{out}(t)} w_i$$

a. Poids normalisé par type de relation.

La distribution des termes sur un échantillon de 51 784 termes suit une loi de puissance comme représenté sur la figure 4.3.

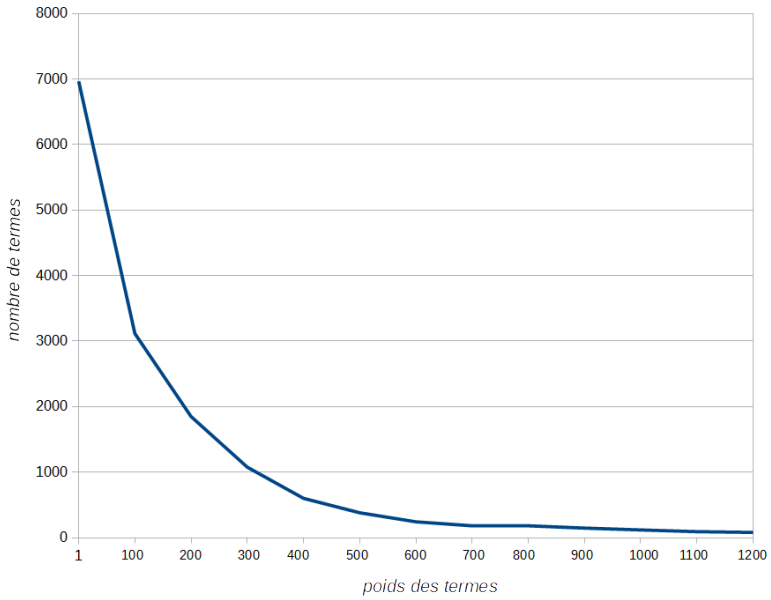


FIGURE 4.3 – Distribution des termes du $RLSM_{PI}$ en fonction de leur score de poids basé sur le calcul du score de poids global (SPG) pour ces relations entrantes et sortantes.

Sur la figure 4.3, l'on remarque beaucoup de termes du $RLSM_{PI}$ avec un SPG peu élevé donc ayant relativement peu de relations. Peu de termes possèdent beaucoup de relations. Le calcul de poids proposé s'applique principalement aux items lexicaux, interlingues et catégoriels car les mécanismes d'acquisition exogène et d'inférence s'appliquent principalement à ces types de termes fournissant les indices de confiance. En effet, les règles des mécanismes d'inférence produisent une action qui concerne les termes lexicalisés et interlingues (ajout des relations, création de nouveaux termes [3])

Exemple 4.1

Parmi les exemples de termes ayant le poids le plus élevé : *consommer* (212 540), *aliment* (37 391), *fruit* (32 560), *boisson* (37 222) *végétal* (36 304), *préparation culinaire* (26 496), *légume* (25 888), *substance* (22 776), *qualité* (21 549), etc.

En pratique, l'indice de confiance accordé à des mécanismes de peuplement exogènes, en particulier, extraction depuis les corpus de textes, sera souvent plus élevé que dans le cas des mécanismes d'inférence. Cependant, le mécanisme d'inférence ascendante des relations sémantiques prend en compte les poids des relations lors du filtrage statistique. Ainsi le mode d'acquisition des poids des termes unique peut être utilisé pour les termes interlingues et lexicalisés.

3. La création des termes par inférence peut uniquement concerner les raffinements des termes interlingues et lexicalisés ainsi que la création des termes interlingues.

Nombre de termes raffinés par rapport au nombre de termes identifiés comme polysémiques. Dans le cadre du RLSM$_{PI}$, pour qu'un terme puisse être considéré comme polysémique, il doit posséder plusieurs termes couvrants dans le pivot interlingue. L'écart entre la polysémie identifiée et le nombre de raffinements potentiels se présente comme représenté sur les figures 4.4 et 4.5 au sein du RLSM$_{PI}$.

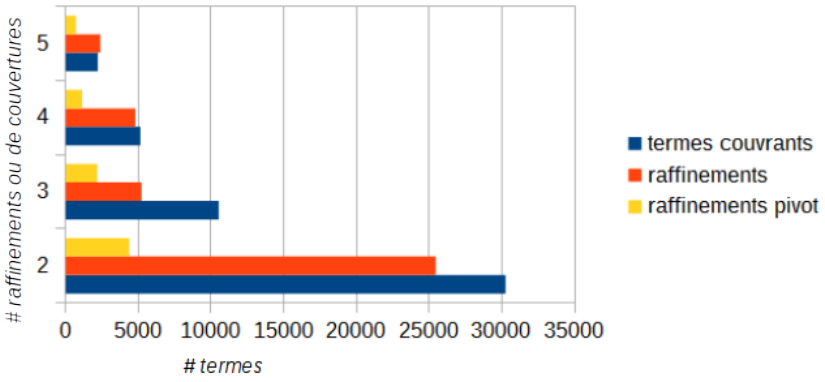


FIGURE 4.4 – Distribution de l'ensemble de termes identifiés comme polysémiques (termes qui ont plusieurs "couvertures") et de l'ensemble de termes raffinés (termes qui possèdent des raffinements).

En ce qui concerne le nombre de raffinements par terme, nous observons la répartition telle qu'elle est représentée sur la figure 4.5.

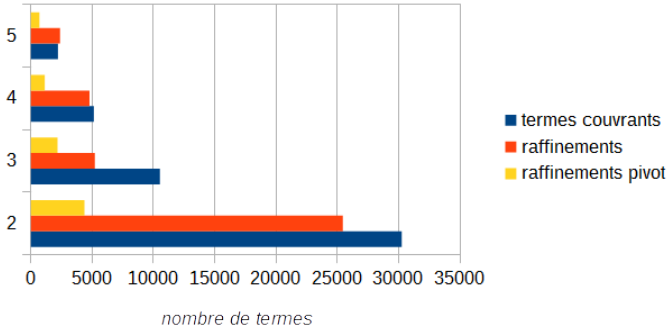


FIGURE 4.5 – Distribution des termes identifiés comme polysémiques, des termes raffinés lexicalises et des termes raffinés interlingues par nombre de distinction de sens (nombre de raffinements ou nombre de couvertures) observé (axe Y). Il y aurait presque autant de termes raffinés que polysémiques. Cependant, il ne s'agit pas forcément des mêmes termes.

Sur ces figures, nous remarquons une corrélation relativement importante entre le nombre de raffinements et le nombre de "couvertures" interlingues pour les termes qui possèdent 4 sens différents ou plus. Cependant, pour les termes ayant

2 ou 3 raffinements potentiels ou avérés [4], le degré de raffinement (nombre de relations sortantes typés *r_refinement*) est inférieur au degré de couverture (nombre de relations entrantes typées *r_covers*). Cette observation révèle la présence potentielle des couvertures superflues soit le taux de raffinement encore bas du pivot interlingue (également reflété sur la figure 4.5).

Nombre de relations. Le nombre de relations est de **2 231 197** (à l'heure où nous écrivons). L'analyse des relations permet de mesurer à la fois la ***pertinence du modèle*** et l'***impact des différentes méthodes de construction de la ressource***.

Dans le présent paragraphe, nous ne nous attarderons pas sur la répartition globale des relations actuellement présentes dans le RLSM_{PI} car elle a été fournie dans la section 2.5. En revanche, il semble intéressant de comparer cette répartition avec les ressources langagières qui ont été partiellement intégrées dans le RLSM_{PI}.

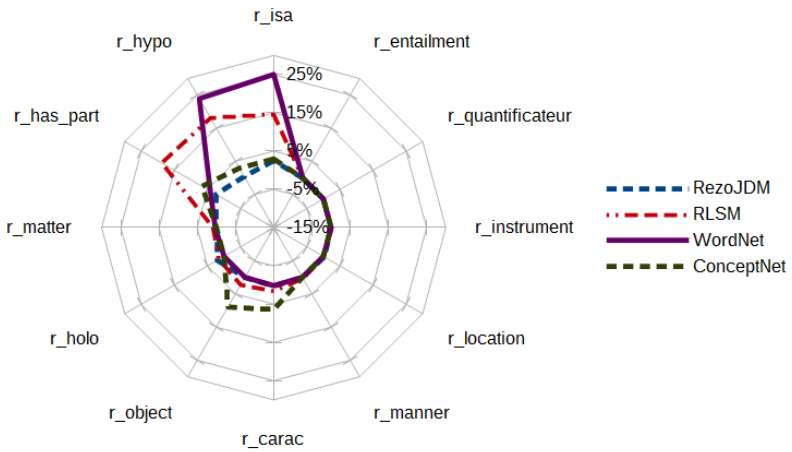


FIGURE 4.6 – Répartition des principales relations du RLSM_{PI} en comparaison avec les autres ressources langagières qui en contiennent.

La ***pertinence du modèle*** est liée à l'utilisation en pratique des différents types de relation prévus. Sont-ils peuplés ? Leur peuplement est-il équilibré ? Pour affiner l'échelle de comparaison, nous distinguons plusieurs sous-ensembles de relations et faisons les comparaisons à l'intérieur de ces sous-ensembles de relations sémantiques :

1. relations hiérarchiques : *r_isa*, *r_hypo*, *r_has_part*, *r_matter*, *r_holo*. Sur la figure 4.6, on remarque un déséquilibre en faveur de ces types de relations au sein du RLSM_{PI}. Ceci est dû, d'une part au nombre réduit de types de relations et d'autre part au caractère récent du RLSM_{PI} ;

4. Les raffinements peuvent être inférés non-glosés ou glosés, les couvertures interlingues peuvent être non encore fusionnées compte tenu du fait que le pivot interlingue du RLSM_{PI} est amorcé à partir d'un pivot naturel puis transformé en pivot interlingue de façon incrémentale.

2. relations thématiques : *r_carac*, *r_manner* ;
3. relations prédicat-argument : *r_object*,*r_instrument*, *r_location*.

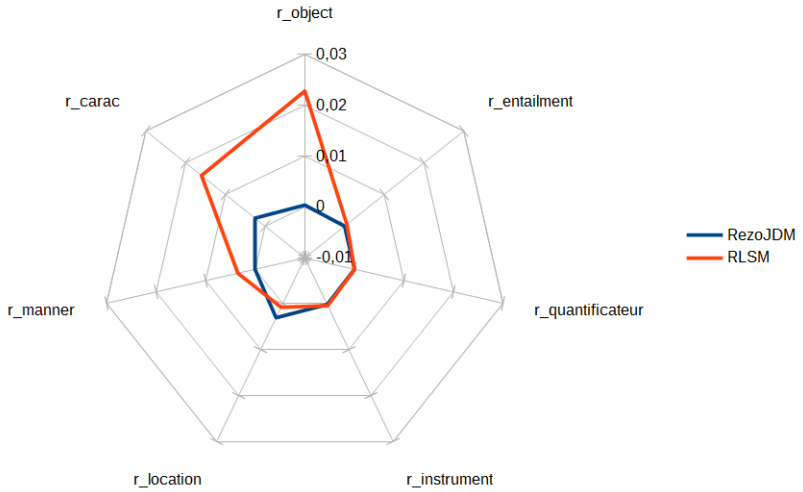


FIGURE 4.7 – Répartition des principales relations thématiques du $RLSM_{PI}$ en comparaison avec RezoJDM.

Sur la figure 4.7, l'on remarque un déséquilibre en faveur des relations typées *r_object* et *r_carac* ce qui est dû au caractère spécialisé du $RLSM_{PI}$, à sa taille encore modeste et au nombre de types de relations sensiblement moins important que dans RezoJDM, ressource de connaissance générale.

Pour estimer l'impact des différents processus de construction et du mode d'amorçage par intégration des ressources pré-existantes guidée par le corpus de spécialité (décrits dans le chapitre 2 du présent mémoire) sur la construction du $RLSM_{PI}$ nous pouvons nous baser sur les observations suivantes :

— intersection entre les relations extraites depuis le corpus et les relations acquises par intégration, intersections entre les différentes ressources intégrées ;

— corrélation entre le nombre de relations acquises par des méthodes exogènes et la productivité d'inférence (processus endogène) des relations.

L'intersection entre les ensembles de relations acquises par extraction depuis le corpus est de 11 164 relations tous types confondus. Les types de relations principalement acquises depuis le corpus sont *r_manner*, *r_has_part*, *r_object*, *r_location*, *r_carac*.

Les ressources qui présentent le plus d'intersection en termes de relations sémantiques sont des ressources anglophones WordNet, ConceptNet et DBNary (partition *en*). Quelques intersections ont pu être observées entre les ressources anglophones et francophones au détour des expériences d'inférence des relations sémantiques.

Le tableau 4.3 montre que l'impact des mécanismes d'inférence cross-lingue diffère selon la couverture des ressources intégrées. Dans le cas des types de relations très présentes dans le corpus utilisé pendant la phase d'amorçage ou dans

Ressources	DBNary(en)	WordNet	ConceptNet	IATE	RezoJDM
DBNary (en)	-	-	-	6 925	23
WordNet	-	-	33 563	4	19
ConceptNet	-	33 563	-	4	-
IATE	6 925	4	4	-	77
RezoJDM	-	-	-	77	-

TABLE 4.2 – Intersections entre les ensembles de relations sémantiques intégrés dans le $RLSM_{PI}$ depuis les ressources extérieures.

R_{type}	corpus	intég.	avant inf	inf	prod
r_isa	67 894	544 632	612 526	27 546	4,5%
r_hypo	688	797 783	798 471	41 053	5,14%
r_has_part	662 737	172 287	835 024	48 015	5,75%
r_matter	606	35 597	36 203	1 893	5,23%
r_holo	224	67 081	67 305	51 360	76,31%
r_object	42 280	29 262	71 542	15 512	21,68%
r_carac	8 300	69 236	77 536	9 521	12,28%
r_manner	2 854	3 250	6 104	250	4,10%
r_location	2086	3 573	5 659	146	2,58%
r_instrument	58	2 738	2 796	402	14,31%
r_refinement	221	29 441	29 662	182 135	614,03%
Totaux	787 948	1 754 880	2 542 828	377 816	15%

TABLE 4.3 – Caractéristiques quantitatives du $RLSM_{PI}$

le cadre des ressources intégrées seront à priori de faible productivité. Une autre explication peut être la couverture insuffisante du pivot interlingue.

4.2 Évaluation qualitative : problématiques et exemples

Pour notre ressource et pour d'autres ressources similaires, les tâches d'évaluation qualitative peuvent se subdiviser en plusieurs ensembles selon différents critères.

En prenant le critère de la ***langue***, elles peuvent être monolingues et cross-lingues.

— tâches monolingues (centrées sur une des langues de la ressource même si d'autres partitions de la ressource peuvent être sollicitées à titre complémentaire). Il s'agit de tester la richesse de la ressource pour une langue donnée (analyse sémantique et ses extensions) ;

— tâches cross-lingues (qui nécessitent nécessairement le recours à plusieurs partitions de la ressource). Il s'agit de tester la qualité d'alignement, l'accent est mis sur la richesse du pivot interlingue.

En prenant la ***nature de la tâche*** comme critère, il est possible de distinguer :

— les tâches d'évaluation ouvertes : analyse sémantique généraliste, évaluation incrémentale des terminologies (glossaires, ressources terminologiques multilingues) notamment validation humaine via une interface homme-machine dédiée ;

— les tâches d'évaluation bornées : analyse sémantique bornée par un critère prédéfini (détection des sentiments, détection des incompatibilités alimentaires, etc.). Globalement, il s'agit des tâches de classification.

Les tâches les plus informatives sur l'état des sous-parties du $\mathrm{RLSM_{PI}}$ à un moment t sont des tâches d'évaluation bornées. Le critère d'arrêt des processus est clairement identifié, le traitement statistique exogène (calcul des scores, analyse exogène des résultats) est possible. L'inconvénient de ces tâches est leur caractère limité à un ensemble de critères d'arrêt ce qui implique l'exploration seulement partielle du $\mathrm{RLSM_{PI}}$.

Dans le cas des tâches d'évaluation ouverte, il s'agit de l'évaluation incrémentale qui participe à l'apprentissage permanent et permet d'enrichir le $\mathrm{RLSM_{PI}}$. Avec ce type d'évaluation, il est possible de détecter les lacunes (termes manquants, relations manquantes) et d'en comparer le nombre avec le nombre de relations trouvées ou inférées au cours de l'analyse.

Dans cette section, nous allons introduire l'analyse sémantique informée en prenant l'exemple d'analyse et explicitation des instructions de cuisine. Puis nous nous focaliserons sur la tâche d'analyse sémantique bornée, la détection des incompatibilités plat-régime et sur la tâche de pré-validation cross-lingue des contributions RezoJDM en attente (relations sémantiques proposés par les joueurs mais non validés par des processus automatiques).

L'analyse sémantique basée sur un réseau lexico-sémantique peut également s'inscrire dans une démarche de conceptualisation à partir des textes dans la mesure où elle permet d'expliciter le contenu textuel.

1. **analyse sémantique**. Cette analyse sous-tend l'ensemble des méthodes d'extraction d'informations à partir d'un ensemble de documents. Cette tâche permet d'évaluer la ressource en termes de sa richesse sémantique (présence des termes et des relations, modélisation de la polysémie lexicale) ainsi que la qualité de son alignement. L'analyse sémantique est aussi une tâche productive car elle permet de constamment enrichir la ressource grâce à des termes ou relations manquantes identifiés ;
2. **détection des incompatibilités plat-régime** est une tâche qui permet d'évaluer l'adéquation de la ressource par rapport à une problématique concrète qui consiste à situer les plats par rapport à un régime alimentaire spécifique. Il s'agit d'une spécialisation de l'analyse sémantique. Cette tâche permet d'évaluer la qualité de représentation des recettes et des aliments, processus, ingrédients ;
3. **pré-validation cross-lingue des contributions en attente** (relations sémantiques). La validation des contributions en attente soit des contri-

butions qui n'ont pas pu être validés par des mécanismes par règles et nécessite un traitement manuel peut difficilement être automatisée. Nous proposons une pré-validation des contributions cross-lingue afin de faciliter la validation des contributions en attente par réduction du volume de données à traiter.

4.2.1 Analyse sémantique des instructions de cuisine

Les recettes de cuisine sont fréquemment prises comme exemple pour l'analyse des instructions et la modélisation des processus.

Les bases de ce type d'analyse sémantique ont été posées par Lafourcade [2011] sous appellation "analyse holistique des textes". Dans sa version originale, ce modèle implique les particularités suivantes :

— analyse par émergence (suppression de contrôle supervisé) ;

— suppression des phases syntaxique et sémantique d'analyse. Remplacement de ces phases par des micro-tâches itérées dont le déroulement dépend des objets présents dans la structure de calcul à un moment donné. L'approche est ainsi indépendante de la disponibilité des informations nécessaires pour telle ou telle phase d'analyse ;

— approche *anytime* (possibilité de suspendre le calcul à tout moment afin de lire la structure telle qu'elle est à un moment précis) ;

— processus d'analyse qui cherche à créer des structures sémantiques à partir des éléments présents dans l'environnement à savoir dans le segment textuel en cours de l'analyse et dans la ressource de connaissance (réseau lexico-sémantique) qui sous-tend l'analyse.

Une approche similaire à l'analyse des textes a été proposée et développée par Poria et al. [2014].

Données utilisées. Nous avons utilisé 50 instructions de cuisine par langue du RLSM_{PI} pour l'expérience d'analyse sémantique des instructions.

Déroulement de l'algorithme. De façon approximative, nous pouvons supposer que notre algorithme suit un déroulement séquentiel pour décrire ses différentes aspects.

1. ***segmentation du texte fourni en entrée*** (détection des termes multimots, suppression des mots-outils). Par exemple, pour l'instruction *roll out the dough*, nous obtenons à la sortie de segmentation une liste ordonnée *roll out, dough* ;
2. ***création d'une chaîne de termes (graphe de surface)*** dans lequel les termes apparaissent sous forme de séquence dans l'ordre du texte fourni en entrée. Par exemple, pour l'instruction *in a food processor, finely chop*

the garlic, parsley, and drained beans, le graphe de surface est composé de termes recopiés à partir du RLSM_{PI} vers le graphe d'analyse de l'instruction et se présente comme suit :

$$in \xrightarrow{r\_succ} food\ processor \xrightarrow{r\_succ} finely \xrightarrow{r\_succ} chop \xrightarrow{r\_succ} garlic \xrightarrow{r\_succ} parsley \xrightarrow{r\_succ} drained \xrightarrow{r\_succ} beans$$

La création de cette chaîne de termes permet d'identifier les termes manquants et de déclencher leur acquisition par des mécanismes exogènes.

3. ***recherche et recopie des relations grammaticales*** (relations forme-lemme et des parties de discours), utilisation des marqueurs éventuellement présents dans le segment textuel fourni à l'entrée pour structurer l'arbre syntaxique de l'instruction. Les parties de discours sont représentées sous forme de termes interlingues. Ces termes spécifiques peuvent comporter des raffinements, par exemple *in :Prep>loc* ou *in :Adv>duration*. Ainsi, le graphe de surface introduit dans (2) prend forme telle qu'elle est présentée sur la figure 4.8.

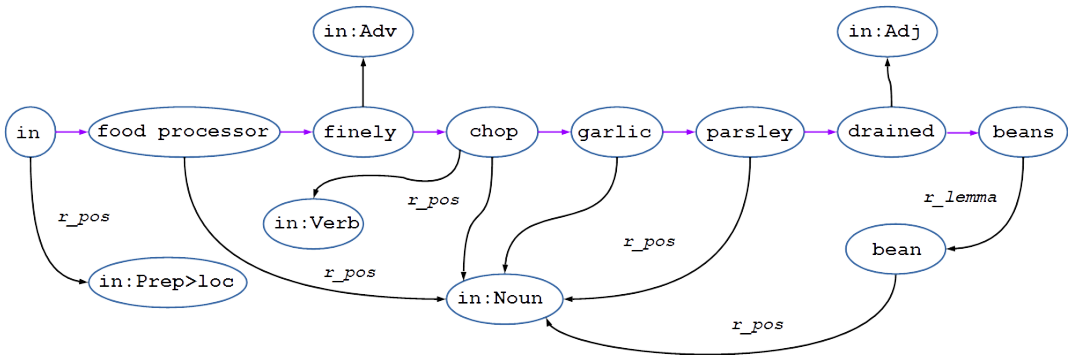


FIGURE 4.8 – Graphe de surface augmenté par l'ensemble des relations grammaticales. ce graphe laisse apparaître la polysémie morpho-syntaxique du terme *chop* ainsi que la détection d'un "marqueur" de lieu *in*.

4. ***recherche et recopie des relations sémantiques*** afin de désambiguïser et enrichir la structure de graphe de surface précédemment augmenté par des relations grammaticales. Celui ci prend la forme représentée sur la figure 4.10.

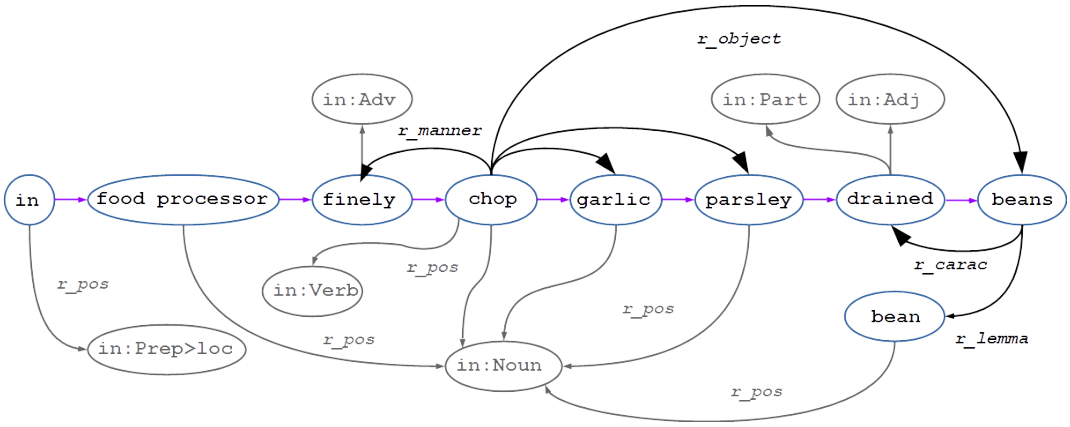


FIGURE 4.9 – Graphe de surface augmenté par recherche et recopie de l'ensemble des relations sémantiques existant entre ses termes au sein de RLSM_{PI}.

Grâce à la relation *chop* $\xrightarrow{r\_object}$ *garlic* ou encore *chop* $\xrightarrow{r\_object}$ *parsley*, le terme *chop* est désambiguïsé en tant que verbe, la relation typée *r_pos* vers le terme *in :Noun* est négativée (Annexe A, exemple A.12). il s'agit à cette étape d'augmenter le graphe par recherche des relations qui existeraient entre les termes du graphe de surface au sein de RLSM_{PI} et leur recopie dans le graphe d'analyse.

5. ***augmentation par la découverte des sous-graphes de voisinage, inférence des relations manquantes.*** L'exploration de la ressource de connaissance à la recherche des éléments à intégrer dans le graphe d'analyse inclut également le parcours du voisinage à la recherche des chemins de longueur $l = 2$ qui relieraient les termes du graphe d'analyse ainsi que à la recherche des termes qui permettraient de déclencher les mécanismes d'inférence par triangulation [5] ou encore inférence par abduction interlingue (les candidats termes font partie du pivot interlingue).

5. Dans le cadre d'un réseau lexico-sémantique, l'inférence par triangulation est un processus qui permet de proposer un type de relation donné entre deux termes à partir d'une série de substitutions d'une relation par un terme intermédiaire. Trianguler revient à "éliminer" successivement les candidats-termes et à créer les relations de type recherché lors de ces éliminations.

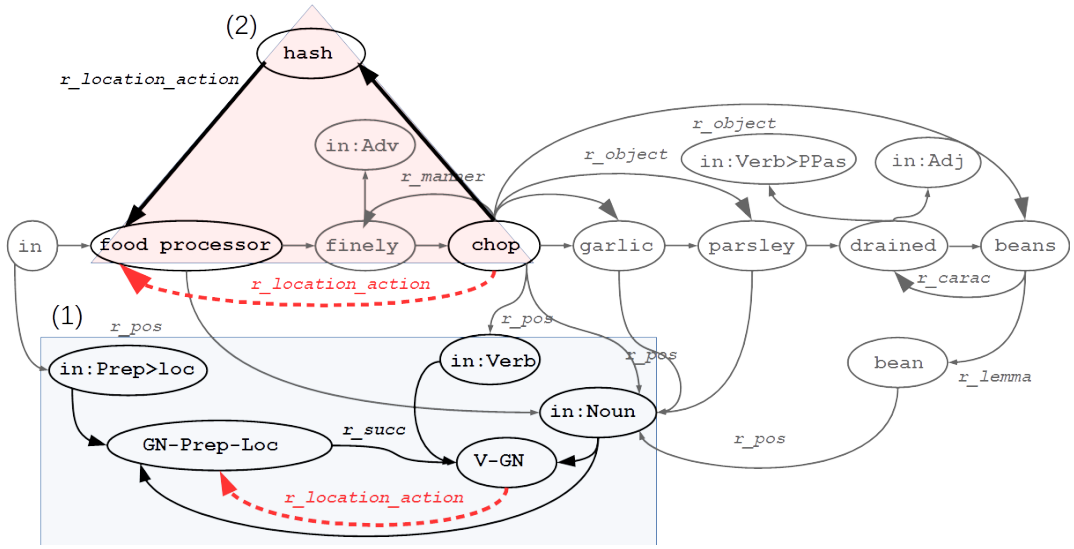


FIGURE 4.10 – Graphe de surface augmenté par recherche et recopie de l'ensemble des termes et relations sémantiques existant au sein de RLSM$_{PI}$. (1) illustre le mécanisme d'inférence par règles, (2) illustre l'inférence par triangulation.

6. ***inférence hors segment, explicitation.*** Il s'agit de sortir des limites de l'ensemble des termes présents au sein du segment textuel à analyser pour expliciter les instructions (ex. processus antérieur, lieu, instrument sous-entendu etc.).

L'intérêt de la tâche d'analyse sémantique globale pour l'évaluation est lié aux aspects suivants :

1. il s'agit d'une analyse multi-critères qui sollicite des types de connaissances variés (relations grammaticales, lexicales, sémantiques) et qui mobilise les différents mécanismes d'inférence présents au sein de la ressource de connaissance sous-jacente ;

2. il s'agit d'une « micro-tâche » qui permet de voir les lacunes et les tendances d'une ressource de connaissance à partir de peu de données de test ;

3. il s'agit d'une tâche qui peut être spécialisée ;

4. c'est une tâche d'évaluation productive qui permet de déclencher les mécanismes d'augmentation de la ressource de connaissance.

Algorithme 5 : Analyse sémantique des instructions de cuisine (ASIC)

input : RLSM_{PI}, texte_instruction
output : $Triplets[]$

```
// initialisation
Triplets[]← ∅ ;
GrapheAnalyse[]← ∅ ;
Terms[]← ∅ ;
// identifier les termes ainsi que les termes multi-mot, en cas de
   découpages multiples
Terms[]←Segmenter (texte_instruction)
// construction du réseau de travail
GrapheAnalyse← ConstruireGrapheDeSurface (Terms[]) ;
// taux de croissance du graphe d'analyse
TauxAugmentation=Taille(GraheAnalyse)/100 ;
while TauxAugmentation >0 do
    t_debut=Taille (GrapheAnalyse) ;
    FaireAnalyse (GrapheAnalyse,RLSM_PI) ;
    t_fin=Taille (GrapheAnalyse) ;
    TauxAugmentation=t_fin/t_debut − TauxAugmentation ;
Triplets[]←GrapheAnalyse
retourner Triplets[]
```

L'algorithme démarre avec la constitution d'un graphe d'analyse (graphe de surface) qui reflète l'ordre des mots pleins dans le texte à analyser. Il et constitué des termes dont les étiquettes correspondent aux formes trouvées dans le texte à analyser. Puis le graphe de surface est enrichi grâce à la recherche et la recopie des termes et des relations depuis le RLSM_{PI}. Cet enrichissement se poursuit tant que l'analyse permet de découvrir de nouvelles relations. L'algorithme d'évaluation par la tâche d'analyse sémantique a été appliqué à un jeu de données de test contenant 200 instructions de cuisine sélectionnées à la main, 50 pour chaque langue du RLSM_{PI}. L'évaluation de la tâche se base principalement sur le critère de **précision** (quel est le pourcentage des relations trouvées et inférées qui sont pertinentes [6]). Le **rappel** (proportion des relations pertinentes ont pu être récupérées et inférées pour un segment textuel donné) est relativement difficile à estimer. Pendant l'analyse le réseau est considéré comme « complet »(hypothèse du monde fermé). Le rappel met en jeu les valeurs qui correspondent à des vrais positifs (relations trouvées et inférées correctes) et des faux négatifs (les silences, relations pertinentes absentes) où la présence des faux négatifs vient diminuer le nombre de vrais positifs.

$$Rappel = \frac{Relations\ pertinentes}{Relations\ pertinentes + Silences}$$

Concernant ces silences, nous ne savons pas pour quelle raison les relations pertinentes n'ont pas pu être récupérées. Elles peuvent être soit absentes soit

6. En se basant sur l'évaluation manuelle

présentes ou inférables mais non identifiées lors de l'analyse. Dans le cadre de notre méthode, en plus de la recherche des termes et relations *présentes* dans le RLSM_{PI}, nous procédons à l'identification des relations *inférables*. Lors du calcul de précision et de rappel, les relations inférables sont traitées comme s'il s'agissait des relations existantes. Les observations détaillées dans le tableau 4.4 nous

Lang	**Surface**	**Recopie**	**$\text{Inf}_{\text{segment}}$**	**$\text{Inf}_{\text{hors segment}}$**	**Filtrage**	**Précision**
en	7	7,3	3,5	172	14	76%
fr	5.5	5.25	3	84	33	72%
ru	5	5,75	6	252	159	80%
es	4,6	3	2,2	81	16	56%

TABLE 4.4 – Moyennes concernant les différentes étapes d'analyse sémantique.

.

renseignent sur la qualité du peuplement du RLSM_{PI}. **Recopie** rend compte des silences en ce qui concerne les termes simples et multi-mot (TMM). Les résultats du test ont été satisfaisants sur ce point car, globalement, la consistance du segment textuel est maintenue lors du passage du texte au graphe de surface. $\textbf{Inf}_{\textbf{segment}}$ montre l'efficacité des mécanismes basés sur les patrons et sur la connaissance. Il s'agit de vérifier dans quelle mesure le mécanisme d'analyse s'approche d'une clique ou quasi-clique qui correspond au segment à analyser. $\textbf{Inf}_{\textbf{hors segment}}$ permet d'analyser la richesse sémantique des termes qui composent le segment à analyser enfin, **Filtrage** et **Précision** permettent d'évaluer la stratégie de filtrage choisie. En effet, le filtrage doit augmenter la précision.

4.2.2 Détection des incompatibilités plat-régime

La détection des incompatibilités plat-régime est une spécialisation de la tâche d'analyse sémantique. L'exploration et la recopie des relations sémantiques ainsi que l'inférence des relations sont bornées par la découverte de la relation typée *r_ incompatible* qui relie les aliments incompatibles avec tel ou tel régime alimentaire spécifique aux termes qui portent l'étiquette de ce régime alimentaire. Par exemple, pour *porc>viande*, nous avons $\textit{porc>viande} \xrightarrow{r\_incompatible} \textit{halal}$.

Dans le cadre de cette tâche, nous nous sommes fixée l'objectif d'opérer un rapprochement entre la connaissance linguistique et la connaissance du monde (notamment, la connaissance sur la composition nutritionnelle des aliments et sa modification lors des transformations). Nous avons tenté de détecter les incompatibilités régime à partir des noms des plats tels qu'ils peuvent être trouvés dans un menu de restaurant.

Données utilisées. Afin de donner un cadre à notre expérience, nous nous sommes intéressée à la propriété *suitableForDiet* qui s'applique au type *Recipe* dans schema.org [7] et, par conséquent, au type *RestrictableDiet* dont les

7. schema.org/Recipe

instances sont : *DiabeticDiet* (diabète), *GlutenFreeDiet*(sans gluten), *HalalDiet* (halal), *HinduDiet* (hindou), *KosherDiet* (kasher), *LowCalorieDiet* (hypocalorique), *LowFatDiet* (sans graisse), *LowLactoseDiet* (sans lactose), *LowSaltDiet* (sans sel), *VeganDiet* (végan), *VegetarianDiet* (végétarien). Pour ces régimes, nous avons extrait automatiquement puis validés manuellement les ensembles totalisant environ 200 ingrédients interdits obtenus à partir des ressources spécialisées. Ces interdits alimentaires divergent sur le plan sémantique (composition nutritionnelle, type de produit, type de découpe etc.).

Nous avons exploité un ensemble de 5 000 titres de recettes en français soit un corpus de 19 000 mots dont 2 900 termes (hors mots-outils) [8]. Nous avons utilisé uniquement la partition *fr* du RLSM_{PI} pour cette expérience. Les termes les plus fréquents du corpus [9] : *salade* (157), *tarte, poulet* (124), *soupe* (112), *facile* (93), *chocolat* (89), *saumon, gâteau* (87), *légume, confiture* (86).

Notre corpus a été annoté semi-automatiquement afin de permettre une évaluation. Pour une partie des titres, le contenu des recettes et les méta-informations correspondantes ont permis de récupérer l'information liée à la compatibilité des recettes avec un sous-ensemble des régimes considérés dans le cadre de notre expérience ce qui a facilité l'annotation.

Déroulement de l'algorithme. La recherche des incompatibilités plat-régime prend en entrée le nom du plat et le RLSM_{PI}. A la sortie, l'algorithme renvoie un vecteur d'incompatibilités pour l'ensemble des régimes considérés. La représentation sous forme d'un vecteur permet des traitement ultérieurs tels que entraînement d'un classifieur.

Les étapes d'analyse de titres des recettes correspondent à celles décrites dans la sous-section 4.2.1 mais comportent aussi des spécificités liées à la détection des incompatibilités. Dans le cadre de cette expérience, l'analyse s'appuie sur le *contexte* (graphe de surface correspondant au nom de plat fourni à l'entrée et le voisinage des termes du graphe de surface) et une liste de termes qui définissent le *domaine* concerné (ex. *art culinaire*, *alimentation* etc.).

Pour chaque terme du contexte $t_i \in C$, l'algorithme effectue un parcours de graphe typé en fonction de la catégorie locale (préparation ou ingrédient simple) du terme t_i. Si t_i est une préparation, le parcours utilise les types {hypo, part-whole, matter}. C'est le cas des mélanges, mets, autres ingrédients complexes. Si t_i est un ingrédient simple (par exemple, *tomate*), les types de relations explorées sont {isa, syn, part-whole, matter, carac}. Le poids de toutes les relations utilisées lors du parcours type du RLSM_{PI} doit être strictement positif.

Dans son ensemble, l'analyse se déroule comme représenté sur la figure 4.11. Lors du parcours typé, les relations candidates sont filtrées selon deux straté-

8. Répartition du corpus : *15% www.cuisineaz.fr, 20% www.cuisinelibre.org et 65% www.allrecipe.fr*

9. *terme*(fréquence).

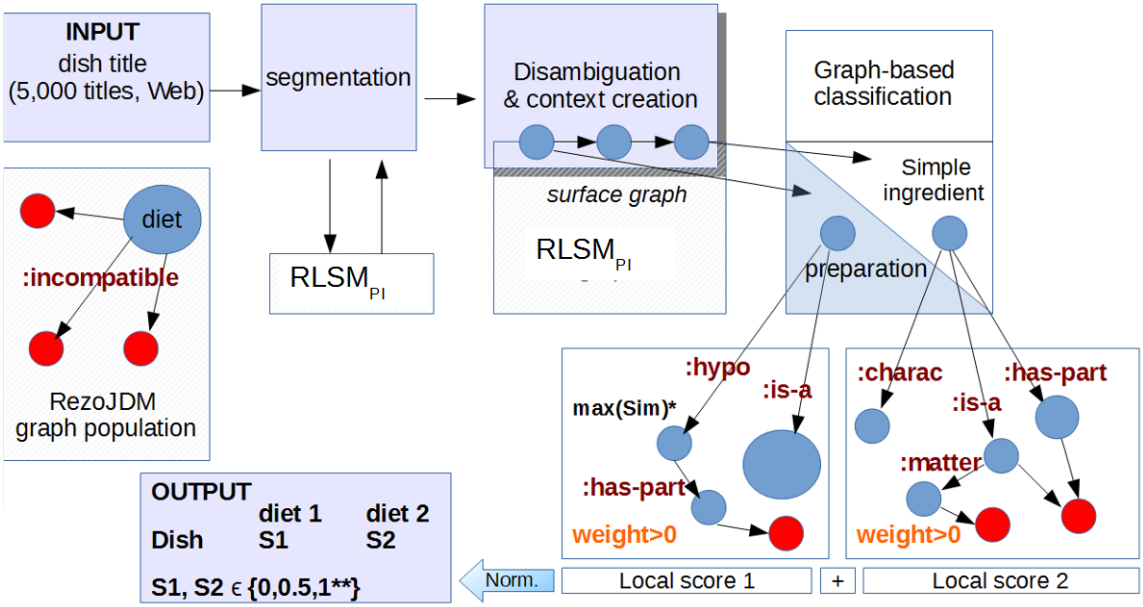


FIGURE 4.11 – Déroulement de l'analyse sémantique des noms de plats afin de déterminer leur compatibilité avec un régime alimentaire donné.

gies principales : pertinence par rapport au domaine d'intérêt et validation par triangulation [10]

L'exploration des ingrédients typiques des plats dont les titres sont analysés se base sur les hyponymes du terme qui correspond à un plat donné (par exemple pour "quiche thon tomate", "quiche" correspond à une préparation, ses hyponymes sont "quiche saumon épinard", "quiche lorraine" etc.). Au sein du RLSM_{PI} on obtient via les relations typées r_has_part et r_matter tous les ingrédients et composants possibles d'une préparation donné. Pour ne sélectionner que les résultats pertinents pour la préparation en train d'être analysée P, nous ne récupérons que les méronymes que le terme-préparation plus général (ex."quiche") partage avec ses hyponymes similaires à P (ex. "quiche thon moutarde"). Pour le calcul de similarité, un indice de Jaccard normalisé sur toutes les relations sortantes typées r_has_part, r_isa, r_matter est utilisé.

Définition 4.3

Indice dit *de Jaccard* (calcul de similarité entre deux termes). Pour un contexte C, l'indice de Jaccard normalisé est calculé selon la formule :

$$J(S_C, S_{C_{hypo}}) = \frac{S_C \cap S_{C_{hypo}}}{S_C \cup S_{C_{hypo}}}$$

10. Lorsqu'on se trouve en présence d'une relation sémantique candidate qui lie les termes x et y sachant que x fait partie du *contexte* (segment en train d'être analysé), on tente de retrouver un chemin typé de longueur 2 qui relie x et y en passant par un des termes présents dans le voisinage de x. Dans le cadre d'analyse sémantique tel qu'il est décrit dans le présent chapitre, si un tel chemin existe, la relation candidate est recopiée dans le graphe d'analyse.

où

C_{hypo} est le contexte de l'hyponyme à tester. Ce contexte est construit à la volée ;
S_C est l'ensemble des relations sortantes de tous les termes du contexte à analyser ;
$S_{C_{hypo}}$ est l'ensemble de relations sortantes de l'hyponyme à tester.

Dans sa version simple, ce score se base sur la *présence des relations* et le calcul du score qui ne détaille pas l'importance de chaque relation (ne prend pas en compte son poids ou son score de poids global). Déjà avec ce mode de calcul, il est possible d'obtenir des résultats satisfaisant et de discriminer les hyponymes non pertinents.

Dans sa version pondérée, à l'instar de l'approche de Lafourcade and Joubert [2009], il s'agit d'utiliser la somme des poids normalisés par type de relation. On calcule le poids de l'ensemble des relations que les deux contextes ont en commun, puis on le ramène à la somme des poids de toutes les relations que les termes des deux contextes à comparer possèdent.

$$J(S_C, S_{C_{hypo}}) = \frac{w(S_C \cap S_{C_{hypo}})}{w(S_C \cup S_{C_{hypo}})}$$

Une fois le contexte enrichi, le calcul de compatibilité peut être lancé. Ce calcul se base de nouveau sur le parcours de graphe qui tente de relier les parties du contexte enrichi à l'ensemble des termes qui modélisent les régimes alimentaires. Les chemins typés *r_holo, r_has_part, r_matter, r_hypo* sont explorés. Le sens de l'exploration est arbitraire, à savoir il est possible de partir d'un régime et remonter vers le terme du contexte enrichi à tester ou l'inverse. Le parcours en partant du terme correspondant au régime alimentaire donné réduit la complexité de l'algorithme de recherche de compatibilité car il s'agit d'un terme spécialisé, souvent monosémique dont la richesse sémantique (nombre de relations sortantes) est réduite.

Définition 4.4

Le chemin typé. Lors du parcours de graphe, le chemin typé est un chemin élémentaire qui contient les relations dont le type est défini pour un parcours de graphe donné.

Pour un ensemble de types $T = \{t_1, t_2, t_3, t_\gamma\}$, le chemin typé de longueur l S_t appartient à l'ensemble des chemins qui ont une forme suivante :

$$\{S_1 = ((n_1, r_{t_1}, m_1),\ (m_1, r_{t_1}, o_1 ..., (o_l, r_{t_\gamma}, p_l))\}$$

Le test de multiples chemins typés génère une combinatoire qui dépend de la longueur du chemin et du degré typé (ne considérant que les types d'intérêt)

des termes qui constituent le chemin. Ainsi, pour n types de relations, la longueur du chemin typé l, les degrés des termes notés $d_1,\ d_2,\ d_n$, la combinatoire serait de

$$C = \frac{\left(\sum_{i=1}^{l} d_i\right)!}{(d_1! \times\ d_2!, \times ...,\ d_n!)}$$

si l'on considère qu'il peut y avoir des répétitions parmi les types de relations utilisées lors du parcours (c'est-à-dire un seul et même type de relation peut apparaître plusieurs fois).

Ainsi la combinatoire *type-longueur-degré* pour $l = 2,\ d_1 = 10,\ d_2 = 3,\ C = \frac{13!}{10! \times 3!} = 2860$

La combinatoire concernant les types de relations et la longueur (*type-longueur*) serait, quant à elle, peu significative. En prenant comme formule de base :

$$C_n^l = n! + \frac{n!}{(l!(n-m)!))}$$

on trouverait que, pour un ensemble de 5 types de relations à tester avec répétitions (chemin dont toutes les relations auraient le même type) et une longueur de 2, nous aurions $C_5^2 = 5! + \frac{5!}{(2!(5-2)!))} = 130$ Compte tenu de ces observations, le filtrage statistique prend toute son importance dans le cadre des tâches d'analyse sémantique.

Un score de compatibilité est construit pour chaque terme du contexte enrichi C (tel qu'il a été précédemment décrit). Ce score est déterminé par la distance d à laquelle se trouve le terme correspondant au régime alimentaire (terme opposé du chemin). Au fur et à mesure que l'on s'éloigne du terme de départ, le score local d'incompatibilité s_{loc} d'un aliment potentiellement incompatible avec un régime alimentaire s'incrémente selon la règle :

$$s_{loc} = s_{loc} + \frac{1}{1+d}$$

On obtient le score global pour un contexte enrichi en additionnant les scores obtenus localement pour chacun de ses composants.

Exemple 4.2

Calcul du score d'incompatibilité pour le nom de plat *quiche thon tomate*.

entrée :
`quiche thon tomate`

graphe de surface
`quiche>preparation→thon>poisson→tomate>legume`

contexte :
`r_has_part pâte d=1 w=6`
`r_has_part oeufs d=1 w=105`
etc.

scores intermédiaires :
`Diabetes 1.3, LactoseFree 3.0, Halal 0.0,`
`Kosher 0.3, LowCalories 1.5, LowSalt 0.5,`
`GlutenFree 1.8, Hindu 0.5, LowFat 0.8,`
`Vegan 0.5, Vegetarian 0.3`

Compte tenu du nombre fixe de régimes alimentaires explorés (N régimes, $N = 11$) et du fait que les scores pour les différentes paires aliment-régime sont indépendants, nous pouvons considérer l'ensemble des scores locaux comme un vecteur. A ce titre, il peut être normé en utilisant la *norme euclidienne* (norme-2) :

$$X_{norme-2} = \sqrt[2]{\sum_{i=1}^{N=11} score_i}$$

On aboutit à un vecteur de compatibilité normé, par exemple :

Exemple 4.3

Vecteur d'incompatibilité normé pour le nom de plat *quiche thon tomate*. La norme est de 4,21.

`Diabetes 0.30, LactoseFree 0.71, Halal 0.00,`
`Kosher 0.07, LowCalories 0.36, LowSalt 0.12,`
`GlutenFree 0.42, Hindu 0.12, LowFat 0.19,`
`Vegan 0.12, Vegetarian 0.07`

La sortie de l'algorithme de calcul d'incompatibilité plat-régime a été "simplifiée" afin de n'obtenir que 3 classes : compatible (0,0), incertain (0,5) et incompatible (1,0). ce choix est motivé par le fait que notre ressource est encore très récente et ne permet pas d'obtenir des scores très fiables. Pour aboutir à ces classes, nous avons appliqué la règle de calcul du score final $score_{fin}$ suivante :

$$\text{si } s \geq 0.5,\ score_{fin} = 1, \text{ sinon } score_{fin} = 0,5$$

Il est également possible de garder les différentes nuances des scores plat-régime pour une analyse plus nuancée, notamment, analyse experte.

Un parcours du $RLSM_{PI}$ spécifique impliquant la recherche des caractéristiques (exploitation de la relation typée *r_carac*) est effectué pour le régime sans sel.

Les résultats de cette expérience ont reflété l'état du réseau notamment en ce qui concerne le peuplement des relations typées r_has_part, r_matter, r_isa, r_holo ainsi que la représentation des contraintes liées aux différents régimes alimentaires dans le $RLSM_{PI}$. Ces points se reflètent dans le nombre de scores d'incompatibilité qui ont pu être produits ainsi que dans la "confiance" de ces scores. Cette confiance peut être mesurée à travers l'analyse de écart type des scores obtenus.

	Corpus (scores)		
Régime	**0**	**0.5**	**1**
Diabetes	344	1 667	2 989
LactoseFree	1 410	1 856	1 733
Halal	1 629	3 140	231
Kosher	1 307	3 453	240
LowCalories	2 540	122	2 338
LowSalt	3 491	1 312	197
GlutenFree	589	2 961	1 450
Hindu	568	3 939	493
LowFat	2 161	2 636	203
Vegan	454	4 261	285
Vegetarian	360	1 655	2 985
totaux	**14 852**	**27 004**	**13 144**

TABLE 4.5 – Répartition entre 3 classes d'incompatibilité dans l'ensemble du corpus.

Pour un régime donné, les valeurs de moyenne (M) et de l'écart type (SD) font apparaître les cas de figure suivants :

- faible moyenne et faible écart type révèlent une grande quantité de scores "incertain". Le peuplement de la ressource est sans doute insuffisant ;
- moyenne forte et écart-type faible témoignent d'une confiance importante concernant le score. Les paire plat-régime concernée peuvent servir à générer de règles d'inférence pour augmenter la couverture de la ressource ;
- une moyenne assez élevée et écart-type élevé (le cas des régimes *GlutenFree* et *LowCalories*) indiquent que les données sont inégalement réparties dans la ressource, qu'il existe des zones insuffisamment peuplées. Par conséquent, des mécanismes d'inférence peuvent être nécessaires pour équilibrer les structures sémantiques relatives à un régime alimentaire donné ;
- moyenne et écart type très faibles révèle une stratégie de parcours inadaptée ou un peuplement de la ressource insuffisant. La contextualisation (notamment, sous forme d'annotation des relations) est possiblement nécessaire dans ce cas.

Sur l'ensemble de la partition ***fr*** du $RLSM_{PI}$, la moyenne et l'écart type ont révélé une confiance faible (beaucoup de scores « incertain ») et le résultat uniforme (peu d'écart de distribution).

Diet	*M*	*SD*
Diabetes	0.785	0.295
LactoseFree	0.200	0.215
Halal	0.240	0.239
Kosher	0.275	0.243
LowCalories	0.487	0.484
LowSalt	0.031	0.094
GlutenFree	0.549	0.349
Hindu	0.447	0.244
LowFat	0.380	0.202
Vegan	0.406	0.219
Vegetarian	0.809	0.253
overall	**0.406**	**0.219**

TABLE 4.6 – Moyenne et écart type observés lors du calcul des scores d'incompatibilité concernant les différents régimes alimentaires.

Résultat de l'expérience . Les résultats de l'expérience ont été exprimées en termes de précision, rappel et F-score. L'évaluation la plus importante concerne la précision car, pour une personne intolérante, même une petite quantité d'aliment interdit peut être dangereuse.

	Données d'évaluation		
Régime	**Précision**	**Rappel**	**F1 score**
Diabetes	92%	92%	92%
LactoseFree	71%	73%	72%
Halal	65%	75%	70%
Kosher	67%	60%	63%
LowCalories	60%	75%	67%
LowSalt	88%	65%	75%
GlutenFree	80%	73%	76%
Hindu	86%	80%	83%
LowFat	67%	70%	68%
Vegan	80%	90%	85%
Vegetarian	83%	90%	86%
macro-average [11]	**76%**	**77%**	**76%**

TABLE 4.7 – Évaluation basée sur le corpus. les corpus d'évaluation et un sous-corpus du corpus principal utilisé pour l'expérience annoté en incompatibilités. Nous n'avons pas tenu compte des relations annotées présentes dans le RLSM_{PI} au moment de l'expérience.

4.2.3 Pré-validation translingue des contributions en attente (relations sémantiques)

La validation des contributions en attente peut devenir un point « bloquant » pour les ressources qui utilisent les processus contributifs d'acquisition des don-

nés structurées. Dans les cadre de notre ressource de référence, RezoJDM, la validation des contributions est faite par un ensemble de mécanismes basées sur des règles. Les contributions restées en attente pour cette ressource sont validées manuellement. Parmi celles-ci, environ 70% sont des contributions justes, 20% sont peu pertinentes et 10% sont des relations fausses.

Donnée utilisées. Nous avons exploité un ensemble de 2 491 triplets typés issus des contributions en attente RezoJDM (nous n'avons pas considéré la relation typés r_associated, relation sous-spécifiée).

Déroulement de l'expérience. Pour tester notre ressource, nous avons fait l'expérience de pré-validation automatique des contributions en attente pour RezoJDM en utilisant un mécanisme translingue organisé en deux étapes.

1. Vérification de l'existence de la relation à tester dans les autres partitions du RLSM_{PI}. Si la relation existe dans une autre langue que le français ou dans plus de une langue, elle est considérée comme pré-validée ;
2. Vérification des relations inférables en exploitant le voisinage des termes qui correspondent aux termes de la relation fournie en entrée du processus dans d'autres langues.

Le processus proposé a fourni les résultats décrits dans le tableau 4.8.

type	**candidats**	**pré-validés**	**%pré-validés**	**précision**
r_isa	136	59	43%	88%
r_has_part	361	170	47%	93%
r_mater>object	118	59	50%	80%
r_carac	1 024	466	45%	79%
r_carac-1	131	70	53%	83%
r_patient-1	395	177	38%	99%
r_agent-1	57	14	25%	86%
r_lieu	583	190	33%	98%
r_instr	136	30	22%	19%
Total (moyen.préc.)	2 491	1 235	50%	(80%)

TABLE 4.8 – Pré-validation des contributions en attente. La précision dans la ligne *Total* est la moyenne arithmétique de la précision observée pour les types de relations traitées.

Résultats de l'expérience. Les résultats de l'expérience de pré-validation translingue des contributions en attente permettent de se rendre compte de l'état de maturité du RLSM_{PI} ainsi que de son caractère spécifique à un domaine particulier. Les observations résumées dans le tableau 4.8 se traduisent en termes de :

— ***richesse (cardinalité) de l'ensemble des types de relations sémantiques pouvant être présents dans la ressource***. Plus la ressource est généraliste et mature, plus elle a de types de relations. Dans le cas d'une ressource de spécialité, certains types de relations n'ont que peu d'utilité. Les pré-validations des relations suivantes n'ont pas pu être faites : *r_conseq, r_magn, r_anti_magn*. La relation r_agent-1 est très peu peuplée car la connaissance qui concerne les instructions de cuisine est centrée sur les actions et les patients (objets) des actions ;

— ***couverture du pivot interlingue (qualité de l'alignement)*** Dans une ressource immature, il existe un nombre important de silences au niveau du pivot. Le processus de pré-validation des contributions en attente étant dépendant de la possibilité de passer d'une langue à l'autre et de disposer des structures adéquates en termes de granularité (présence des raffinements de sens alignés), des vrais négatifs fournis par le système sont principalement dus à la polysémie des termes perdue lors du passage d'une langue à l'autre ;

— ***qualité de peuplement*** de la ressource. On remarque la relation typée *r_instr* qui a fourni très peu de relations pré-validées ce qui est dû aux silences quant à la relations de ce type.

Malgré la performance encore insuffisante en terme du nombre de relations pouvant être validées par type (due à des silences des termes et relations), il s'agit du diagnostic du RLSM_{PI} utile afin de guider son peuplement et son amélioration par des mécanismes endogènes et exogènes.

Par ailleurs, la validation basée sur un RLS permet de fournir une explication concernant la décision de validation positive ou négative proposée.

Exemple 4.4

Exemples des donnés de sortie du processus de pré-validation :

1. validation grâce à l'*existence de la relation sémantique correspondante* dans d'autres partitions de la ressource :

chocolat r_meaning marron>couleur true because *chocolate -r_carac-> brown*

2. validation grâce à l'analyse du voisinage et aux processus d'inférence par triangulation (présence d'un chemin typé entre les termes du triplet fourni en entrée das d'autres partitions que celle de la langue d'origine), par déduction/induction (en utilisant des relations transitives telles que r_isa)

pâte_à_tartiner r_has_part chocolat true because chocolate –r_patient-1–> spread

chocolat r_lieu verre true because *in :chocolate r_isa –> in :drink* & *in :drink r_quantifier in :glass*

L'algorithme de validation des contributions en attente peut être résumé comme suit :

Algorithme 6 : Pré-validation des contributions en attente des relations sémantiques.

input : RLSM_{PI}, triplets_à_valider
output : Validés[]

```
1 // initialisation
2 Validés[]← ∅ ;
3 validation← ∅ ;
4 explication← ∅ ;
5 for relation_FR ∈ triplets_à_valider do
6     for terme_source, terme_cible do
7         if ∃relation_¬FR then
8             validation=true ;
```

Discussion. Un mode d'évaluation possible pour une ressource lexico-sémantique serait de la comparer à une autre ressource lexico-sémantique. Cependant, pour pouvoir procéder à une évaluation de ce type, il serait nécessaire de disposer d'une ressource de référence qui devrait être une ressource de spécialité multilingue interopérable avec RLSM_{PI}. De façon générale la comparaison directe des ressources est difficile car elle nécessite la mise en place d'une procédure de leur mise en correspondance. En revanche, lors de l'intégration des ressources, une telle comparaison devient possible. Il devient également possible d'estimer l'apport de la ressource en cours d'augmentation et de consolidation en termes de récolte de nouvelles connaissance (non répertoriées dans les ressources existantes).

Par conséquent, l'évaluation des ressources de connaissances se fait couramment par la tâche. Les campagnes d'évaluation sont alors organisées afin de permettre l'évaluation des ressources multilingues. Les tâches peuvent être de nature très variée : substitution lexicale, désambiguïsation, similarité translingue, etc.

De tels types d'évaluation sont au-delà du cadre applicatif de la ressource que nous avons construite étant donné que notre ressource n'est pas une ressource de connaissance générale même si elle en intègre une grande partie. En effet, l'amorçage de la ressource a été guidé par le corpus de spécialité ce qui la rend ciblée pour la construction ontologique et l'analyse des textes qui relève du domaine de la cuisine de nutrition.

Dans le contexte industriel de gestion des terminologies multilingues, il est possible d'évaluer la qualité d'alignement des unités terminologiques présentes au sein du RLSM_{PI} en faisant l'extraction des termes et en les reversant dans un environnement industriel de gestion terminologique et d'aide à la traduction. La validation des alignements entre les termes serait alors faire de façon incrémentale, au fur et à mesure que le glossaire terminologique est utilisé.

Par ailleurs, les critères structurels issus de la théorie de graphes sont peu informatifs quant à la qualité d'un réseau lexico-sémantique multilingue car d'une part une ressource telle que RLSM_{PI} comporte des relations typées et valuées grâce à l'attribution de score de poids ou d'annotation. Par conséquent, pour un protocole d'évaluation pertinent, il devient indispensable de disposer de méthodes pour évaluer la représentation de la polysémie et des relations sémantiques et l'impact des différentes méthodes de construction.

Conclusion du chapitre

Nous avons introduit dans ce chapitre les critères d'évaluation d'une ressource multilingue de type réseau lexico sémantique. Ce type de ressource sémantiquement riche est difficile à évaluer en se basant sur une ressource pré-existante. Des métriques spécifiques (notamment, SPG) ont été proposées pour intégrer les méta-informations fréquentielles et les informations sur l'origine de l'information présente au sein du réseau. Une attention particulière a été accordée à l'évaluation des différentes méthodes de construction du réseau afin de se donner les moyens de juger sur l'apport de la ressource par rapport à l'état de l'art. Les expériences d'évaluation qualitative par la tâche ont donné

Contributions du chapitre 4

Les contributions du chapitre sont les suivantes :

1. identification des paramètres quantitatifs d'évaluation de la ressource lexico-sémantique multilingue avec pivot interlingue ;
2. proposition du score de poids global (SPG) pour permettre de faire évoluer le poids des relations du RLSM_{PI} dans le temps et en prenant en compte l'impact des différents mécanismes de peuplement du RLSM_{PI} et en absence d'acquisition contributive ;
3. proposition des grands ensembles de tâches d'évaluation qualitative pour une ressource multilingue de spécialité et, en particulier, des tâches d'analyse ouverte et limitée (classification) et des tâches translingues. Il s'agit des tâches qui contribuent à la construction permanente de la ressource car leur résultats peuvent être exploités enrichir cette dernière.

Chapitre 5

Vers un système semi-automatique de construction termino-ontologique

Le présent chapitre vise à présenter la méthode de construction ontologique assistée basée sur un réseau lexico-sémantique multilingue avec pivot interlingue. Les expériences présentées permettent de mettre en perspective l'exploitation du $RLSM_{PI}$ dans le cadre de construction des ressources termino-ontologiques menée par les experts humains.

Termes et notations utilisés dans le chapitre 5

système multi-agent (SMA) : système basé sur plusieurs programmes autonomes (appelés « agents ») qui interagissent avec leur environnement.

environnement : *(SME)* Structure de données perçue et pouvant être modifiée par un agent.

rôle : 1. *(SME)* Fonction qu'à l'agent dans le système. Le type d'interaction avec l'environnement dépend de cette fonction. Type d'interaction avec l'environnement. 2. *(Ontologie)* Ensemble de couples d'individus d'ontologie. *Dans le présent chapitre la définition (1) sera utilisée.*

Les exemples cités dans le présent chapitre qui viennent du $\mathrm{RLSM_{PI}}$ et détaillent les relations sémantiques sont notés sous forme relationnelle, par exemple :

$$\textit{cocky-leeky} \xrightarrow{r\_location} \textit{Scotland}.$$

où *r_location* est le type de la relation.

Une relation annotée faisant partie du $\mathrm{RLSM_{PI}}$ est notée de la façon suivante :

$$\textit{cocky-leeky} \xrightarrow{r\_location::region} \textit{Scotland}.$$

où *région* est une annotation de la relation. Une annotation est une méta-information ajoutée au type de la relation. Il peut y avoir plusieurs annotations pour une seule et même relation.

Les propriétés d'ordre ontologique (*Object Properties*) présentes dans l'ontologie de référence ou proposées à l'expert pour enrichir l'ontologie sont notées comme dans l'exemple suivant :

borshch `aPourPays` *Russie*.

où `aPourPays` est le nom de la propriété d'ordre ontologique. Cette *Object Property* est représentée dans le RLSM_{PI} comme suit :

$$borshch \xrightarrow{r\_location} pays$$
$$borshch \xrightarrow{r\_location} Russie$$
$$Russie \xrightarrow{r\_isa} pays$$

Dans un premier temps, nous allons présenter la ressource termino-ontologique qui sous-tend les expériences décrits dans le présent chapitre. *Dans un second temps*, nous introduirons les expériences et les outils qui visent à accompagner la construction d'ontologie. *Dans un troisième temps*, nous allons résumer l'architecture du système d'exploitation du réseau lexico-sémantique multilingue avec pivot interlingue (RLSM_{PI}) et nous focaliser sur le système d'aide à la construction termino-ontologique en tant que piste d'ouverture de nos travaux.

La termino-ontologie MIAM a été immergée [1] au sein du RLSM_{PI} pour permettre la recherche d'une passerelle entre le modèle de structuration des connaissances propre aux réseaux lexico-sémantiques et les termino-ontologies. La structure modulaire de MIAM a été introduite dans le chapitre 2. Puis, dans le chapitre 3, nous avons détaillé nos expériences quant à l'immersion de MIAM dans un réseau lexico-sémantique et l'enrichissement de ses *Object Properties*.

L'ensemble des expériences présentées dans le présent chapitre s'appuie d'une part sur la représentation de la termino-ontologie MIAM en termes du réseau lexico-sémantique multilingue avec pivot interlingue (RLSM_{PI}) et, d'autre part, sur le module SensoMIAM [2] de cette termino-ontologie que nous chercherons à enrichir en soumettant automatiquement des propositions aux experts à partir du contenu de RezoJDM, réseau lexico-sémantique du français (notre ressource de référence tout au long du présent mémoire, plus mature et stable après 11 années de construction que le RLSM_{PI} multilingue encore récent).

5.1 Présentation de la termino-ontologie *SensoMIAM*

SensoMIAM est un module de MIAM [3]. Le périmètre de ce module concerne la modélisation des descripteurs sensoriels. La représentation des caractéristiques

1. L'immersion est entendue ici comme expression conformément à un modèle donné. Voir le chapitre 3 pour le détail sur l'immersion de l'ontologie MIAM dans le RLSM_{PI}.
2. `http://www-limics.smbh.univ-paris13.fr/sensoMIAM/`
3. MIAM est une ontologie modulaire où chaque module modélise un domaine particulier : aliment, processus, nutrition, personne, matériel, etc.

sensorielles suscite un intérêt particulier chez les professionnels de l'alimentation. Nous avons souhaité proposer une méthode qui utilise un réseau lexico-sémantique monolingue pour enrichir une ontologie spécifique à cette représentation et supposée « en cours de développement ».

SensoMIAM comporte un ensemble de classes qui modélisent les ***descripteurs sensoriels***. Toutes ces classes sont des « classes énumérées » (*enumerated class*) [4] soit des classes définies par énumération des individus (instances de ces classes). Chaque descripteur est une sous-classe de la classe `DescripteurSensoriel`. Les classes énumérées servent de point de départ à notre expérimentation sont données ci-dessous.

Parmi les descripteurs toucher (`DescripteurToucher`) :

`DescripteurThermique` = {*brûlant, chaud, frais, froid, glacé, tiède*}

`DescripteurTact` = {*astringent, fibreux, filandreux, granuleux, grumeleux, lisse, nerveux*}

`DescripteurConsistance` = {*cassant, collant, coulant, dense, farineux, ferme, feuilleté, fluide, fondant, gélatineux, gras, juteux, moelleux, mou, onctueux, pâteux, sec, tendre*}

`DescripteurComposition` = {*crème, herbe, légume, œuf*}

`DescripteurBruit` = {*craquant,croquant, croustillant*}

Parmi les descripteurs d'aspect (`DescripteurAspect`) :

`DescripteurSurface` = {*brillant, gras, gratiné, grillé, lisse, nappant*}

`DescripteurSubstance` = {*aéré, bouillant, dense, épais, fin*}

`DescripteurMacroForme` = {*carré, ovale, plate, rectangulaire, ronde, sphérique, triangulaire*}
`DescripteurCouleur` = {*blanc, brun, clair, coloré, doré, jaune, noir, orange, pourpre, rose, rougeâtre, sombre,vert*}

`DescripteurContraste` = {*hétérogène, uniforme*}

4. `https://www.w3.org/TR/owl-ref/#EnumeratedClass`

Parmi les descripteurs de flaveur [5] (`DescripteurFlaveur`) :

`DescripteurSaveur` = {*acide, amère, salée, sucrée, umami*}

`DescripteurSensationTrigeminale` = {*fraîcheur, fraîcheur mentholée, pétillant, piquant*}

`DescripteurArome` = {*agrume, aillé, alcool, amande, anisé, beurré, brûlé, caramélisé, champignon, crémeux, crucifère, épicé, fleuri, fort, fromage, fruit à coque, fruité, fumé, gras, herbe aromatique, iodé, légume, noisette, noix, œuf, oignon, poireau, olive, piquant, poisson, poivron, terreux, tomate, végétal, viande*}

Le module d'ontologie comprend aussi les ensemble de classes qui modélisent les ***aliments*** : (Légume, PommeDeTerre etc.) et les ***caractéristiques sensorielles des aliments*** (umami, uniforme etc). Elle comporte également quelques *Object Properties* : `aPourConsistance`,`aPourCouleurChair`, `aPourTypeChair`, `aPourModeCuisson`.

Les questions que nous pouvons nous poser quant à l'enrichissement de l'ontologie *SensoMIAM* (ainsi que des ressources similaires) sont les suivantes :

1. quelle est la spécificité d'un descripteur sensoriel par rapport à une simple caractéristique telle qu'elle peut être trouvée dans une ressource langagière comportant le type de relation « caractéristique » et comment capturer l'aspect qualifiant d'un descripteur ?
2. quelles caractéristiques des aliments peuvent être intégrées dans l'ontologie en cours de construction sous forme de *Object Properties* (quels types de relations du RLS (RezoJDM) doivent être explorés) ?

5.2 Enrichissement et construction ontologique

5.2.1 Enrichissement

L'expérience d'enrichissement s'est déroulée comme suit :

1. analyse des structures présentes au sein de l'ontologie à enrichir, identification des points à améliorer ;
2. passage des concepts de l'ontologie aux termes (l'ébauche d'ontologie ne comporte pas nécessairement d'étiquettes de concepts, une phase de mise en correspondance manuelle peut être nécessaire) ;

5. Ensemble des sensations olfactives, gustatives et tactiles ressenties lors de la dégustation d'un produit alimentaire.(Larousse, `https://www.larousse.fr/dictionnaires/francais/flaveur/34069`)

3. identification des types de relation et des moyens de contextualisation (annotation des relations, définition des ensemble de départ et d'arrivée de la relation réelle ou inférable) ;
4. production des triplets candidats ;
5. évaluation et discussion des résultats obtenus.

Au départ du processus d'enrichissement et de conceptualisation nous disposons de 115 termes issus de l'expression des URI des concepts SensoMIAM [6] sous une forme lexicalisée. Dans le cadre de nos expérimentations, nous faisons l'hypothèse que le module SensoMIAM est encore un module qui nécessite d'être construit et structuré.

Compte tenu des informations présentes dans le module d'ontologie à enrichir, nous pouvons procéder de façon suivante :

— vérifier la convergence entre les ensembles des termes de RezoJDM par type de descripteur (*arôme*, *aspect*, *toucher*) qui correspondent aux listes des individus des classes énumérées (calcul de l'intersection) et proposer des individus-descripteurs supplémentaires ;

— identifier les aliments possédant les caractéristiques listées dans les descripteurs sensoriels, annoter les relations correspondantes [7] et récupérer leurs relations sémantiques contextualisées par rapport au domaine de spécialité de l'ontologie à enrichir.

Exemple 5.1

Par exemple, si lors du parcours de RezoJDM, nous retrouvons le terme *pomme* qui possède les relations réelles ou inférables suivantes :

$pomme \xrightarrow{r\_carac::bruit} croquante$

$pomme \xrightarrow{r\_carac::saveur} sucrée$

$pomme \xrightarrow{r\_carac::saveur} fruitée$

$pomme \xrightarrow{r\_isa} fruit$

nous pouvons proposer la relation

$pomme \xrightarrow{r\_has\_part::substance} fructose$

sachant que la pomme est un fruit qui contient de la fructose et qu'un fruit peut être sucré et fruité. Soit, en termes de l'ontologie à enrichir

pomme `aPourComposantSaveur` *fructose*.

6. Ces concepts ne possèdent pas d'étiquettes dans la version du module dont nous disposons à l'entrée du processus d'enrichissement.

7. Ainsi, $pomme \xrightarrow{r\_carac} croquante$ devient $pomme \xrightarrow{r\_carac::bruit} croquante$, le terme bruit vient contextualiser la relation.

Pour calculer la **convergence** entre les descripteurs sensoriels et les termes cibles de la relations typée « caractéristique » (r_carac) de RezoJDM, nous avons utilisé une méthode basée sur la contextualisation par la sémantique observée au niveau de l'ensemble des termes source de la relation typée r_carac. Si l'ensemble de la sémantique (relations sémantiques sortantes) d'un terme source donné le relie à un type d'aliment et s'il possède un ensemble de caractéristiques partagées avec d'autres termes avec un générique $\approx$ « aliment », l'individu-descripteur non couvert par l'ébauche d'ontologie peut être proposé pour être relié à la super-classe correspondante.

Dans cette partie de l'expérience, nous faisons face à cinq cas de figure distincts :

1. ***intersection*** (#**inter**) : descripteurs présents dans RezoJDM, reliés à un ensemble de termes définis comme $\xrightarrow{r\_isa} \approx aliment$ et à un ensemble de termes définis comme $\xrightarrow{r\_isa} \approx$ { $arôme$, $surface$, $odeur$...} par une relation entrante typée r_carac ;
2. ***candidats+*** (#**cand+**) : descripteurs *SensoMIAM* présents dans RezoJDM reliés à un ensemble de termes définis comme $\xrightarrow{r\_isa} \approx aliment$ mais non reliés à un ensemble de termes définis comme $\xrightarrow{r\_isa} \approx$ { $arôme$, $surface$, $odeur$...}. Ce sont des termes qui pourraient faire partie des instances des classes-descripteurs *SensoMIAM* ;
3. ***candidats-*** (#**cand-**) : descripteurs *SensoMIAM* présents dans RezoJDM mais non reliés à un ensemble de termes définis comme $\xrightarrow{r\_isa} \approx aliment$. Ce sont des termes qui pourraient servir à décrire les aliments mais qui ne sont pas utilisés à cette fin ;
4. ***silences-t*** (#**silences-t**) : descripteurs *SensoMIAM* absents du RezoJDM.
5. ***silences-r*** (#**silences-r**) : descripteurs *SensoMIAM* présents mais relations attendues absentes dans RezoJDM.

La répartition des descripteurs entre les différents ensembles est présentée dans le tableau 5.1.

inter	**cand+**	**cand-**	**silences-t**	**silences-r**
38	24	17	1	35
33%	*20%*	*15%*	*0.86%*	*30%*

TABLE 5.1 – Intersection entre les lexicalisation des individus (instances de classes) *SensoMIAM* et les termes cible de la relation typée r_carac dans RezoJDM.

Exemple 5.2

Exemples de termes appartenant aux différents sous-ensembles.

inter : *légume, coulant, cassant, épicé, nerveux,* etc.

cand+ : *ferme, fort, épais, brun,collant, froid, sombre,* etc.

cand- : *dense, triangulaire, uniforme, hétérogène, sphérique, vert*, etc.

silences-t : *fraîcheur mentholée*

silences-r : *brûlant, jaune, doré*, etc.

L'*intersection* peut servir à peupler la classe *Aliment* et à proposer de nouveaux descripteurs.

Les *candidats+* peuvent permettre d'enrichir les listes de descripteurs si le contraste entre ces termes et les termes déjà utilisés comme descripteurs est suffisamment important (dissimilarité suffisante).

Les *candidats-* peuvent permettre d'enrichir les listes de descripteurs si la similarité entre les termes source de leur relations entrantes typées r_carac est suffisante. Ex. $olive \xleftarrow{r\_carac} couleur$

Les *silences-t* déclenchent les mécanismes exogènes (à circuit ouvert) d'acquisition des termes (validation des contributions en attente, acquisition contributive, acquisition à partir des textes. C'est le cas où le processus qui utilise un réseau lexico-sémantique participe à continuellement améliorer la couverture de ce réseau. Dans le cadre de notre expérience, nous avons découvert 1 terme manquant : *fraîcheur mentholée*.

Les *silences-r* déclenchent les mécanismes endogènes (à circuit fermé) de peuplement en testant les relations inférables.

Dans le cadre de notre expérience, nous nous sommes appuyés sur tous les sous-ensembles. Les sous ensemble intersection et candidats+ ont été les plus productifs.

desc	**#orig**	**#cand**	**% pr (cand)**	**#val**	**% pr (val)**	**%val**
Aspect	33	2 099	6 360%	**145**	**439%**	6%
Flaveur	44	1 157	2 629%	**95**	**215%**	8%
Toucher	38	1 499	3 944%	**102**	**268%**	7%
Total	115	4 284	3 725%	**342**	**297%**	8%

TABLE 5.2 – Proposition des descripteurs.

Exemples des descripteurs proposés pour chaque classe listée dans le tableau 5.2. Ces chiffres laissent apparaître une très grande productivité des méthodes d'acquisition des descripteurs basées sur l'utilisation d'un RLS quant aux candidats proposés par le système avant le filtrage. Le filtrage (avant validation notée **val**) est principalement statistique et logique proche de celui qui a été décrit par Zarrouk [2015] dans le cadre d'inférence par raffinement. Pour l'ensemble des structures comparables entre SensoMIAM et RezoJDM (intersection), ce processus est basé sur le poids des relations RezoJDM et sur l'existence de plusieurs exemples de termes identifiés comme aliments ayant pour caractéristique le descripteur de départ et le descripteur candidat.

Exemple 5.3

Pour le descripteur candidat ***laiteux***, nous avons :

$tiède \xleftarrow{r\_carac} sauce \xrightarrow{r\_carac} laiteux$

$tiède \xleftarrow{r\_carac} boisson \xrightarrow{r\_carac} laiteux$

ainsi que

$blanc \xleftarrow{r\_carac} fromage \xrightarrow{r\_carac} laiteux$

$blanc \xleftarrow{r\_carac} eau \xrightarrow{r\_carac} laiteux$

Par conséquent, nous pouvons proposer le descripteur **laiteux.**

Parmi les sous-classes de `DescripteurFlaveur` :

`DescripteurSensationTrigeminale` = {*laiteux, tannique*}

`DescripteurArome` = {*violette, sucré-salé, persillé, miellé, musqué, moutardé, limoneux, saumuré, vinaigré*}

Parmi les sous-classes de `DescripteurToucher` :

`DescripteurThermique` = {*tempéré*}

`DescripteurTact` = {*pâteux, écailleux, spongieux, spumeux, floconneux, velouté*}

`DescripteurConsistance` = {*pâteux, plâtreux, velouté, pulvérulent, friable*}

`DescripteurComposition` = {*huile, vinaigre, moutarde, miel, vanille, noisette,poivre, persil, saumure, citron, musc, meringue, lait, safran*}

`DescripteurBruit` = {*sourd*}

Parmi les descripteurs d'aspect (`DescripteurAspect`) :

`DescripteurSurface` = {*mat, flambé, saisi, saupoudré*}

`DescripteurSubstance` = {*sirupeux, laiteux, frémissant*}

`DescripteurMacroForme` = {*oblong, rectiligne*}
`DescripteurCouleur` = {*turquoise, rouge et blanc, noir et blanc, tricolore, bicolore, mazouté, mordoré, mauve, multicolore, écru, verdâtre, crème, pêche, indigo, incolore, jaunâtre*}

`DescripteurContraste` = {*tigré, rayé, zébré, tacheté, teinté*}

5.2.2 Conceptualiser à partir d'une ébauche d'ontologie en utilisant un RLS

Outre la construction collaborative par un groupe d'experts (approche descendante), la construction ontologique peut être amorcée sur la base des ressources terminologiques (liste de termes, mémoires de traduction) ou une ébauche d'ontologie. L'expérience d'enrichissement que nous avons décrite permet de se tourner vers une possible aide aux experts quant à la conceptualisation (suggestion des éléments remarquables qui permettent de proposer des classes et des propriétés d'ontologie).

Proposition des descripteurs. Au cours de l'expérience d'enrichissement, nous avons pu voir émerger les classes énumérées candidates suivantes que nous avons nommées (à titre de démonstration) :

- `DescripteurAppréciation` = {*raffiné, horrible, sublime, trop cuit, trop sucré, trop salé, trop liquide, travaillé, mauvais*} ;
- `DescripteurIntensité` = {*concentré, subtile, fort, serré, saturé, sapide, passé*} ;
- `DescripteurEffet` = {*roboratif, laxatif*} ;
- `DescripteurMicroForme` = {*piriforme, ovoïde, patatoïde, tubulaire*}.

En pratique, il serait impossible de proposer ces items à des experts car ce serait contraire à la méthode descendante de construction de MIAM. En effet, les classes sont définies de façon « économique » pour éviter les redondances et garantir ainsi la consistance de la ressource par rapport aux axiomes. Dans le cadre de cette méthode, ces descripteurs-candidats seraient reversés dans une folksonomie [8] adossée à l'outil de construction ontologique. Cette démarche prend toutes son importance dans le contexte multilingue où la proposition des ensembles de « classes » folksonomiques partagées appuieraient la construction collaborative multilingue entre les experts. Par ailleurs, notre démarche se veut générique et applicable à d'autres domaines. Par conséquent, nous validons le principe plutôt que les résultats fermes et définitifs de la démarche.

Proposition des propriétés. L'expérience de proposition des propriétés pour une pseudo-ébauche d'ontologie a permis de produire 6 020 triplets qui relient un aliment à un descripteur présent dans *SensoMIAM*. Parmi ces triplets, certains sont des triplets valides (la relation est sémantiquement vraie). En termes de pertinence, il est nécessaire d'éliminer les termes génériques tels que *soupe*, *préparation culinaire*, *sauce* car ils peuvent potentiellement avoir toute une gamme

8. Une folksonomie est un système de classification collaborative décentralisée spontanée.

de descripteurs et subsument les autres termes plus spécifiques. Pour les détecter, nous avons exploré les relations hiérarchiques typées *r_isa* et *r_hypo* des termes qui correspondent aux aliments. En prenant les différentes arbres hiérarchiques auxquels appartiennent ces termes-aliments, nous avons éliminé les termes nous n'avons gardé que les feuilles de l'arbre hiérarchique avec leur parent direct.

Pour illustrer le mécanisme de proposition des propriétés ontologiques, nous nous sommes restreints aux sous-ensembles des descripteurs *intersection* et *candidats+*.

Nous avons tenté de proposer automatiquement les termes pour peupler les propriétés hypothétiques `aPourComposantToucher`, `aPourComposantAspect`, `aPourComposantFlaveur`. L'inférence des relations correspondant à ces propriétés à partir du RLSM_{PI} se base sur les relations typées *r_has_part* et *r_matter* (partie-tout et partie-tout substance).

Exemple 5.4

Les propriétés hypothétiques sont des relations thématiques. Le domaine de ces propriétés serait une sous-classe de la classe *Aliment* tandis que son co-domaine serait une sous-classe de la classe que l'on pourrait désigner comme *ComposantSensoriel* et qui correspond à l'ensemble des composants qui définissent les qualités sensorielles d'un aliment.

Dans le cadre du modèle RLS (RezoJDM), les termes correspondant au co-domaine de la propriété `aPourComposantSensoriel` doivent permettent de relier le terme qui correspond au domaine de cette propriété à ses descripteurs par triangulation.

Par exemple, si nous avons

pain grillé $\xrightarrow{r\_carac::couleur}$ *doré*

pain grillé $\xrightarrow{r\_carac::bruit}$ *croustillant*

pain grillé $\xrightarrow{r\_has\_part}$ *croûte*

croûte $\xrightarrow{r\_carac::bruit}$ *croustillant*

croûte $\xrightarrow{r\_carac::couleur}$ *doré*

la relation telle que

pain grillé/*r_has_part*/*croûte* $\xrightarrow{r\_annotation}$ *composant toucher*

peut être inférée. Conformément aux règles de correspondance RLS-Ontologie, nous pouvons proposer :

pain grillé `aPourComposantToucher` *croûte*

En effet, pour peupler les propriétés de test, nous utilisons la *généralisation*

méronymique[9] car nous avons *pain grillé* $\xrightarrow{r\_has\_part}$ *croûte* où le tout et sa partie partagent un sous-ensemble de caractéristiques. Sur la base d'un sous-ensemble des caractéristiques subsumées par l'aliment-tout et de la catégorie de la caractéristique (descripteur) qui prend forme de la relation r_carac qui va du terme qui désigne une caractéristique sensorielle (*saveur* $\xrightarrow{r\_carac}$ *umami*, *bruit* $\xrightarrow{r\_carac}$ *sourd* etc), nous pouvons identifier les structures qui permettent de proposer les propriétés à valeur objet d'ontologie.

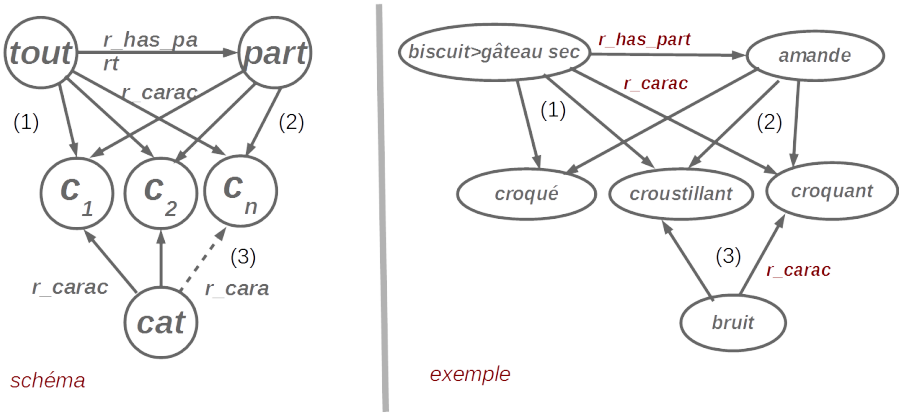


FIGURE 5.1 – Schéma et exemple de détection des structures nécessaires la proposition automatique des descripteurs.

Nous avons introduit la distinction entre trois grands ensembles de descripteurs que nous avons explorés et nous avons exprimé le schéma (figure 5.1) sous forme d'une règle de production (définition 5.1.).

Définition 5.1

Règle de production de propriété d'ontologie basée sur le parcours d'un RLS.

(1)

$$\forall x \xrightarrow{r\_has\_part} y \wedge x \xrightarrow{r\_carac::cat} \{c_1, c_2, ..., c_n\}$$

(2)

$$\exists y \xrightarrow{r\_carac::cat} \{c_1, c_2, ..., c_n\}$$

(3)

$$\exists cat \xrightarrow{r\_carac} \{c_1, c_2, ..., c_n\}$$
$$\Rightarrow x \xrightarrow{r\_has\_part::cat} y$$

9. Nous entendons par **généralisation méronymique** une relation fondée sur la relation partie-tout et les caractéristiques qu'un tout et sa partie peuvent partager. Il s'agit de généraliser à un tout certaines caractéristiques de ses parties. Par exemple, « tout aliment qui contient du sucre est sucré », « tout pain qui a une croûte est caractérisé par un bruit croquant » etc.

propriété (règle)	**#instances**	**#prop**	**% prop**	**#val**	**%val**
aPourComposantFlaveur	22 751	8 576	38%	857	10%
aPourComposantToucher	31 679	11 941	19%	441	7%
aPourComposantAspect	45 798	17 263	23%	410	4%
Total	81 908	30 874	-	1709	-
Moyenne	-	-	27%	-	7%

TABLE 5.3 – Proposition des propriétés **aPourComposant**.

Dans le tableau 5.3, #**instances** correspond au nombre de structures qui satisfont la prémisse notée (1) de la règle ; #**prop** correspond au nombre de structures qui satisfont la prémisse (2) de la règle et permettent de fournir la conclusion de la règle sémantiquement valide, #**validés** sont des propositions ontologiquement valides (la contrainte sur le co-domaine de la relation est remplie, l'intersection entre l'ensemble des relations typées *r_ carac* de l'aliment-tout et celles de composant-partie est est de cardinalité suffisante, elle permet de faire émerger une catégorie de composants sensoriels (Toucher, Aspect ou Flaveur)(3). Exemples des propriétés hypothétiques proposées :

— *veau Orloff* `aPourComposantFlaveur` *lard* à partir de {*gras, viande*} ;

— *jus* `aPourComposantFlaveur` *fruit* à partir de {*sucré, acide, sirupeux*} ;

— *gratin savoyard* `aPourComposantAspect` *fromage* à partir de {*gratiné, gras, brûlé*} ;

— *sauce* `aPourComposantAspect` *graisse végétale* à partir de {*gras, fluide,nappant*} ;

— *baclava* `aPourComposantToucher` miel *à partir de* {*collant, parfumé, fondant*} ;

— *thé* `aPourComposantToucher` *eau bouillante* à partir de {*chaud,bouillant*}.

Les processus de construction d'un RLS ainsi que les processus d'enrichissement et de conceptualisation d'ontologie fondés sur l'utilisation de celui ci font partie du système exploitant cette ressource. Dans la section suivante, nous allons résumer les principes de fonctionnement d'un tel système.

5.3 Système exploitant RLSM$_{PI}$ en tant que système multi-agent (SMA)

De façon générale, un système multi-agent s'appuie sur une population d'agents autonomes qui se trouvent en interaction directe ou indirecte. Les mécanismes de communication et d'interaction entre les agents sont globalement de deux types :

— envoi des messages directs ;

— modification de l'environnement (par exemple, stigmergie ou imitation des traces de phéromone dans les systèmes bio-inspirés) [10].

10. Par exemple, les algorithmes dits *à colonies de fourmis* « imitent » le cycle de vie des

En ce qui concerne son architecture, le système d'exploitation du RLSM$_{\text{PI}}$ se rapproche du système défini par Ferber [1995] comme une architecture basée sur les *systèmes de production*.

> *Un système de production est défini par la combinaison d'une base de faits (BF), d'une base de règles de production (BR) et d'un interprète, le moteur d'inférence (MI).* [11]

Les règles de production ont la forme générale

« si `liste de conditions` alors `liste d'actions` ».

Le choix d'un SMA est motivé par le fait que ce type de système permet de faciliter la modélisation car il permet la mise en œuvre des tâches multiples s'exécutant en parallèle où chacune des tâches est relativement peu complexe.

Nous distinguons deux types d'agents dans le cadre de nos expériences : *agent endogène* (figure 5.2) qui modifie le RLSM$_{\text{PI}}$ et *agent d'analyse* (figure 5.3) dont les actions portent sur le graphe d'analyse (dans le cadre d'analyse sémantique des textes). Lorsqu'il est produit, l'agent « connaît » RLSM$_{\text{PI}}$ et dispose d'une liste de critères (conditions) qui définissent l'action (ou les ensembles d'actions) qu'il produit. Cette liste de conditions peut correspondre à une règle qui décrit un agent ou à un ensemble de conditions fourni par un programme ou un ensemble d'observations sur le RLSM$_{\text{PI}}$.

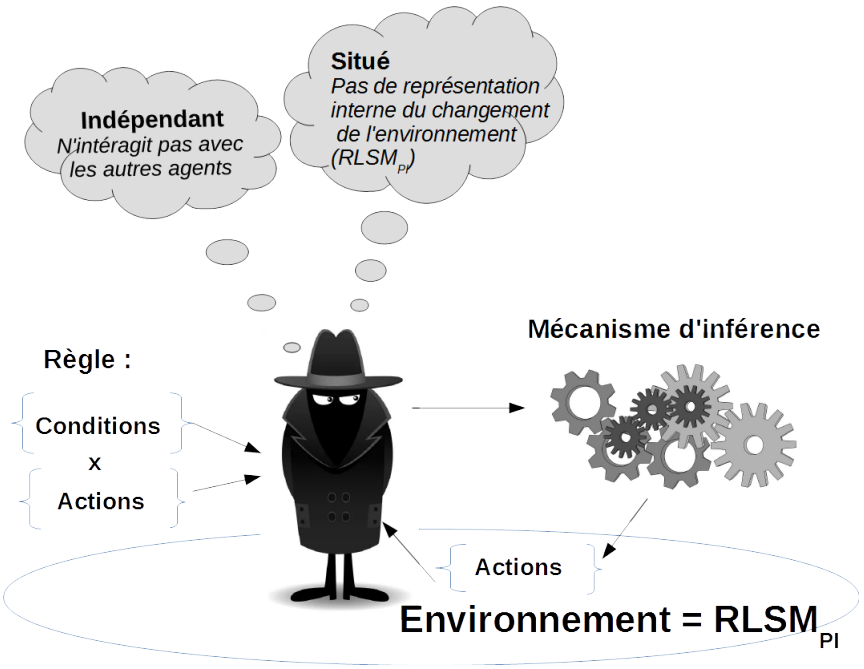


FIGURE 5.2 – Agent au sein du système multi-agent d'exploitation du RLSM$_{\text{PI}}$

fourmis artificielles ainsi que leur manière de communiquer en laissant des traces de phéromone qui peuvent se renforcer et s'évaporer.

11. *Op.cit.* p. 137.

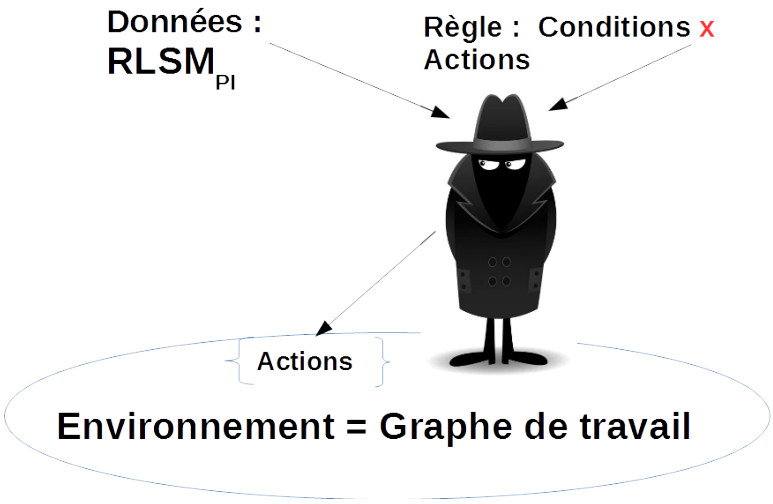


FIGURE 5.3 – Agent d'analyse des textes relié au système d'exploitation du RLSM$_{PI}$

Conformément à cet objectif, sur le plan fonctionnel, chaque agent RLSM$_{PI}$ remplit un rôle précis. Il s'agit, en particulier, des rôles suivants :

- ***agglomération*** : association en unités polylexicales ;
- ***recherche et recopie*** : recherche des termes et des relations dans le but de les utiliser dans un processus d'analyse comme, par exemple, l'analyse sémantique des instructions, leur recopie dans le graphe d'analyse (réseau d'analyse).
- ***recherche et modification*** : recherche des termes et des relations et modification des informations telles que le poids de la relation, la liste des origines du terme ;
- ***recherche et création*** : vérification de l'existence d'un élément et création d'un nouvel élément ;
- ***fusion*** : rapprochement des termes au sens proche dans le réseau ;
- ***différenciation*** : distinction de plusieurs relations ou termes et création des éléments correspondants. par exemple, dans le cas de raffinement d'un terme ou de précision des relations via notamment l'annotation ;
- ***inférence*** *des relations*, proposition des nouvelles relations à partir de celles observées entre les termes présents dans le texte et leur voisinage.

Exemple 5.5

Illustration des différents rôles des agents. Dans cadre de l'énoncé *Battre les œufs en omelette et les mélanger aux Carré Frais et aux herbes de Provence.*

- ***agglomération*** des termes *herbes de Provence*, *en omelette* à partir des critères purement syntaxiques et des patrons lexico-sémantiques [a] ;
- ***recherche et recopie*** des termes : *battre*, *œuf*, *omelette*, *mélanger*, *Carré Frais*, *herbes de Provence* et de leurs relations *battre* $\xrightarrow{r\_object}$ *œuf*, *omelette* $\xrightarrow{r\_has\_part}$ *œuf*, *battre* $\xrightarrow{r\_manner}$ *en omelette* etc.

— en supposant dans le cadre de l'exemple que le terme *Carré Frais* n'existe pas dans le RLSM_{PI}, ***création*** du terme *Carré Frais*

Plus généralement :

— ***fusion*** : rapprochement des sens des termes `in:white`, `in:albumen`, `in:glair` en qualité de terme couvrant interlingue par opposition à `in:protein`. En effet, les quatre termes couvrent le terme `ru:белок` qui est polysémique : `ru:белок>blanc d'oeuf` et `ru:белок>protéine`;

— ***différenciation*** des sens du terme `ru:белок` [b] fondé sur ses multiples termes couvrants ;

— ***inférence*** des relations sémantiques par les agents qui agissent d'une manière isolée en se basant sur l'état du réseau tel qu'ils le trouvent au moment où ils sont produits.

a. Les patrons lexico-sémantiques décrits notamment par Ramadier [2016] combinent les patrons syntaxiques dits de surface et l'identification des relations sémantiques dans un réseau lexico-sémantique tel que le RLSM_{PI}. Par exemple, dans le cas du terme *salade de chou*, le patron lexico-sémantique serait « N_1 PREP N_2 & $N_1 \xrightarrow{r\_has\_part} N_2$ ».

b. ce terme se traduit à la fois par"blanc (blanc d'œuf)" et par "protéine"

Un système multi-agent d'exploitation du RLSM_{PI} permet de répondre à des besoins qui peuvent surgir dans les cas suivants :

— acquisition des connaissances qui nécessite des échanges entre les experts et les ontologues ;

— acquisition des connaissances auprès des traducteurs et terminologues et à partir des textes de spécialité dans le contexte industriel ;

— construction termino-ontologique semi-automatisée.

Par conséquent, le SMA peut se décliner en fonction du mode d'acquisition des ensembles de conditions qui mènent à des actions produites par les agents. La démarche centrale présentée dans ce mémoire a concerné la génération automatique des conditions basée sur les observations des termes et relations présentes dans le RLSM_{PI} (ou du RLS monolingue). Dans le contexte applicatif qui sera présenté à titre d'ouverture, dans la suite du présent chapitre, nous explorons l'acquisition des conditions auprès des experts (via une interface de création et de test des règles de construction termino-ontologique).

5.4 Aide à la construction ontologique (ACO) : un outil d'assistance

La présente section est centrée sur une application d'aide aux spécialistes en ingénierie de connaissances et aux experts du domaine :

— via une ***interface*** qui permet de saisir et de valider les règles de création des *Object Properties* d'ordre ontologique lors de la construction termino-

ontologique[12] manuelle. La particularité de ces règles réside dans le fait qu'elles peuvent être spécifiées en utilisant une langue parmi les langues du RLSM_{PI} et permettent d'obtenir des propositions de concepts (à titre expérimental et hypothétique) et de *Object Properties* qui reflètent une conceptualisation partagée (ex. structures présentes dans plusieurs langues du RLM_{PI} ainsi que dans le pivot) si cela est nécessaire ;

— via un ***processus automatique*** d'enrichissement de module d'ontologie (processus valable pour une ébauche d'ontologie) où les conditions d'inférence des classes et *Object Properties* sont générées automatiquement à partir du module d'ontologie fourni et des éléments (termes et relations) présents dans un réseau lexico-sémantique (RLS).

Dans le contexte d'aide à la construction termino-ontologique, l'expert du domaine est au centre du processus. Ses besoins vis-à-vis du système d'exploitation d'une ressource de connaissance sont principalement les suivants :

— visualisation du contenu du RLSM_{PI} ;

— recherche sémantique riche (avec RLSM_{PI} comme base de connaissances il s'agit de la recherche qui utilise un index augmenté grâce aux relations sémantiques des termes de la requête identifiés dans la base) dans le but d'analyse des segments textuels ;

— conceptualisation, par exemple, proposition des relations d'ordre ontologique en fonction des critères fournies par un expert.

Le système d'aide à la construction ontologique (ACO) répond à ces besoins en proposant :

1. des possibilités de visualisation du contenu du RLSM_{PI} ;
2. une recherche multicritères ;
3. une interface spécifique de saisie de règles de conceptualisation (relations thématiques d'ontologie).

Compte tenu du caractère peu innovant des deux premiers aspects, nous nous focalisons sur la proposition d'interface de conceptualisation. Le défi central de celle-ci est de concilier la simplicité d'utilisation et la pertinence des critères obtenus pour la construction termino-ontologique.

Principe. Le principe de fonctionnement d'un environnement ACO est de permettre à un ou plusieurs experts de fournir un exemple simple ou enrichi qui permet de spécifier une *Object Property* en utilisant une des langues naturelles du RLSM_{PI} et d'obtenir des structures similaires à l'exemple fourni. Ces structures peuvent trouver leur expression à travers les termes présents dans les langues naturelles du RLSM_{PI} ou à travers les termes interlingues (conceptualisation partagée) et leurs relations.

12. Construction des ressources ontologiques dont les étiquettes sont des unités terminologiques.

Déroulement. De façon générale, le point de départ du processus est la saisie des informations (étiquettes de termes, types de relations choix des langues) dans un interface dédiée. Deux cas de figure sont possibles : le cas simple qui ne comporte pas de paramètres qui sont des variables proprement parler et le cas complexe qui peut comporter des variables.

Cas simple de recherche d'exemples. Dans le cas simple, l'expert fournit un triplet qui relie des classes désignées par un terme : domaine (étiquette de classe), nom de propriété existante (propriété qui a déjà été immergée au sein du RLSM_{PI}), co-domaine (étiquette de classe).

Ce processus est utile lorsqu'une propriété (relation thématique) de l'ontologie de référence a été spécifiée mais peu renseignée. Tel est le cas de la propriété à valeur objet `aPourProduitAuxiliaire`.

L'ensemble des relations sémantiques (dont les hyperonymes et les hyponymes) des termes source et cible de la propriété fournie ainsi que éventuellement l'annotation de la relation sont considérés comme une instance de règle d'inférence par abduction. Cette instance de règle est la structure qualifiante qui permet d'identifier et de proposer de nouvelles instances de règle. Les triplets correspondant à ces instances sont proposés à la validation.

A titre d'exemple, nous pouvons considérer le cas de la propriété `aPourPays` qui permet d'identifier les spécialités gastronomiques et, en particulier, le triplet *minestrone* `aPourPays` *Italie*.

FIGURE 5.4 – Interface simple de recherche d'exemples.

Suite à la saisie des étiquettes de termes source et cible correspondants et le choix de la propriété à tester dans a liste déroulante, le système va d'abord rechercher les termes source (notés x) qui, par imitation des termes fournis,

— ont pour hyperonyme : $x \xrightarrow{r\_isa} \{soupe,\ entrée,...\}$;

— ont des relations sémantiques telles que
$x \xrightarrow{r\_carac} \{liquide,\ chaud\}$;
$x \xrightarrow{r\_matter} \{umami,\ légume,\ viande\}$
$x \xrightarrow{r\_object} \{verser,\ réchauffer, consommer\}$.

et les termes cible (que nous pouvons noter y qui peuvent aussi être identiques au terme saisi en entrée (*Italie*) ou qui ont des relations suivantes :

— ont pour hyperonyme : les termes tels que *pays* ;

— ont des relations sémantiques telles que
$y \xrightarrow{r\_has\_part} \{culture, tradition,\ cuisine,\ gastronomie\}$.

x et y peuvent avoir une relation typée $r\_location$ qui peut être annotée ou non. De plus, le système explore les relations typées $r\_location$ *existantes* et *inférables*.

Il est supposé que les ensembles dont le terme source et cible font partie sont des ensembles disjoints.

Avec le cas simple `aPourPays` et la langue cible *anglais*, nous obtenons le résultat présenté dans la table 5.4.

#inst origine	**#proposés**	**%productivité**	**#val**	**%val**
41	75	182%	52	69%

TABLE 5.4 – Cas d'application du schéma simple.

Dans le cadre du schéma simple (table 5.4, nous avons recherché des spécialités culinaires des pays différents dont le générique serait *soupe* (ou terme équivalent au sein du $RLSM_{PI}$). La validation a été faite manuellement sur la base des critères suivants : *soupe épaisse, spécialité d'un pays, soupe aux légumes*. Le terme *Italie* pourrait permettre d'identifier les spécialités des pays latins (Méditerranée) uniquement mais l'état de notre ressource (récente qui nécessite encore d'être peuplée) ne nous permet pas de fournir ce degré de précision au niveau des propositions du système.

Exemple 5.6

Exemples des propriétés proposées à partir des relations présentes dans le $RLSM_{PI}$ de forme « $x \xrightarrow{r\_location} y$, $x \xrightarrow{r\_location} pays$, $y \xrightarrow{r\_isa} pays$ » :

marmite `aPourPays` *France*

mulligatawny `aPourPays` *India*

pepper pot `aPourPays` *United States*

borshch `aPourPays` *Russia*

gumbo `aPourPays` *United States*

Exemple des propriétés proposées à partir de relations inférables, utilisées dans le calcul :

cocky-leeky `aPourPays` *UK*

à partir de

$cocky\text{-}leeky \xrightarrow{r\_location} Scotland$

Cas complexe. Dans le cas complexe, l'expert ne connaît que partiellement ou ignore les lexicalisations qui pourraient correspondre à la classe « domaine » et à la classe « co-domaine » de la relation d'ontologie.

Il peut spécifier de façon plus ou moins détaillée les hyperonymes (structure d'appartenance) et d'autres relations sémantiques (structure de contextualisation) importantes pour la recherche des lexicalisations et des structures sémantiques qui équivaudraient à des instances de la propriété d'ontologie en train d'être construite.

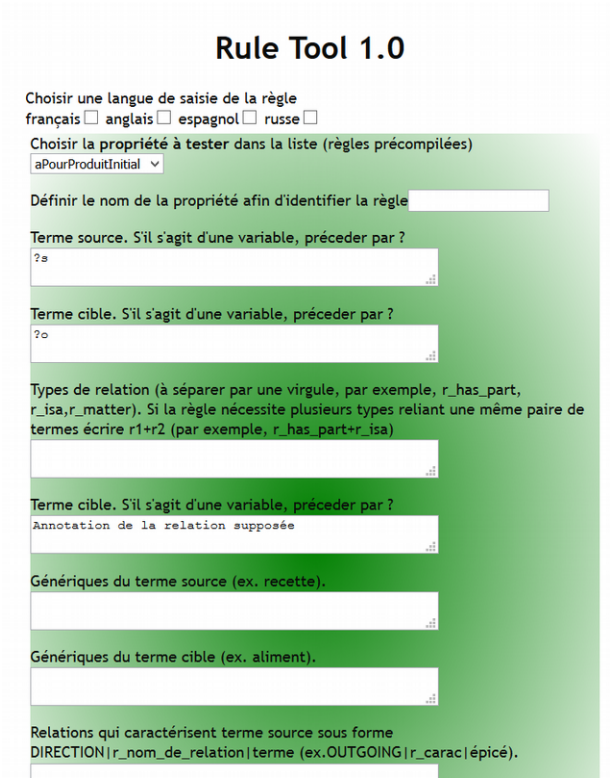


FIGURE 5.5 – Vue de formulaire de saisie de la règle de production dans le but de tester la propriété d'ontologie en cours de construction.

A titre d'ouverture un autre type d'approche pourrait être mis en œuvre quant à la saisie de la règle de production. Elle pourrait prendre forme d'un exemple ou d'une série d'exemples telle que « pâte & sabler, pâte & pétrir, pâte & mixer » si l'on souhaite mettre en évidence les différentes façons de mélanger les ingrédients

d'une pâte et découvrir les différents aspects sensoriels et de composition qui y sont liés. La chaîne de caractères serait transformée en règle de production pour être traitée conformément au mécanisme décrit dans le chapitre 3 du présent mémoire.

Une ouverture sous forme de *chat-bot* (agent conversationnel) destiné aux experts semble également une piste intéressante. En effet, ce type d'outil permettrait la construction intuitive et contributive des ressources termino-ontologiques adossées à l'ontologie dans lesquels les experts pourraient puiser les termes et les relations utiles pour la conceptualisation.

L'avantage d'une saisie via l'interface dédiée réside dans la possibilité d'exprimer naturellement sa vision de la future propriété d'ontologie. De façon plus globale, l'intérêt d'un système ACO est de pouvoir tester et affiner les propriétés en cours de création via la visualisation et la validation des cas inférables de ces propriétés. Ce dernier point semble très utile dans le cas de construction des ontologies dites *core* (indépendantes d'une langue donnée) car, en effet, leur propriétés peuvent être modélisées spécifiquement sur la base du pivot interlingue.

A titre d'exemple, considérons la propriété hypothétique telle que `aPourManièreTransformation` que nous souhaitons tester en tant que propriété potentielle d'une ontologie noyau (sur les termes présents dans le pivot interlingue). Cette propriété permettrait de découvrir des classes d'aliments en termes de manière (relation typée r_manner) qui relie le terme qui désigne l'action à la manière typique dont cette action peut être effectuée. La partie du discours associée à la manière typique est l'adverbe. Plusieurs types d'adverbes peuvent être distingués. Notamment, dans le cadre de du projet SMM, Mahlow and Piotrowski [2009] ont spécifié les types d'adverbes suivants : général, temporel, modal, local.

Exemple 5.7

Au niveau du pivot interlingue, nous observons les structures suivantes dont nous avons pu annoter la relation typée r_manner :

$cooked\ potato \xleftarrow{r\_object} cut \xrightarrow{r\_manner::shape} \{evenly,\ unevenly,\ thinly,\ coarsely,\ horizontally,\ lengthwise\}$

$cooked\ potato \xleftarrow{r\_object} cut \xrightarrow{r\_manner::speed} \{just,\ immediately,\ rapidly\}$

$cooked\ potato \xleftarrow{r\_object} cut \xrightarrow{r\_manner::aspect} \{individually,\ diagonally\}$

$dairy\ product \xleftarrow{r\_object} whisk \xrightarrow{r\_manner::mode} \{gradually,\ constantly,\ slightly\}$

$dairy\ product \xleftarrow{r\_object} drain \xrightarrow{r\_manner::timemode} \{once,\ twice,\ meanwhile,\ together,\ then\}$

Sur la base de ces structures, les ébauches de classes ontologiques suivantes peuvent être proposées automatiquement à l'attention d'un expert : `aPourManièreTransformation` {découpe rapide, découpe lente, cuisson rapide, cuisson lente, drainage répété...}

#premisses	#proposés	%proposés	#val	%val
230 850	46 170	20%	3 220	0,7%

TABLE 5.5 – Cas d'application du schéma complexe pour la détection des instances de **aPourManièreTransformation**.

L'expérience utilisant le schéma complexe dont le résultat est présenté dans la table 5.5 s'est focalisée sur la suggestion automatique d'une prorpriété à des expert à partir des relations sémantiques (ici r_manner) contenues dans le réseau lexico-sémantique mais non couvertes par l'ontologie à enrichir et potentiellement pertinentes pour celle ci. L'interface permettant de décrire une propriété d'ontologie non encore existante, le système ACO est destiné à supporter cet type de démarche.

Le faible pourcentage des propositions validées automatiquement et manuellement est dû à la polysémie observée au sein du RLSM_{PI} ainsi qu'au caractère répétitif des structures.

L'objectif du système exploitant le RLSM_{PI} est l'amélioration continue de la ressource par les différents processus qui l'utilisent. Cette amélioration comprend l'ajout de nouveaux termes et de nouvelles relations, la modification des relations existantes afin de les contextualiser (ajout d'une annotation), de les invalider (attribution du poids négatif à une relation jugée fausse), de les valoriser (augmentation du poids de la relation).

Les ensembles de tâches de construction, d'évaluation, d'exploitation d'une ressource multilingue s'appuient sur des ensembles de critères multiples. En effet, des critères linguistiques mais aussi logiques et statistiques peuvent être pris en compte dans le cadre d'un seul et même processus (analyse des textes, extraction terminologique, inférence des relations, conceptualisation etc.). Un système multi-agent est, par conséquent, adapté à ce type de contexte.

5.5 Analyse des résultats et discussion

Les expériences présentées tout au long du présent chapitre permettent de détailler les points principaux liés à l'utilisation d'un RLS dans le cadre de la construction termino-ontologique.

Le point saillant des expérimentations est la très grande productivité des différentes méthodes qui utilisent un RLS couplée à un très faible taux de proposition validées par filtrage statistique et logique. Elles permettent en effet de

parcourir une très grande quantité de données (exemples) avant de fournir des propositions, une démarche impossible à faire manuellement compte tenu de la quantité de relations à traiter. La structure d'un RLS (termes et relations soit des composants discrets qui permettent de modéliser les aspects sémantiques variés) permet un filtrage logique efficace en particulier lorsque l'on s'appuie sur les relations sémantiques verticales telles que la méronymie.

Compte tenu de divers mécanismes de filtrage et des caractéristiques du $\mathrm{RLSM_{PI}}$, la complexité des algorithmes présentés mérite d'être détaillée.

Définition 5.2

Complexité des algorithmes de filtrage logique.

Opération de base : l'opération de base que nous avons considérée est la recherche informée d'une relation. L'information dont on dispose dans le cadre de cette recherche sera, selon les différents types de filtrage, les types de relations à parcourir et le terme opposé de la relation.

Particularités quantitatives et structurelles du $\mathbf{RLSM_{PI}}$:

- degré moyen constaté d_{av}=4 ;
- présence des *hubs*, termes très fortement connectés tels que, par exemple *animal*, *ADN* etc. ;
- graphe petit monde : diamètre (excentricité maximale de ses sommets) $D \approx 6$. Par conséquent le parcours du réseau s'effectue en largeur et en profondeur avec une profondeur maximale de parcours fixée à 2.

Selon le type de parcours de graphe choisi, **la complexité d'une opération de filtrage** peut être déterminée :

- par son *facteur de branchement* (degré du terme de départ) b et par la limite de profondeur définie l : $O(b^l)$ dans le cas de parcours en profondeur (ex. filtrage par triangulation). Si $l = 2$, la complexité de l'opération de filtrage serait $O(n^2)$;
- par son facteur de branchement b et par la profondeur du terme d (soit la distance jusqu'au terme à tester le plus proche) : $O(b^{d+1})$ (ex.filtrage par abduction). Étant donné qu'on teste les voisins du terme de départ, $d = 1$, la complexité de l'opération de filtrage dans ce cas serait également $O(n^2)$.

Les prémisses d'une règle qui définissent le parcours à mettre en œuvre dans le cadre du filtrage logique sont destinées à réduire le nombre de relations à tester (et la complexité) en ne parcourant que certains types de relations. Cette opération de recherche des relations doit être exécutée m fois. Ainsi, la complexité globale de filtrage logique (sans spécification de type de filtrage) serait de $O(m \times n^2)$

La **complexité moyenne** correspondrait dans ce cas au degré moyen du $\mathrm{RLSM_{PI}}$: $d_{av} = 4 \Rightarrow O(16 \times m)$.

La **complexité dans le pire des cas** serait déterminée par le degré du *hub* le plus important du RLSM_{PI}. Le degré de ce hub est $d_{max} = 33085$. Par conséquent, la complexité dans le pire des cas serait de $O(m \times 1094617225)$.

La complexité théorique dans le pire des cas serait $O(n^3)$. Cependant ce cas de figure est très peu probable car cela voudrait dire que le terme n aurait dans son voisinage tous les termes du RLSM_{PI}.

Dans le cadre de nos expériences concernant, notamment, les descripteurs, la complexité moyenne du filtrage logique pour les descripteurs *Aspect* qui a généré le plus de candidats serait 2 099 $\times$ 16= 33 584 soit $1,461698175 \times 10^{-8}$ la complexité dans le pire des cas. Pour la proposition des propriétés à valeur objet [13] (à titre hypothétique), dans le cas de la *Object Property* `aPourComposantAspect` qui a généré 45 798 instances à filtrer, la complexité de filtrage logique a été de 45 798 $\times$ 16 = 716 768 soit $1,429781974 \times 10^{-8}$ la complexité dans le pire des cas.

Conclusion du chapitre

Dans ce dernier chapitre, nous avons proposé une ouverture vers un ensemble d'outils et méthodes d'aide à la construction des ressources termino-ontologiques. Contrairement au chapitre 3, dans le présent chapitre nous nous sommes focalisée non plus sur les méthodes d'aide à l'acquisition des éléments remarquables (traces éventuelles de concepts et propriétés d'ontologie au sein d'une ressource de connaissance langagière), mais sur les méthodes d'aide à la modélisation, notamment lorsqu'il s'agit de la modélisation par les experts. Nous avons détaillé un outil qui permet à un expert de rechercher des exemples (instances) d'une propriété en train d'être modélisée. Nous avons également exploré l'hypothèse d'ébauche d'ontologie pour laquelle la constitution d'une base de suggestions obtenues par calcul sur le réseau lexico-sémantique et soumis à un expert peut être utile. Les perspectives d'évolution de ce type d'outils se situent dans le domaine des outils collaboratifs et interactifs (tels que les agents conversationnels à intelligence artificielle) qui permettraient à une communauté d'experts d'un domaine donné de partager leurs connaissances et leur vision du modèle à construire de façon spontanée.

Contributions du chapitre 5

Les contributions du présent chapitre concernent les perspectives d'évolution quant à l'utilisation des ressources de type « réseau lexico-sémantique » pour la construction des ressources termino-ontologiques. Elles peuvent être résumées comme suit :

13. *Object Properties.*

— proposition d'une méthode d'aide à la construction d'ontologie (ACO) destinée aux experts humains ;

— proposition d'une méthode de modélisation à partir d'une ébauche d'ontologie.

Ces deux approches sont basées sur l'exploitation d'une ressource de connaissance langagière.

Conclusion

Dans le présent mémoire nous avons décrit nos travaux qui ont concerné la modélisation des connaissances issues du domaine de la cuisine et de la nutrition sous forme d'un réseau lexico-sémantique multilingue avec pivot interlingue ($RLSM_{PI}$) et l'exploitation de cette ressource non ontologique pour accompagner la construction et la localisation d'ontologie.

L'architecture et les méthodes de construction du $RLSM_{PI}$ ont été proposées après une étude attentive des ressources de connaissance existantes et dans l'optique d'interopérabilité sémantique avec certaines de ces ressources. Le peuplement a été, et continue à être, effectué par les ensembles de méthodes qui agissent de manière indépendante sur le $RLSM_{PI}$ afin de l'améliorer continuellement.

Nous avons défini et mis en œuvre une méthode de construction d'une ressource structurée qui pourrait, dans le cadre de la construction d'ontologie à partir des textes en langue naturelle, prendre la place du corpus et fournir des structures sémantiques pour la construction d'ontologie par les experts humains sous forme d'une ressource médiatrice. Cette ressource serait alors la projection du modèle souhaité sur le $RLSM_{PI}$. Si, dans le cadre des domaines de spécialité comme celui que nous avons abordé dans le présent travail, il semble impossible de se passer de la participation des experts du domaine (cuisiniers, nutritionnistes), nous considérons que l'effort humain quant à l'acquisition des connaissances pour la construction d'ontologie ainsi que la structuration de ces connaissances en vue de modélisation ontologique peut être réduit grâce à l'utilisation de ressources de connaissance structurées. Le contexte industriel de notre travail a conditionné également notre engagement en faveur des méthodes automatiques d'acquisition des connaissances.

Contributions

Outre la ressource de connaissance structurée et spécifique au domaine de la cuisine et de la nutrition que nous avons obtenue à l'issue de nos expérimentations, notre contribution à l'état de l'art concerne les aspects suivants :

— définition de la méthode de construction de la ressource de type réseau lexico-sémantique qui comprend aussi bien des méthodes d'acquisition exogènes (extraction depuis les corpus, intégration de ressources existantes)

que endogènes (inférences y compris inférences cross-lingues) ;

— application de méthodes d'inférence sur les réseaux lexico-sémantiques à la construction d'une ressource multilingue ;

— proposition d'un protocole d'évaluation de la ressource de type réseau lexico-sémantique spécialisé qui comprend l'estimation de son apport à l'état de l'art (apport de connaissance par rapport aux ressources existantes, gestion de la polysémie) et la tâche d'analyse sémantique ;

— proposition de la passerelle inter-modèle entre l'ontologie et le réseau lexico-sémantique : le modèle d'ontologie est acquis par immersion dans le réseau lexico-sémantique, les éléments et les structures pertinentes par rapport à ce modèle sont découverts dans le réseau, puis extraites sous forme d'une ressource médiatrice ;

— proposition des méthodes et outils pour accompagner la construction d'ontologie par les experts basée sur une ressource non ontologique.

Perspectives

Amorçage automatique d'une ressource de spécialité de type $\mathbf{RLSM_{PI}}$

L'interopérabilité des ressources continue à se heurter à la problématique d'interopérabilité de format et de difficulté à rendre interopérables via les formats du Web Sémantique certaines ressources de connaissance historiques riches en données de qualité. En effet, le passage dans les formats de donnés liées peut entraîner une perte d'information considérable. Il s'agit parfois de mise en œuvre trop complexe car les modèles des ressources lexicales et sémantiques contiennent les informations difficiles à exprimer.

Étant donné que les ressources structurées de base restent souvent les mêmes et que souvent ces ressources existent dans un format non interopérable mais permettent un accès programmatique spécifique, la démarche de transition (en attendant la généralisation de l'interopérabilité de format) pourrait être de :

— concevoir une ressource de type RLS qui intégrerait les ressources structurées par immersion et qui permettrait des sorties (projections) sous forme des ressources médiatrices en format souhaité dont un format de données liées ;

— fournir via l'exploitation de ce type de ressource un accès programmatique unique à un ensemble de données contenues dans les ressources de connaissances structurées éventuellement enrichi grâce aux extractions depuis les corpus de spécialité.

Nous avons pu constater lors de nos expérimentations la polyvalence de la structure sous forme de graphe orienté, typé et valué (réseau lexico-sémantique inspiré de RezoJDM). La structuration de ce type de réseau basée sur la sémantique fournit le degré d'abstraction nécessaire pour représenter les connaissances issues des ressources existantes et des textes. La présence des méta-structures sous

forme d'éléments discrets (nœuds, arcs) permet de spécialiser à souhait le réseau et d'éviter lors de l'intégration la perte des (méta)informations qui constituent la valeur ajoutée des ressources de connaissance structurées.

Formalisation de la méthode de construction des architectures avec pivot évolutif.

Dans la littérature, le choix est imposé entre l'utilisation d'un pivot naturel ou l'utilisation d'un pivot artificiel (interligue). Une des perspectives relatives à notre travail serait de formaliser une méthode qui permet d'amorcer le pivot en tant que pivot naturel afin qu'il se transforme en pivot interlingue de façon incrémentale, au fur et à mesure de l'évolution de la ressource. Cette démarche s'inscrit dans les tendances actuelles de conception industrielle des produits. Elle semble compatible avec l'objectif de réduction de coût humain lors de la conception des ressources multilingues.

Formalisation et test des méthodes d'évaluation des ressources de type réseau lexico-sémantique multilingue.

Les ressources lexico-sémantique sous forme de réseau sont des graphes pas comme les autres. En effet, il s'agit de graphes dont les arcs sont orientés, typés, comportent des méta-informations comme des annotations, des poids etc. Les critères d'évaluation propres à la théorie des graphes s'appliquent, mais restent peu informatives quant à la qualité d'une ressource lexico-sémantique sous forme de graphe. La modélisation de la polysémie et de la sémantique d'une langue naturelle, la qualité des liens translingues (dans le cadre d'une ressource multilingue) sont les critères qui doivent permettre de déterminer la qualité d'une telle ressource. Une évaluation par référence (comparaison à une autre ressource) est souvent compliquée notamment pour des raisons linguistiques ou pour des raisons de format. Par conséquent, un protocole d'évaluation statistique propre défini pour ce type de ressources pourrait constituer une piste d'évolution d'une partie de nos travaux. Un protocole d'évaluation par la tâche mérite également un développement ultérieur.

Perfectionnement de l'outil d'aide à la construction d'ontologie sous forme d'agent conversationnel

La construction des ressources termino-ontologiques devient de plus en plus collaborative, axée sur le partage des informations. Dans ce contexte, l'échange informel entre les experts est favorisé. Lorsque l'on souhaite utiliser un système d'aide à la construction d'ontologie qui se charge d'acquérir des connaissances à partir des textes et des ressources structurées ainsi que de proposer des structures sémantiques potentiellement intéressantes en vue de modélisation d'ontologie, il semble indispensable pour un expert humain de pouvoir communiquer avec ce

système de façon la plus spontanée possible. À ce titre, la conception d'un agent conversationnel (*chatbot*) basé sur la connaissance (le RLSM_{PI} qui a fait objet de nos travaux) semble être un développement naturel de nos expériences décrites dans le dernier chapitre du présent mémoire.

Liste des tableaux

Table des figures

Annexe A

Fonctions et Algorithmes

A.1 Fonctions

La structure de RLSM_{PI} suppose que chaque nœud possède une liste de relations sortantes et une liste de relations entrantes. Les nœuds faisant partie du pivot interligue comptent parmi leurs relations sortantes la relation spécifique typée r_covers ; par conséquent ils possèdent une liste des nœuds *couverts* (toutes langues confondues). Les nœuds appartenant aux graphes des langues ont, parmi leur relations entrantes, les relations typées r_covers et possèdent une liste des nœuds interlingues *couvrants*.

Les fonctions de base sont liées à la manipulation du RLSM_{PI} et, d'autre part, au fonctionnement de la passerelle RLSM_{PI} - Ontologie.

Notation. Pour alléger l'écriture des fonctions et des algorithmes, nous adoptons les conventions de notation suivantes :

- la chaîne de caractères qui correspond à l'étiquette des termes ou à l'étiquettes des langues est notée entre guillemets, par exemple, "sorbet à la framboise", "en" ;
- l'objet qui correspond à un terme du RLSM_{PI} est noté sans guillemet et en italique, par exemple, *sorbet à la framboise* ;
- le type de l'objet est noté entre parenthèses et commence par une lettre majuscule, par exemple, (Terme) ;
- le contenu d'un objet (terme ou relation) est noté entre crochets, par exemple, [id :123,name :"sorbet"] ;
- le contenu d'une liste ou d'un ensemble est noté entre crochets, par exemple, {*sorbet*, *glace*, *parfait glacé*} ;
- les termes appartenant au pivot interlingue (PI) ont une étiquette qui commence par un préfixe "in :" afin de les différencier des termes qui appartiennent à une langue naturelle ;

— les relations sont notées $fruit \xrightarrow[w=200]{r\_hypo} pomme$ où *fruit* et *pomme* sont des termes, w est le poids de la relation, r_hypo est le type de la relation.

A.1.1 Fonctions de base

`RechercherTerme`(étiquette_terme, étiquette_langue) ⇒ *terme*. Fonction qui retourne le terme (nœud du $RLSM_{PI}$) qui a pour étiquette une chaîne de caractères et appartient à un graphe donné. La valeur de retour peut correspondre aussi bien à un mot d'une langue, à un objet lexical qui modélise une catégorie morpho-syntaxique (ex. "in :Noun, in :GenitiveCase), un groupe de mots ou une expression. Complexité *O(1)*.

Exemple A.1

`RechercherTerme` ("soupe","fr")
⇒ (Terme) *id* :"79311", *name* :"soupe", *origin* :"JDM", *label* :{$fr, miam$}

`RechercherRelation` (*type*, *étiquette_langue*, *annotation*,*poids minimum*)={r_1,r_2, ...r_n} Fonction qui retourne une liste de relations de type *type* au sein du graphe désigné par l'étiquette_langue. Chaque relation possède un identifiant unique, ***elle modélise un triplet*** pourvu de terme source, terme cible, type, poids, parfois annotation. Complexité *O(1)*.

Exemple A.2

`RechercherRelation`(r_matter,"fr","in :aroma",150)
⇒ (Relation) [

$id\_relation$: 806496, //id unique de la relation

$source$: 79311, //*soupe*

$cible$: 262917, //*umami*

$type$: "r_matter", //type

$annotation$: [

$reification$: $soupe[r\_matter]umami$,

$annotation$: 62647 // *in :aroma*

$prop$: ["$weight$" : "150",

$origin$: "$MIAM$"]

]

]

`Raffinements`(terme)=[raff$_1$,raff$_2$, ...raff$_n$]. Une fonction qui retourne la liste des raffinements d'un terme. Complexité *O(1)*.

Exemple A.3

`Raffinements`(*soupe*)⇒(Terme) {*soupe>potage*, *soupe>neige*, *soupe>repas*}

`TermesCouvrants`(terme)=[c$_1$,c$_2$, ...c$_n$]. Fonction qui retourne la liste de termes couvrants interlingues pour un terme d'une langue présente au sein d'un graphe de langue. Complexité *O(n)*.

Exemple A.4

`TermesCouvrants`(*soupe*) ⇒ (Terme){*in :broth*, *in :soupe*}

`TermesCouverts`(terme_interlingue,*langue*)=[tc$_1$,tc$_2$, ...tc$_n$]. Fonction qui retourne la liste de termes couverts interlingues pour un terme d'une langue présente au sein d'un graphe de langue. Le paramètre *langue* est optionnel. Complexité *O(1)*.

Exemple A.5

`TermesCouverts`(*in :soup*)
⇒ (Terme){*panade,caldo,sopa,potage,velouté,soupe,soup,похлёбка,суп*}

`RelationsSortantes`(terme,*type_relation*, *ensemble_de_relations*, Tab[])=[Tab[e$_1$], [Tab[e$_2$], ...[Tab[e$_n$]. Cette fonction renvoie en tableau les termes du tableau Tab[] qui modélise toutes les relations sortantes du *terme* qui vérifient la relation entrante dont le terme de départ est *terme* fourni en paramètre. Le paramètre *type_de_relation* est optionnel, il est possible de choisir un sous-ensemble de relations tel que relations **lexicales** (ex. *r_variant*, *r_syn*), **prédicatives** (relations prédicat argument ex. *r_instrument*, *r_object*), **sémantiques**(ex. *r_has_part*, *r_manner*, *r_carac*) en fournissant le paramètre optionnel *ensemble_de_relations*. Complexité *O(n)*.

Exemple A.6

`RelationsSortantes`(*soup*) // *sortie partielle*

(Relation) {

$soup \xrightarrow{r\_pos} in : Noun$,

$soup \xrightarrow{r\_isa} dish$,

$soup \xrightarrow{r\_location} bowl$,

$soupe \xrightarrow{r\_hypo} julienne$,

$soupe \xrightarrow{r\_hypo} vichyssoise$,

$soupe \xrightarrow{r\_carac} hot$,

$soupe \xrightarrow{r\_time} beginning\ of\ the\ meal$

}

`RelationsEntrantes`(Tab[],*type_ relation*,*ensemble_ de_ relations*,terme)=[Tab[e_1], [Tab[e_2], ...[Tab[e_n]. Cette fonction renvoie en tableau les termes du tableau Tab[] qui modélise toutes les relations sortantes du *terme* qui vérifient la relation entrante dont le terme de départ est *terme* fourni en paramètre. Le paramètre *type_ de_ relation* est optionnel, il est possible de choisir un sous-ensemble de relations tel que relations **lexicales** (ex. *r_ variant*, *r_ syn*), **prédicatives** (relations prédicat-argument ex. *r_ instrument*, *r_ object*), **sémantiques**(ex. *r_ has_ part*, *r_ manner*, *r_ carac*) en fournissant le paramètre optionnel *ensemble_ de_ relations*. Complexité *O(n)*.

Exemple A.7

`RelationsEntrantes`((Terme) soup) //*sortie partielle*

⇒(Relation) {

$peasoup \xrightarrow{r\_isa} soup$,

$broth \xrightarrow{r\_isa} soup$,

$borscht \xrightarrow{r\_isa} soup$,

$gazpacho \xrightarrow{r\_isa} soup$,

$reheat \xrightarrow{r\_object} soup$,

$pure \xrightarrow{r\_object} soup$,

$pour \xrightarrow{r\_object} soup$

}

`Similaire`(terme_1,terme_2, seuil, α,β). Une fonction qui retourne un réel qui correspond à la similarité entre le terme_1 et le terme_2 fournis en paramètre dans les limites d'un score minimum requis en fournissant le paramètre *seuil*. Le calcul de similarité est le calcul de l'indice de Jaccard (dont le cas général est l'indice de Tversky. Cet indice est calculé pour les ensembles des demi-relations sortantes et entrantes. Le choix des relations peut être réduit aux seules relations sémantiques, lexicales, grammaticales, rôles thématiques. α et β sont des paramètres optionnels qui permettent de varier l'importance accordée à l'intersection entre les ensembles de demi-relations par rapport à leur union. Pour le cas par défaut où $\alpha = \beta = 1$ le seuil σ est fixé empiriquement à 0,30 (le seuil dépend fortement de l'état du peuplement du réseau lexico-sémantique et dépend du degré moyen des termes du réseau).

$$J(terme\_1, terme\_2) = \frac{\alpha|R_{terme\_1} \cap R_{terme\_2}|}{\beta|R_{terme\_1} \cup R_{terme\_2}|} \tag{A.1}$$

Complexité $O(n^2)$.

Exemple A.8

`Similaire`("pain","viande") ⇒ ***0.10***

(intersection) (Terme){*produit*, *aliment périssable*, *blanc*,..., *bio*}

`Similaire`("glace","sorbet") ⇒ ***0.31***

(intersection) (Terme){*aliment*, *produit glacé*, *in :ice*, *sucré*, ...*entremets*}

`Polysémique`(terme). Fonction qui retourne un booléen qui indique si le terme fourni en paramètre est polysémique. La polysémie correspond à la présence des raffinements si le terme est raffiné ou à la présence de multiples termes interlingues couvrant le terme fourni en paramètre. Complexité *O(1)*.

Exemple A.9

`Polysémique`(*chocolate*) ⇒ ***vrai***

(présence des relations entrantes typées r_covers et relations sortantes

de type r_refinement)

(Relation){

$chocolate \xleftarrow{r\_covers} in :chocolate>aroma$

$chocolate \xleftarrow{r\_covers} in :chocolate>beverage$

$chocolate \xleftarrow{r\_covers} in :chocolate>color$

$chocolate \xleftarrow{r\_covers} in :chocolate>candy$

$chocolate \xrightarrow{r\_refinement} chocolate>adj$

$chocolate \xrightarrow{r\_refinement} chocolate>noun$

$chocolate \xrightarrow{r\_refinement} chocolate>food\ product$

}

`Raffiné`(terme). Fonction qui retourne un booléen qui indique si le terme fourni en paramètre a des raffinements.

Exemple A.10

`Raffiné`(*chocolate*) ⇒ ***vrai***

(relations sortantes typées r_refinement)

(Relation){

$chocolate \xrightarrow{r\_refinement} chocolate>aroma$

$chocolate \xrightarrow{r\_refinement} chocolate>adj$

$chocolate \xrightarrow{r\_refinement} chocolate>noun$

$chocolate \xrightarrow{r\_refinement} chocolate>food\ product$

$chocolate \xrightarrow{r\_refinement} chocolate>color$

}

Complexité *O(1)*.

`SemantiquementLié`(*terme_1*,*terme_2*). Fonction qui retourne un booléen dont la valeur correspond à la présence d'une relation (indépendante d'une langue donnée) entre terme_1 et terme_2 fournis en paramètre. Toutes les relations de chemin doivent être des relations sémantiques. La fonction vérifie également l'existence d'un lien sémantique cross-lingue entre les termes fournis en paramètre. A savoir, s'il existe ou non un lien sémantique entre les termes alignés avec terme_1 et terme_2 dans une autre langue du RLSM_{PI} ou au niveau du

pivot interlingue.

Exemple A.11

`SemantiquementLié`("pain","viande") ⇒ ***vrai***

(au niveau cross-lingue)

$$bread \xleftarrow{r\_object} slice \xrightarrow{r\_object} meat$$

(au niveau monolingue)

$$pain \xleftarrow{r\_matter} sandwich \xrightarrow{r\_has\_part} viande$$

$$pain \xleftarrow{r\_carac} escalope\ panée \xrightarrow{r\_matter} viande$$

$$pain \xrightarrow{r\_carac} cuit \xleftarrow{r\_carac} viande$$

Complexité *O(n)*.

`CréerTerme`("terme", "langue") ⇒ *terme*. Fonction qui créé et retourne un nouveau terme à lequel est associée l'étiquette de "langue" et l'étiquette fournie en paramètre. La fonction vérifie la présence d'un terme possédant l'étiquette et l'étiquette langue fournies en paramètre. Si le terme n'existe pas, il est créé en temps constant (sinon, il est retourné). Complexité *O(1)*.

```
1 Fonction CréerTerme(nom_terme, langue) :
2     // initialisation
3     Terme ← null
4     if ∄ Terme (nom_terme, langue) then
5         Terme.id ← ObtenirDernierId + 1 ;
6         Terme.name ← nom_terme ;
7         Terme.label ← langue ;
8     retourner Terme ;
```

`CréerRelation`("nom_source","nom_cible",*type*,"nom_annotation","langue") ⇒ (Relation) {r_1,r_2, ...r_n}. Fonction qui renvoie un ensemble de relations qui peut contenir une seule relation binaire (si le paramètre annotation n'est pas fourni) ou contenir plusieurs relations et impliquer la création d'un terme spécifique correspondant à la réification de la relation en train d'être créée afin qu'elle puisse être annotée (si la relation à créer est une relation annotée).

Dans un premier temps, la fonction `CréerRelation` identifie à partir des étiquettes de termes fournies en paramètre et de l'étiquette de langue les termes *source*, *cible* et *annotation* s'ils existent dans $RLSM_{PI}$ et si la relation typée *type* existe entre *source* et *cible*. Le cas échéant, ces termes sont créés. Si le paramètre "étiquette_annotation" est fourni, une relation annotée est créée. La création

d'une relation annotée débute par la création de la relation typée *type* entre *source* et *cible*.

Puis un terme spécifique (*réification*) qui réifie cette relation est créé et relié à *source*, *cible* et *annotation* par des relations spécifiques typées respectivement r_source, r_cible et r_annotation.

La fonction `CréerRelation` implique l'incrémentation de du dernier identifiant attribué car une relation possède un identifiant unique.

La fonction fait appel à des fonctions de plus bas niveau telles que `CréerTerme`, `RechercherTerme` et `RechercherRelation` qui fournissent le résultat en temps constant. Par conséquent, la fonction `CréerRelation` s'exécute également en un temps constant. Complexité *O(1)*.

---

1 **Fonction** `CréerRelation`(*nom_source, nom_cible, type,nom_annotation, langue*) **:**
2 `// initialisation`
3 Terme source $\leftarrow \varnothing$;
4 Terme cible $\leftarrow \varnothing$;
5 Terme annotation $\leftarrow \varnothing$;
6 Terme réification $\leftarrow \varnothing$;
7 Terme source $\leftarrow$ `CréerTerme` (nom_source,langue) ;
8 Terme cible $\leftarrow$ `CréerTerme` (nom_cible,langue) ;
9 Terme annotation $\leftarrow$ `CréerTerme` (nom_annotation,langue) ;
10 Terme réification $\leftarrow$ `CréerTerme` (nom_source[type]nom_cible,langue) ;
11 Relation $rtype \leftarrow \varnothing$;
12 Relation $rannot \leftarrow \varnothing$;
13 Relation $rsource \leftarrow \varnothing$;
14 Relation $rcible \leftarrow \varnothing$;
15 **if** $\nexists$ *(Relation)source* $\xrightarrow{type}$ *cible* **then**
16 $rtype \leftarrow \{$`ObtenirDernierId` $+ 1$, *source* $\xrightarrow{type}$ *cible*$\}$;
17 **if** $\nexists$ *(Relation) réification* $\xrightarrow{r\_annotation}$ *annotation* $\wedge$ *annotation* $\neq \varnothing$ **then**
18 lastID $\leftarrow$ `ObtenirDernierId` $+1$;
19 $rannot \leftarrow \{$ lastID,*réification* $\xrightarrow{r\_annotation}$ *annotation*$\}$;
20 lastID$++$;
21 $rsource \leftarrow \{$ lastID,*réification* $\xrightarrow{r\_source}$ *source*$\}$;
22 lastID$++$;
23 $rannot \leftarrow \{$ lastID,*réification* $\xrightarrow{r\_cible}$ *cible*$\}$;

`Négativer`(Relation r). Fonction qui modifie le poids d'une relation fournie en paramètre pour le rendre négatif (cas d'une relation fausse). Complexité O(1).

Exemple A.12

`Négativer`((Relation) *casser* $\xrightarrow[w=150]{r\_object}$ *framboise*)
⇒(Relation)*casser* $\xrightarrow[w=-150]{r\_object}$ *framboise*

`ModifierPoids`(Relation r, opérateur, valeur). Fonction qui modifie le poids d'une relation fournie en paramètre. Complexité O(1).

Exemple A.13

`ModifierPoids`((Relation) *pain* $\xrightarrow[w=150]{r\_isa}$ *boulot >moyen de subsistance*, /, 3)
⇒ (Relation) *pain* $\xrightarrow[w=50]{r\_isa}$ *boulot >moyen de subsistance*

A.2 Algorithmes

A.2.1 Inférence des raffinements glosés

Le processus d'inférence des raffinements glosés est décrit dans le chapitre 2 du présent manuscrit. Il exploite deux leviers : la présence éventuelle des raffinements dans une des langues du RLSM_{PI} et les informations présentes au sein du pivot interlingue notamment la présence des termes couvrants multiples pour un seul et même terme en langue source. Dans le premier cas, le processus d'inférence correspond au choix de la glose appropriée pour le raffinement dans la langue source à partir des raffinements des termes alignés avec le termes fourni en paramètre (`Correspondance`) présents dans la langue cible (`Raffinements`(*terme_lié*)). Si la glose est validée le *terme raffiné* et la relation du *terme* à *terme raffine* typée r_refinement est créée dans la langue source. Dans le second cas, le processus d'inférence est l'identification de la glose appropriée pour chaque sens supposé en admettant, dans un premier temps, que chaque terme couvrant correspond à un sens différent (`SansGlose`), puis en regroupant les gloses potentielles redondantes (`Fusionner`). Ce processus peut être ascendant ou descendant.

Exemple A.14

`InférenceCrosslingueDesRaffinements`(*adoucir*)

`Raffinements`(`Correspondance`(adoucir, "en") ⇒ ∅
`SansGlose` ⇒ {

adoucir>in :alleviate //"soulager,réduire,calmer"

adoucir>in :dull //"assourdir"

adoucir>in :soften //"ramollir"

adoucir>in :sweeten //"sucrer"

}

`Fusionner` ⇒ {

adoucir>[alleviate,dull] //"réduire, "

adoucir>in :soften //"mollir,ramollir"

adoucir>in :sweeten //"sucrer,édulcorer"

}

`CréerTerme` ⇒ {*adoucir>réduire*, *adoucir>ramollir*, *adoucir>sucrer*}
`Inférer` ⇒ {

$adoucir \xrightarrow{r\_refinement} adoucir{>}réduire$

$adoucir \xrightarrow{r\_refinement} adoucir{>}ramollir$

$adoucir \xrightarrow{r\_refinement} adoucir{>}sucrer$

}

Algorithme 7 : Inférence des raffinements glosés

input : RLSM_{PI}, terme, langue_source, langue_cible
output : R_{raff_cible}

```
1  // initialisation
2  // glose qui sert à définir le sens dans la langue_cible
3  glose_candidate=∅ ;
4  // relation candidate de raffinement
5  relation_candidate=∅;
6  // ensemble de raffinements candidats pour langue_cible
7  R_raff_cible ← ∅
8  if Polysémique(terme) ∧ !Raffiné(terme) then
9      foreach terme_lié ∈ Correspondance (terme, langue_cible) do
10         // inférence des raffinements à partir d'un raffinement
              existant dans une autre langue ou au niveau du pivot
              interlingue
11         if Raffiné(terme_lié) then for glose ∈ Glose (Raffinements
            (terme_lié)) do
12             if SemantiquementLié (terme,Correspondance (glose,
                langue_source) then
13                 glose_candidate ← Correspondance (glose, langue_source);
14                 raffinement ← CréeTerme (Nom (terme, glose_candidate),
                    langue_cible));
15                 relation_candidate ← Inférer (terme --raff--> raffinement);
16                 R_raff_cible.first()← relation_candidate ;
17         ;
18         // inférence des raffinements à partir des nœuds couvrants
              existant au niveau du pivot interlingue
19         else
20             S ← ∅ ;
21             foreach n ∈ TermesCouvrants (terme) do
22                 S ← SansGlose (terme, n) ; Fusionner (S) ; foreach u∈S do
23                     raffinement ← CréeTerme (Nom (terme, u), langue_cible));
24                     relation_candidate ← Inférer (terme --raff--> raffinement);
25                     R_raff_cible ← relation_candidate ;
26 retourner R_raff_cible
```

A.2.2 Découverte des éléments remarquables de type "classe d'ontologie"

Détail sur le déroulement de l'algorithme d'inférence des classes d'ontologie.

L'algorithme d'inférence des classes d'ontologie se déroule en trois temps.

Premièrement, il recherche et valide toutes les chaînes hiérarchiques présentes au sein du $RLSM_{PI}$ pour un terme donné quelle que soit la place de ce terme dans chacune des chaînes. La chaîne est ensuite totalement ou partiellement validée selon le mécanisme décrit dans le chapitre 3.

Deuxièmement, l'algorithme produit une liste de triplets candidats. Ce triplets peuvent être suggérés aux experts du domaine pour une validation manuelle, validés automatiquement servir pour l'inférence cross-lingue des classes d'ontologie.

Troisièmement, l'algorithme valide les triplets candidats par abduction dans la langue source.

Exemple A.15

`InférenceClasses`(*pain*)

`ObtenirChaîneHiérarchiques`(*pain*)

(exemples de chaînes après filtrage sémantique) *pain de campagne–>pain–>ingrédient de cuisine–>aliment baguette complète–>pain complet–>pain–>ingrédient de recette de cuisine–>aliment*

africain–>pain suisse–>pain–>aliment

Algorithme 8 : Découverte des éléments remarquables de type « classe d'ontologie »

input : terme, langue_source, langue_cible
output : Terme[]

```
1  // l'algorithme permet de proposer à la fois des éléments
      candidats en langue cible (langue de l'ontologie de départ) et
      dans d'autres langues, langue_source == ou != langue_cible
2  // initialisation
3  Terme[] ← ∅ ; // liste de termes (éléments candidats) pour la
      langue_cible
4  Candidats[] ← ∅ ;// liste de triplets candidats
5  foreach c ∈ ObtenirChaînesHiérarchiques (terme) do
6  |  ValiderChaîne (c) ;
7  |  Candidats[] ← ObtenirCandidats () ;
8  // utilisation du processus par règles
9  foreach candidat ∈ Candidats[] do
10 |  // validation par abduction (en exploitant les exemples
   |     proches) dans la langue source, s'il existe suffisamment
   |     d'exemples
11 |  if Match (candidat) then
12 |  |  // recherche des structures proches dans la langue cible
```

```
1  Fonction ValiderChaine(Chain[]) :
2      // initialisation
3      LocalChain[]← ∅ ;
4      i=0 ;
5      l_node m=∅ ;
6      l_node n=∅ ;
7      l_node o=∅ ;
8      if Longueur (c) ≥ 3 then
9          while i≤Longueur (c)-3 do
10             m←c(i) ;
11             n←c(i+1) ;
12             o←c(i+2) ;
13             if lié (m,n)&lié (n,o) then
14                 LocalChain[].first() ← m ;
15                 LocalChain[].first() ← n ;
16                 LocalChain[].first() ← o ;
17     retourner LocalChain[] ;
```

Détail des fonctions utilisées dans la cadre de l'algorithme de découverte des éléments remarquables de type "classe d'ontologie".

```
1  Fonction Candidats(Chain[]) :
2      // initialisation
3      Candidats[] ← ∅ ;
4      triplet← ∅ ;
5      foreach chaîne ∈ LocalChain[] do
6          j←Longueur (chaîne)-1 ; while j≠0 do
7              if chaîne(j) ∈ MIAM then
8                  if chaîne(j-1) notin MIAM then
9                      triplet ← chaîne(j-1) subClassOf chaîne(j) ;
10                     Candidats[] ← triplet ;
11             if chaîne(j) ∉ MIAM then
12                 if chaîne(j-1) in MIAM then
13                     triplet ← chaîne(j-1) subClassOf chaîne(j) ;
14                     Candidats[].first() ← (triplet) ;
15             j - - ;
16     retourner Candidats[] ;
```

Algorithme 9 : Inférence cross-lingue des relations sémantiques

input : $\mathrm{RLSM_{PI}}$, *type*, langue_source, langue_cible
output : R_{type}

```
1  // initialisation
2  relation_candidate=∅;
3  // ensemble de raffinements candidats pour langue_cible
4  R_type ← ∅
5  foreach terme ∈ Termes (RLSM_PI, langue_source) do
6      if RelationsSortantes (terme)> 25 then
7          foreach terme_lié ∈ Correspondance (terme, langue_cible) do
8              if Raffiné(terme_lié) then for glose ∈ Glose (Raffinements
                (terme_lié)) do
9                  if SemantiquementLié (terme,Correspondance (glose,
                    langue_source) then
10                     glose_candidate ← Correspondance (glose, langue_source);
11                     raffinement ← CréerNoeud (Nom (terme, glose_candidate),
                        langue_cible));
12                     relation_candidate ← Inférer (terme --raff--> raffinement);
13                     R_raff_cible ← relation_candidate ;
14             ;
15             // inférence des raffinements à partir des nœuds
                   couvrants existant au niveau du pivot interlingue
16             else
17                 S ← ∅ ;
18                 foreach n ∈ NoeudsCouvrants (terme) do
19                     S ← SansGlose (terme, n) ; Fusionner (S) ; foreach u∈S do
20                         raffinement ← CréerNoeud (Nom (terme, u),
                            langue_cible));
21                         relation_candidate ← Inférer (terme --raff--> raffinement);
22                         R_raff_cible ← relation_candidate ;
23 retourner R_raff_cible
```

Annexe B

Glossaire

B.1 Définitions

acception : sens particulier d'un mot, admis et reconnu par l'usage. Il s'agit d'une unité sémantique propre à une langue donnée. Axie est une acception interlingue.

acquisition lexicale : acquisition du vocabulaire nécessaire à la description des phénomènes réels ;

architecture : *(Ressource)* Organisation des éléments constitutifs d'une ressource de connaissance en vue d'optimiser l'ensemble de sa conception.

bruit : *(RLS)* présence des relations non pertinentes.

cadre : structure cohérente de concepts reliés d'une manière qui impose de les connaître tous pour en comprendre un. (Allan [2001])

classe : ensemble d'individus ayant les mêmes caractéristiques.

concept : *(Terminologie*[1]*)* Unité de connaissance créée par une combinaison unique de caractères.

donnée : *(Terminologie)* informations représentées sous une forme conventionnelle qui conveient à la communication, à l'interprétation, au stockage et au traitement ;

environnement : *(SME)* Structure de données perçue et pouvant être modifiée par un agent.

expression polylexicale : combinaison de mots pour laquelle les propriétés syntaxiques et sémantiques de l'expression entière ne peuvent pas être obtenus à partir de l'analyse des ses parties. Syn. : terme multi-mot.

folksonomie : ressource informelle rattachée à une ontologie (ressource formelle) afin de fournir un "vivier" de termes et de structures sémantiques issus de l'acquisition des connaissances et non intégrables dans l'ontologie (pour des raisons de redondance, par exemple).

forme : aspect sous lequel une vocable apparaît dans un énoncé.

1. Standard ISO 1087-1.

glose : 1.*(RLS)* Terme qui sert à "nommer" un sens d'usage. 2a. *(Lexicographie)* Une définition. 2b. *(Lexicographie)* Un exemple d'utilisation d'une vocable donnée dans un texte.

inférence : opération qui consiste à admettre une proposition en raison de son lien avec une proposition préalable tenue pour vraie [2].

interlingua : *(Traduction automatique)* une langue artificielle intermédiaire entre une langue naturelle source et une langue naturelle cible.

mot : 1.*(énoncé)* forme. 2.*(dictionnaire)* vocable.

ontologie : spécification formelle d'une conceptualisation partagée.

propriété : relation d'ontologie. Cette relation peut exister entre une source (appelé *domaine*) et une cible (*co-domaine*). Si le domaine et le co-domaine de la propriété sont des classes (concepts), il s'agit d'une *Object Property* (propriété à valeur objet). Si lre domaine est un individu tandis que le co-domaine est un littéral, il s'agit de *Data Property* (propriété à valeur donnée).

raffinement : sens d'usage d'un terme donné.

relation : 1.*(RLS)* Relation sémantique faisant partie du RLSM_{PI}, par exemple, r_has_part. 2. *(Ontologie)* Relation hiérarchique ou thématique d'ordre ontologique.

ressource : Ensemble de moyens nécessaires pour accomplir un ensemble de tâches.

ressource de connaissance : Ressource qui regroupe et/ou structure des connaissances.

ressource langagière : Les **ressources langagières** sont des ressources de connaissance qui modélisent la connaissance sur le monde telle qu'elle est exprimée à travers le langage humain. A ce titre, elles incluent à la fois la connaissance d'ordre ontologique et sémantique et la connaissance d'ordre purement lexical et spécifique à une langue donnée. Toute ressource de connaissance n'est pas forcément une ressource langagière.

rôle : 1. *(SME)* Fonction qu'à l'agent dans le système. Le type d'interaction avec l'environnement dépend de cette fonction. Type d'interaction avec l'environnement. 2. *(Ontologie)* Ensemble de couples d'individus d'ontologie. *Dans le présent chapitre la définition (1) sera utilisée.*

silence : (*(RLS)* absence des relations pertinentes dans une réseau lexico-sémantique.

synset : *(Lexicographie)* Ensemble de synonymes ou quasi-synonymes.

système multi-agent (SMA) : système basé sur plusieurs programmes autonomes (appelés "agents") qui interagissent avec leur environnement.

terme : *((Terminologie [3])* Désignation verbale d'un concept général dans un domaine spécifique.*(RLS)* Item lexical(vocable, expression polylexicale).

vocable (mot) : *(Linguistique)* unité significative indépendante, ne pouvant pas toujours être déterminée selon un critère de séparabilité fonctionnelle ni par un critère de délimitation intonative [4] ;

2. `http://www.cnrtl.fr/definition/inf\%C3\%A9rence`
3. Standard ISO 1087-2.
4. `http://www.cnrlt.fr/definition/mot`

B.2 Synthèse des schémas d'inférence utilisés et envisageables

B.2.1 Déduction

La déduction est un schéma descendant qui se base sur la transitivité des relations d'hyperonymie (r_isa) et hyponymie(r_hypo) d'un terme donné. Le schéma peut résumé comme suit :

$$\exists A \xrightarrow{r\_isa} B \wedge \exists B \xrightarrow{R} C \Rightarrow A \xrightarrow{R} C$$

Le blocage logique pour ce type d'inférence concerne les différents sens (raffinements) qui peuvent exister pour le terme B. Si deux raffinements de B sont connectés respectivement à) une des deux prémisses (A et C), l'inférence serait fausse.

B.2.2 Induction

L'inférence par induction propose des relations d'un hyponyme à ses hyperonymes.

$$\exists A \xrightarrow{r\_isa} B \wedge \exists A \xrightarrow{R} C \Rightarrow B \xrightarrow{R} C$$

Le filtrage logique est également lié à la polysémie éventuelle de A.

B.2.3 Abduction

Le schéma par abduction est basé sur les exemples. En présence de deux termes similaires, les relations d'un terme peuvent être proposées à l'autre terme. Dans le cadre de ce schéma, il s'agit de sélectionner un ensemble de termes similaires et à une terme T et de proposer les relations détenues par les termes similaires à T.

Ce schéma s'applique aux termes ayant au minimum deux raffinements et consiste à proposer des relations sémantiques détenues par un raffinement aux hypéronymes, synonymes et hyponymes du terme raffiné.

B.2.4 Inférence par raffinement et inférence interlingue

L'ensemble des sens d'un terme donné (l'ensemble d'usages) peut être structuré sous forme d'un arbre.

Dans le contexte d'une architecture de la ressource multilingue avec pivot, l'inférence ascendante (langue→ pivot) et descendente(pivot→ langue) est très pertinente étant donné qu'un terme peut avoir plusieurs termes couvrants et un terme du pivot peut couvrir plusieurs termes lexicalisés.

Bibliographie

Yong-Yeol Ahn, Sebastian E. Ahnert, James P. Bagrow, and Albert-László Barabási. Flavor network and the principles of food pairing. *CoRR*, abs/1111.6074, 2011. URL `http://dblp.uni-trier.de/db/journals/corr/corr1111.html#abs-1111-6074`.

Emin Akkoç and Nihan Kesim Cicekli. *Semanticook : A Web Application for Nutrition Consultancy for Diabetics*, pages 215–224. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-24731-6. doi : 10.1007/978-3-642-24731-6_23. URL `http://dx.doi.org/10.1007/978-3-642-24731-6_23`.

Keith Allan. *Natural Language Semantics*. Blackwell, 2001.

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 344–354. Association for Computational Linguistics, 2015. doi : 10.3115/v1/P15-1034. URL `http://www.aclweb.org/anthology/P15-1034`.

ATILF. Réseau lexical du français (rl-fr), 2017. URL `https://hdl.handle.net/11403/lexical-system-fr/v1`. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Sophie Aubin and Thierry Hamon. YaTeA - version 2006. 2006. URL `https://hal.archives-ouvertes.fr/hal-00090750`. YaTeA (Yet Another Term ExtrActor) aims at identifying and extracting noun phrases which are potential terms (*i.e.* term candidates). Each term candidate is syntactically analysed in order to identify head and modifier components. YaTeA can integra.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL '98, August 10-14, 1998, Université de Montréal, Montréal, Quebec, Canada. Proceedings of the Conference.*, pages 86–90, 1998. URL `http://aclweb.org/anthology/P/P98/P98-1013.pdf`.

Charles Bally. *Linguistique générale et Linguistique Fran caise*. 1965.

Christian Bizer, Georgi Kobilarov, Jens Lehmann, and Zachary Ives. Dbpedia : A nucleus for a web of open data. In *Proc. 6th Int'l Semantic Web Conf.* Springer, 2007.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase : a collaboratively created graph database for structuring human knowledge. In *In SIGMOD Conference*, pages 1247–1250, 2008.

Francis Bond and Ryan Foster. Linking and extending an open multilingual wordnet. Sofia, 2013.

Dhouha Bouamor. *Constitution de ressources linguistiques multilingues à partir de corpus de textes parallèles et comparables. (Using parallel and comparable corpora for multilingual linguistic resources extraction)*. PhD thesis, University of Paris-Sud, Orsay, France, 2014. URL `https://tel.archives-ouvertes.fr/tel-00994222`.

Didier Bourigault. Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. pages 24–27, 01 2002.

Lynne Bowker and Jennifer Pearson. *Working with Specialized Language : A Practical Guide to Using Corpora*. Number London. Routledge, 2002.

François Brown De Colstoun, Estelle Delpech, and Etienne Monneret. Libellex : une plateforme multiservices pour la gestion des contenus multilingues. In Prince V. Lafourcade, M., editor, *TALN'2011*, volume 2, page 319, Montpellier, France, 2011. URL `https://hal.archives-ouvertes.fr/hal-00912322`.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*, 2010. URL `http://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/view/1879`.

J. Charlet, B. Bachimont, and M.-C. Jaulent. Building medical ontologies by terminology extraction from texts : An experiment for the intensive care units. *Computer in Biology and Medicine*, 36(7-8) :857–870, 2006.

Amélie Cordier, Valmi Dufour-Lussier, Jean Lieber, Emmanuel Nauer, Fadi Badra, Julien Cojan, Emmanuelle Gaillard, Laura Infante-Blanco, Pascal Molli, Amedeo Napoli, and Hala Skaf-Molli. *Taaable : A Case-Based System for Personalized Cooking*, pages 121–162. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. ISBN 978-3-642-38736-4. doi : 10.1007/978-3-642-38736-4_7. URL `https://doi.org/10.1007/978-3-642-38736-4_7`.

Brian A. Davey and Hilary A. Priestley. *Introduction to lattices and order*. Cambridge University Press, Cambridge, 1990. ISBN 0521365848 9780521365840 0521367662 9780521367660. URL `http://www.worldcat.org/search?qt=worldcat_org_all&q=0521367662`.

F. de Saussure and R. Engler. *Cours de linguistique générale*. Number vol. 2 in Cours de linguistique générale. Harrassowitz, 1990. ISBN 9783447015271. URL `https://books.google.fr/books?id=nVznoAEACAAJ`.

Hervé Déjean, Éric Gaussier, and Fatia Sadat. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi : 10.3115/1072228.1072394. URL `http://dx.doi.org/10.3115/1072228.1072394`.

Sylvie Després. Construction d'une ontologie modulaire pour l'univers de la cuisine numérique. In *IC 2014 : 25es Journées francophones d'Ingénierie des Connaissances (Proceedings of the 25th French Knowledge Engineering Conference), Clermont Ferrand, France, May 12-16, 2014.*, pages 27–38, 2014. URL `http://hal.archives-ouvertes.fr/IC\_2014/hal-01010222`.

Sylvie Després. Construction d'une ontologie modulaire. application au domaine de la cuisine numérique. *Revue d'Intelligence Artificielle*, 30(5) :509–532, 2016. doi : 10.3166/ria.30.509-532. URL `https://doi.org/10.3166/ria.30.509-532`.

G. Dikonov. Development of lexical basis for the universal dictionary of unl concepts. 2013.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault : A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 601–610, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2956-9. doi : 10.1145/2623330.2623623. URL `http://doi.acm.org/10.1145/2623330.2623623`.

Zhendong Dong, Qiang Dong, and Changling Hao. Hownet and its computation of meaning. In *Proceedings of the 23rd International Conference on Computational Linguistics : Demonstrations*, COLING '10, pages 53–56, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1944284.1944298`.

Valmi Dufour-Lussier, Florence Le Ber, Jean Lieber, Thomas Meilender, and Emmanuel Nauer. Semi-automatic annotation process for procedural texts : An application on cooking recipes. *CoRR*, abs/1209.5663, 2012. URL `http://arxiv.org/abs/1209.5663`.

Valmi Dufour-Lussier, Florence Le Ber, Jean Lieber, and Emmanuel Nauer. Automatic case acquisition from texts for process-oriented case-based reasoning. *Inf. Syst.*, 40 :153–167, 2014. doi : 10.1016/j.is.2012.11.014. URL `https://doi.org/10.1016/j.is.2012.11.014`.

Roberto Navigli et al. Babelnet : The automatic construction, evaluation and application of a . . ., 2012.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised named-entity extraction from the web : An experimental study. *Artificial Intelligence*, 165(1) :91 – 134, 2005. ISSN 0004-3702. doi : http://dx.doi.org/10.1016/j.artint.2005.03.001. URL `http://www.sciencedirect.com/science/article/pii/S0004370205000366`.

Christiane Fellbaum. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, 1998. ISBN 978-0-262-06197-1. URL `http://mitpress.mit.edu/catalog/item/default.asp?ttype=2{\&}tid=8106`.

J. Ferber. *Les systèmes multi-agents : vers une intelligence collective*. InterEditions, Paris, 1995.

Pascale Fung and Lo Yuen Yee. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*, COLING '98, pages 414–420, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics. doi : 10.3115/980451.980916. URL `http://dx.doi.org/10.3115/980451.980916`.

E. Gaillard, J. Lieber, and E. Nauer. Improving ingredient substitution using formal concept analysis and adaptation of ingredient quantities with mixed linear optimization. In *Computer Cooking Contest Workshop*, Frankfort, Germany, 2015. URL `https://hal.inria.fr/hal-01240383`.

Alfio Gliozzo. Semantic domains and linguistic theory. In *In Proceedings of the LREC 2006 workshop*, 2006.

Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, 43(5-6) :907–928, December 1995. ISSN 1071-5819. doi : 10.1006/ijhc.1995.1081. URL `https://doi.org/10.1006/ijhc.1995.1081`.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. UBY - A large-scale unified lexical-semantic resource based on LMF. In *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, 2012*, pages 580–590, 2012. URL `http://aclweb.org/anthology/E/E12/E12-1059.pdf`.

Zellig Harris. Distributional structure. *Word*, 10(23) :146–162, 1954.

Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics. doi : 10.3115/992133.992154. URL `https://doi.org/10.3115/992133.992154`.

Hirst. Review of "eurowordnet : A multilingual database with lexical semantic networks" by piek vossen. kluwer academic publishers 1998. *Comput. Linguist.*, 25(4) :628–630, December 1999. ISSN 0891-2017. URL `http://dl.acm.org/citation.cfm?id=973226.973236`. Reviewer-Hirst, Graeme.

Jermsak Jermsurawong and Nizar Habash. Predicting the structure of cooking recipes. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *EMNLP*, pages 781–786. The Association for Computational Linguistics, 2015. ISBN 978-1-941643-32-7. URL `http://dblp.uni-trier.de/db/conf/emnlp/emnlp2015.html#JermsurawongH15`.

S. Johansson. *Seeing Through Multilingual Corpora : On the Use of Corpora in Contrastive Studies*. Studies in corpus linguistics. J. Benjamins, 2007. ISBN 9789027223005. URL `https://books.google.fr/books?id=xnvSJSIPLLEC`.

A. Jolkovsky and Igor Mel'čuk. Essai d'une theorie semantique applicable au traitement de langage. In *Second Conference Internationale Sur Le Traitement Automatique Des Langues, COLING 1967, Grenoble, France, August 1967*, 1967. URL `http://aclweb.org/anthology/C/C67/C67-1028.pdf`.

Anne-Laure Jousse. Extension de l'encodage formel des fonctions lexicales dans le cadre de la lexicologie explicative et combinatoire. In *Actes des 9e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 469–478, Toulouse, France, June 2007. Association pour le Traitement Automatique des Langues. URL `http://www.atala.org/taln_archives/RECITAL/RECITAL-2007/recital-2007-long-005`.

David Kamholz, Jonathan Pool, and Susan M. Colowick. Panlex : Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 3145–3150, 2014. URL `http://www.lrec-conf.org/proceedings/lrec2014/summaries/1029.html`.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. Class-based construction of a verb lexicon. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 691–696. AAAI Press, 2000. ISBN 0-262-51112-6. URL `http://dl.acm.org/citation.cfm?id=647288.721573`.

Maxim Kolchin, Alexander Chistyakov, Maxim Lapaev, and Rezeda Khaydarova. Foodpedia : Russian food products as a linked data dataset. In

The Semantic Web : ESWC 2015 Satellite Events - ESWC 2015 Satellite Events Portorož, Slovenia, May 31 - June 4, 2015, Revised Selected Papers, pages 87–90, 2015. doi : 10.1007/978-3-319-25639-9_17. URL https://doi.org/10.1007/978-3-319-25639-9_17.

Mathieu Lafourcade. Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07 : 7th International Symposium on Natural Language Processing*, page 7, Pattaya, Chonburi, Thailand, December 2007. URL https://hal-lirmm.ccsd.cnrs.fr/lirmm-00200883.

Mathieu Lafourcade. *Lexique et analyse sémantique de textes - structures, acquisitions, calculs, et jeux de mots. (Lexicon and semantic analysis of texts - structures, acquisition, computation and games with words)*. 2011. URL https://tel.archives-ouvertes.fr/tel-00649851.

Mathieu Lafourcade and Alain Joubert. Similarity between term senses in a lexical network. *TAL*, 50(1) :177–200, 2009. URL http://www.atala.org/IMG/pdf/TAL-2009-50-1-07-Lafourcade.pdf.

Ora Lassila and Deborah McGuinness. The role of frame-based representation on the semantic web. Technical report, Knowledge Systems Laboratory Report KSL-01-02, Stanford University, Stanford (USA), 2001.

Michael Lesk. Automatic sense disambiguation using machine readable dictionaries : How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, pages 24–26, New York, NY, USA, 1986. ACM. ISBN 0-89791-224-1. doi : 10.1145/318723.318728. URL http://doi.acm.org/10.1145/318723.318728.

Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. BIOTEX : A system for Biomedical Terminology Extraction, Ranking, and Validation. In *ISWC : International Semantic Web Conference*, volume CEUR-WS.org of *Posters & Demonstrations*, pages 157–160, Riva del Garda, Italy, October 2014. URL https://hal-lirmm.ccsd.cnrs.fr/lirmm-01112894.

Lashevich G. Gerasimova A. A. Ivanov V. V. Dobrov B. V. Loukachevitch, N. Creating russian wordnet by conversion. In *Dialog-2016*, pages 405–415, Moscow, 2016.

Cerstin Mahlow and Michael Piotrowski. SMM : Detailed, structured morphological analysis for Spanish. *Polibits*, (39) :41–48, June 2009. URL http://www.gelbukh.com/polibits/2009_39/39_06.pdf.

Jonathan Malmaud, Earl Wagner, Nancy Chang, and Kevin Murphy. Cooking with semantics. In *Proceedings of the ACL 2014 Workshop on Semantic Parsing*, pages 33–38, Baltimore, MD, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W14/W14-2407.

Mausam Mausam, Stephen Soderland, Oren Etzioni, Daniel Weld, Michael Skinner, and Jeff Bilmes. Compiling a massive, multilingual dictionary via probabilistic inference., 01 2009.

John P. McCrae, Francis Bond, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Jorge Gracia, Ilan Kernerman, Elena Montiel-Ponsoda, Noam Ordan, and Maciej Piasecki, editors. *Proceedings of the LDK 2017 Workshops : 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge (LDK 2017), Galway, Ireland, June 18, 2017*, volume 1899 of *CEUR Workshop Proceedings*, 2017. CEUR-WS.org. URL `http://ceur-ws.org/Vol-1899`.

T. Mitchell, W. Cohen, E. Hruscha, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohammad, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. In *AAAI*, 2015. URL `http://www.cs.cmu.edu/~wcohen/pubs.html`. : Never-Ending Learning in AAAI-2015.

Thibault Mondary. *Construction d'ontologies à partir de textes. L'apport de l'analyse de concepts formels.* Theses, Université Paris-Nord - Paris XIII, May 2011. URL `https://tel.archives-ouvertes.fr/tel-00596825`. Equipe RCLN.

Shinsuke Mori, Tetsuro Sasada, Yoko Yamakata, and Koichiro Yoshino. A machine learning approach to recipe text processing. 2012.

Gilbert Müller and Ralph Bergmann. Cookingcake : A framework for the adaptation of cooking recipes represented as workflows. In *Workshop Proceedings from The Twenty-Third International Conference on Case-Based Reasoning (ICCBR 2015), Frankfurt, Germany, September 28-30, 2015.*, pages 221–232, 2015. URL `http://ceur-ws.org/Vol-1520/paper23.pdf`.

Silvio Peroni, Giorgia Lodi, Luigi Asprino, Aldo Gangemi, and Valentina Presutti. Food : Food in open data. In Paul Groth, Elena Simperl, Alasdair Gray, Marta Sabou, Markus Krötzsch, Freddy Lecue, Fabian Flöck, and Yolanda Gil, editors, *The Semantic Web – ISWC 2016*, pages 168–176, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46547-0.

B. Pinter, I. Vassányi, B. Gaál, E. Mák, and Gy. Kozmann. *Personalized Nutrition Counseling Expert System*, pages 957–960. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-23508-5. doi : 10.1007/978-3-642-23508-5_248. URL `http://dx.doi.org/10.1007/978-3-642-23508-5_248`.

Alain Polguère. Lexical systems : graph models of natural language lexicons. *Language Resources and Evaluation*, 43(1) :41–55, 2009. doi : 10.1007/s10579-008-9078-4. URL `https://doi.org/10.1007/s10579-008-9078-4`.

Alain Polguère. Modélisation des liens lexicaux au moyen des fonctions lexicales. page 24, 06 2002.

Soujanya Poria, Basant Agarwal, Alexander Gelbukh, Amir Hussain, and Newton Howard. Dependency-based semantic parsing for concept-level text analysis. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 113–127, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg. ISBN 978-3-642-54906-9.

Lionel Ramadier. *Automatic extraction of semantic information in the radiologic reports for search in of medical imaging.* Theses, Université Montpellier, November 2016. URL `https://hal-lirmm.ccsd.cnrs.fr/tel-01479769`.

Thomas Rebele, Fabian M. Suchanek, Johannes Hoffart, Joanna Biega, Erdal Kuzey, and Gerhard Weikum. YAGO : A multilingual knowledge base from wikipedia, wordnet, and geonames. In *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part II*, pages 177–185, 2016. doi : 10.1007/978-3-319-46547-0_19. URL `https://doi.org/10.1007/978-3-319-46547-0_19`.

Elaine Rich and Kevin Knight. *Artificial Intelligence.* McGraw-Hill Higher Education, 2nd edition, 1990. ISBN 0070522634.

Christophe Roche. Ontology : A survey. *IFAC Proceedings Volumes*, 36, 09 2003. doi : 10.1016/S1474-6670(17)37715-7.

Christophe Roche. Le terme et le concept : fondements d'une ontoterminologie. In *TOTh 2007 : Terminologie et Ontologie : Théories et Applications*, pages 1–22, Annecy, France, June 2007. URL `https://hal.archives-ouvertes.fr/hal-00202645`. 22 pages.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. FrameNet II : Extended theory and practice. *Unpublished Manuscript*, 2006. URL `http://framenet.icsi.berkeley.edu/`.

Gilles Sérasset. lnterlinguai lexical organisation for multilingual lexical databases in NADIA. In *15th International Conference on Computational Linguistics, COLING 1994, Kyoto, Japan, August 5-9, 1994*, pages 278–282, 1994. URL `http://aclweb.org/anthology/C94-1044`.

Gilles Sérasset. Dbnary : Wiktionary as a lmf based multilingual rdf network. In *LREC*, 2012.

Gilles Sérasset. DBnary : Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web – Interoperability, Usability, Applicability*,

pages –, 2014. URL `https://hal.archives-ouvertes.fr/hal-00953638`. To appear.

Jaeho Shin, Sen Wu, Feiran Wang, Christopher De Sa, Ce Zhang, and Christopher Ré. Incremental knowledge base construction using deepdive. *Proc. VLDB Endow.*, 8(11) :1310–1321, July 2015. ISSN 2150-8097. doi : 10.14778/2809974.2809991. URL `http://dx.doi.org/10.14778/2809974.2809991`.

John F. Sowa. Conceptual graphs for a data base interface. *IBM Journal of Research and Development*, 20(4) :336–357, 1976. doi : 10.1147/rd.204.0336. URL `https://doi.org/10.1147/rd.204.0336`.

Robert Speer and Catherine Havasi. Representing general relational knowledge in conceptnet 5. In *LREC Proceedings*, 2012.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago : A core of semantic knowledge. In *IN PROC. OF WWW '07*, pages 697–706, 2007.

Sylvie Szulman. Logiciel Terminae - Version 2012, 2012. URL `https://hal.archives-ouvertes.fr/hal-00719453`. TERMINAE est une plateforme d'aide à la construction de ressources termino-ontologiques à partir de ressources textuelles.

Dan Tasse and Noah A. Smith. Sour cream :toward semantic processing of recipes. *T.R. CMU-LTI-08-005*, page 9, May 2008.

Andon Tchechmedjiev. *Semantic Interoperability of Multilingual Lexical Resources in Lexical Linked Data*. Theses, Université Grenoble Alpes, October 2016. URL `https://tel.archives-ouvertes.fr/tel-01681358`.

Mutsuko Tomokiyo, Mathieu Mangeot, and Emmanuel Planas. Papillon : a Project of Lexical Database for English, French and Japanese, using Interlingual Links. In *JST'00 Journées Science et Technologie*, page 3, National Olympic Memorial Youth Center, Tokyo, Japan, November 2000. URL `https://hal.archives-ouvertes.fr/hal-00968824`.

Michael Völske, Pavel Braslavski, Matthias Hagen, Galina Lezina, and Benno Stein. What users ask a search engine : Analyzing one billion russian question queries. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 1571–1580, 2015. doi : 10.1145/2806416.2806457. URL `http://doi.acm.org/10.1145/2806416.2806457`.

Andreas Witt, Ulrich Heid, Felix Sasaki, and Gilles Sérasset. Multilingual language resources and interoperability. *Language Resources and Evaluation*, 43 (1) :1–14, Mar 2009. ISSN 1574-0218. doi : 10.1007/s10579-009-9088-x. URL `https://doi.org/10.1007/s10579-009-9088-x`.

Ludwig Wittgenstein. *Philosophical Investigations*. Wiley-Blackwell, 2009.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. Probase : A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 481–492, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1247-9. doi : 10.1145/2213836.2213891. URL `http://doi.acm.org/10.1145/2213836.2213891`.

Manel Zarrouk. *Endogeneous Consolidation of Lexical Semantic Networks*. Theses, Université de Montpellier, November 2015. URL `https://hal-lirmm.ccsd.cnrs.fr/tel-01300285`.

Duygu Çelik. Foodwiki : Ontology-driven mobile safe food consumption system. *TheScientificWorldJournal*, 2015 :475410, 07 2015. doi : 10.1155/2015/475410.