UNIVERSITÉ PARIS 13 - SORBONNE PARIS CITÉ
LIPN - UMR CNRS 7030

THÈSE

pour l'obtention du grade de

Docteur en Sciences

de l'Université Paris 13

Discipline : INFORMATIQUE

Présentée et soutenue par

Nouha OMRANE

Méthode de construction d'un réseau termino-conceptuel normalisé : contribution à la construction d'ontologies à partir de textes

Sous la direction : Adeline NAZARENKO et Sylvie SZULMAN

30 Septembre 2013

Jury

M. Jean CHARLET, Chargé de mission Recherche AP-HP, INSERM (*Rapporteur*)
Mme. Marie-Claude L'HOMME, Professeur, Université de Montréal (*Rapporteur*)
Mme. Adeline NAZARENKO, Professeur, Université de Paris 13 (*Directeur*)
Mme. Sylvie SZULMAN, Maître de conférences, Université de Paris 13 (*Co-encadrante*)
Mme. Nathalie AUSSENAC-GILLES, Directrice de recherche CNRS (*Examinatrice*)
M. Patrick ALBERT, Leader IBM CAS France (*Examinateur*)
M. Aldo GANGEMI, Professeur, Université de Paris 13 (*Examinateur*)

Remerciements

Je tiens à remercier ici les personnes qui ont contribué, à divers titres, à l'achèvement de cette thèse et à rendre cette recherche intéressante, à qui je dédie de tout cœur ce travail :

Adeline Nazarenko, ma directrice de thèse. Je ne la remercierai jamais assez pour m'avoir offert l'opportunité d'effectuer cette thèse et d'intégrer l'équipe RCLN et pour avoir scrupuleusement et méthodiquement encadré mon travail. Son soutien, ses encouragements et ses conseils très précieux, tout au long de ma thèse pour bien cerner toute la problématique, m'ont permis d'évoluer tout au long de ces années de thèse et ont contribué à la maturation de mon esprit de chercheur. Je la remercie énormément de m'avoir fait confiance et de m'avoir mis, à plusieurs reprises, au devant de la scène, pour présenter nos travaux de recherche.

Sylvie Szulman, ma co-encadrante, pour avoir accepté de co-encadrer ma thèse. Je la remercie pour son soutien, ses conseils et ses recommandations, pour m'avoir guidé tout au long de cette thèse pour faire évoluer mon travail. Je la remercie énormément pour les relectures multiples de ce mémoire et pour ces corrections minutieuses. Ce fut un honneur pour moi de travailler avec elle et de bénéficier de son expérience dans le domaine de l'Ingénierie des Connaissances. J'en garderai un acquis au delà de cette soutenance.

Merci aux membres du jury qui m'ont fait l'honneur d'accepter de juger ce travail. Nathalie Aussenac-Gilles, Directrice de recherche CNRS à l'IRIT comme présidente du jury, Marie-Claude L'Homme, Professeur à l'université de Montréal et Jean Charlet, Chargé de mission Recherche AP-HP à INSERM comme rapporteurs, Patrick Albert, Leader IBM CAS France et Aldo Gangemi, Professeur à l'université de Paris 13 comme examinateurs.

Les membres de RCLN, mes collègues avec qui j'ai énormément appris. Je les remercie pour cette ambiance chaleureuse et conviviale et pour leur bonne humeur. Un grand merci à François Lévy, qui a su me conseiller, me guider et avec qui j'ai eu le plaisir de discuter de nombreux sujets. Merci à Haifa Zargayouna et Sylvie Salotti qui m'ont beaucoup aidé à nourrir ma réflexion autour des différentes problématiques auxquelles j'ai été confrontée durant mes recherches, durant les séances de l'atelier état de l'art. Il me revient de

remercier Antoine Rozenknop, Laurent Audibert et Thibault Mondary. Leurs critiques m'ont fourni matière à revoir des points importants auxquels touche cette thèse. Je remercie particulièrement le clique des thésards Sarra, Sondes, Jonathan, Hanene, les deux Inés, Nada, Zayd, Amine, Manisha, qui a permis de rendre mon travail plus agréable lors des périodes parfois difficiles.

Les partenaires du projet ONTORULE avec qui j'ai beaucoup discuté, les groupes de IBM, CTIC, Peter de Audi et surtout John Hall un spécialiste de règles métier.

L'équipe pédagogique, pour m'avoir permis d'enseigner dans de bonnes conditions et m'offrir un bon contrepoint à mes activités de recherche.

Brigitte Gueveneux, la secrétaire dans le laboratoire LIPN, qui sait se rendre disponible pour les doctorants, et qui fait preuve de patience pour gérer nos dossiers.

Mes merveilleux amis que j'adore. Je les remercie pour m'avoir soutenu, encouragé et surtout supporté particulièrement pendant la période de rédaction de ma thèse. Une pensé particulière à mes « binônes » dans le projet ONTORULE et jeunes docteurs Amina et Abdoulaye qui étaient à mes côtés depuis le début de cette aventure. J'espère que nos chemins ne cesseront pas d'être proches.

A mes frères et sœurs, Zied, Yosra, Hassan et Asma qui m'ont toujours encouragé, soutenu dans la poursuite de mes études et m'ont entouré d'affection malgré les kilomètres. En particulier, toute ma gratitude et ma reconnaissance à ma grande sœur Amira et son mari Abderrahman de m'avoir accueilli chez eux au début de mon aventure en France.

A mes parents. Je ne remercierai jamais assez mes parents pour tout ce qu'ils m'ont donné, qui ont su me donner toutes les chances pour réussir. Je dédie cette thèse à la mémoire de mon père disparu trop tôt. C'est avec une certaine émotion que j'exprime ma tristesse de ne pas avoir mon père à mes côtés pour ma soutenance de thèse. J'espère que, du monde qui est le sien maintenant, il apprécie le fruit de ce travail. J'espère que mes parents trouvent dans la réalisation de ce travail l'aboutissement de leurs efforts ainsi que l'expression de ma reconnaissance.

Résumé :

Les textes se sont imposés depuis la fin des années 1990 comme une source précieuse de connaissances pour la construction de ces ontologies qui constituent à la fois l'ossature sémantique du web sémantique et son goulot d'étranglement. Les textes sont en effet porteurs de connaissances stabilisées et partagées qui sont plus faciles d'accès que les experts qu'on pourrait vouloir interroger. Le recours aux textes ne remplace pas l'expertise humaine mais elle permet à l'ingénieur de la connaissance de prendre connaissance du domaine à modéliser et d'amorcer le travail de modélisation.

La construction d'ontologies de domaine à partir de textes repose sur des techniques de traitement automatique de la langue (TAL) couplées à des techniques d'ingénierie de connaissances pour aboutir à un modèle formel décrivant les connaissances partagées dans un domaine précis. L'un des enjeux du passage des textes à des ontologies est l'identification du vocabulaire du domaine et sa structuration sous la forme d'un thésaurus avant sa formalisation et ce sont les difficultés inhérentes à cette exploitation du matériau linguistique et à sa normalisation qui ont retenu notre attention dans ce travail de thèse.

Nous proposons une méthode de normalisation qui permet de transformer le matériau linguistique – tel qu'il a été extrait d'un corpus d'acquisition par des outils de TAL – en un réseau sémantique que nous appelons « réseau termino-conceptuel » et qui décrit le vocabulaire normalisé du domaine, c'est-à-dire le vocabulaire désambiguïsé et structuré tel qu'il est stabilisé dans le domaine en question. C'est un réseau de termes non ambigus qui sont interconnectés à travers des relations taxonomiques et associatives. Il sert de base pour la construction d'une ontologie de domaine à partir de textes mais aussi de thesaurus pour l'annotation des documents.

Cette thèse a été conduite dans le cadre du projet européen ONTORULE (ONTOlogy meets business RULEs). Notre approche s'inscrit dans le cadre global de la méthode de construction de ressources ontologiques TERMINAE qui a été initiée par les travaux du groupe *TIA* (Terminologie Intelligence Artificielle). Cette méthode TERMINAE repose sur trois niveaux de connaissances – terminologique, termino-conceptuel et conceptuel – pour la construction d'ontologies de domaine à partir de textes. La première étape d'extraction terminologique permet l'identification du vocabulaire mentionné dans les textes

et sert de point de départ pour la construction d'un modèle formel du domaine. La deuxième étape de normalisation permet de transformer le réseau terminologique initial en un réseau termino-conceptuel. La dernière étape, de formalisation, assure la transformation du réseau termino-conceptuel en un réseau conceptuel représenté sous la forme d'une ontologie. Si la première étape peut être automatisée par des outils d'extraction, les deux autres nécessitent un travail de désambiguïsation et de modélisation qui repose en grande partie sur l'expertise humaine.

Cette thèse a permis d'affiner la méthode TERMINAE en montrant comment décomposer le travail de normalisation en différentes opérations, comment enchaîner ces opérations et comment contrôler le processus global de normalisation. C'est en effet une étape difficile pour l'ingénieur de la connaissance qui se retrouve, à l'issue de la phase d'extraction linguistique, face à une masse d'unités à traiter, dont certaines sont ambiguës et qui ne sont pas toutes pertinentes pour le domaine.

Pour élaborer cette méthode de normalisation, nous nous sommes intéressée dans cette thèse à :

- l'enrichissement du réseau terminologique par la prise en compte notamment des entités nommées là où la méthode TERMINAE exploitait essentiellement les termes ;
- la formalisation de structures de connaissances manipulées dans le processus de construction d'ontologies tel que posé dans la méthode TERMINAE : nous avons défini précisément les structures de connaissances manipulées et mis en évidence les liens de correspondance qui permettent de dériver une structure de connaissances à partir d'une autre et de naviguer de l'une à l'autre ;
- la définition d'un processus de normalisation d'un réseau terminologique en un réseau termino-conceptuel qui permet de guider l'ingénieur de la connaissance dans la détection du vocabulaire du domaine et dans ses choix de normalisation : le réseau terminologique est constitué par des unités terminologiques qui sont des termes et des entités nommées et par des relations terminologiques décrivant des relations syntaxiques, lexicales et spécialisées ; des indicateurs permettent de suive la progression du travail de normalisation.

Cette approche de la normalisation a été testée dans le cadre d'expéri-

mentations visant à évaluer les principales contributions dans cette thèse. Les ontologies créées ont été utilisées dans le cadre du projet ONTORULE sur trois cas d'usage différents. Elles ont servi de vocabulaires conceptuels pour l'écriture des règles métier de différents systèmes d'aide à la décision mais elles ont surtout été utilisées pour annoter sémantiquement les textes réglementaires et ainsi guider le travail d'acquisition des base de règles métier à partir de ces textes.

Mots clés : Normalisation terminologique, Construction d'ontologies à partir de textes, Potentiel terminologique, Entités nommées.

Abstract :

Since the late 1990s, texts have emerged as a precious source of knowledge for building ontologies that are at times a semantic framework of the Semantic Web and sometimes its bottleneck. In fact, texts carry stabilized and shared knowledge which are easier to access than questioning any expert. The use of texts doesn't replace human expertise but allows the knowledge engineer to understand the domain to be modelled and initiate the work of modelling.

Building domain ontologies from text is based on techniques of natural language processing (NLP) coupled with knowledge engineering techniques to construct a formal model describing knowledge shared in a specific domain. One of the challenges of the transformation from texts to ontologies, is to detect a vocabulary of the domain and its structure in the form of a thesaurus before its formalization and these difficulties that are inherent to exploitation of linguistic material and its normalization, caught our attention in this thesis.

We propose a normalization method that transforms the linguistic material - as it was extracted from an acquisition corpus by NLP tools - in a semantic network that we call "termino-conceptual network" and that describes a normalized vocabulary of the domain : a disambiguated and structured vocabulary such that it is stabilized in the concerned domain. It is a network of unambiguous terms that are interconnected through taxonomic and associative relationships. It serves not only as the basis for building a domain ontology from texts but also as a thesaurus for annotating documents.

This thesis was conducted within the European project ONTORULE (ONTOlogy meets business RULEs). Our approach fits within the overall ontological resources construction TERMINAE method that was initiated by the work of the *TIA* group (Terminology Intelligence Artificial). TERMINAE method is based on three knowledge levels - terminological, termino-conceptual and conceptual - to build domain ontologies from texts. The first step of terminology extraction allows the identification of the vocabulary mentioned in texts that serves as a starting point for building a formal model of the domain. The second normalization step transforms the original terminology network into a conceptual network. The final step of formalization ensures the transformation of termino-conceptual network to conceptual network that is represented in the form of an ontology. If the first step can be automated by

using extraction tools, the other two require a disambiguation and modelling work that is largely based on human expertise.

This thesis helps to refine the method by showing how TERMINAE decomposes the normalization work in different operations, how these operations are enchained and how to control the overall normalization process. It is indeed a difficult step for the knowledge engineer who, after the linguistic extraction phase, is facing a mass units to process, some of them are ambiguous and not all are relevant to the domain.

To elaborate this normalization method, we are interested in :

- enrichment of terminological network by taking into account also the named entities where TERMINAE method consideres essentially the terms ;
- formalization of the knowledge structures that are manipulated in the building ontologies process as defined in TERMINAE method : we have precisely defined the knowledge structures manipulated and highlighted the correspondence links that allow deriving a knowledge structure from one another and navigating from one to the other ;
- the definition of a normalization process of a terminological network into a termino-conceptual network that guides the knowledge engineer in detecting the domain vocabulary and his normalization choices : the terminological network consists of terminological units that are formed by terms and named entities, and terminological relationships that describe syntactic, lexical and specialized relationships ; indicators allow to follow the progress of the normalization work.

This normalization approach has been experimented to evaluate the main contributions in this thesis. The ontologies created were used in the ONTORULE project for three different use-cases. They served as conceptual vocabularies for writing business rules related to different decision based systems but especially they were used to semantically annotate business documents and to guide the acquisition work of the database business rules from these texts.

Keywords : Terminology normalization, Construction of ontologies from texts, Termhood, Named entities.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Sommaire

1.1 L'enjeu de l'acquisition de connaissance à partir de textes

Ce travail s'inscrit à la croisée de l'ingénierie des connaissances et du traitement automatique de la langue. Cette thèse s'intéresse à la construction d'ontologies de domaine à partir de textes. D'après (Gruber, 1993), une ontologie est définie comme : « une spécification explicite d'une conceptualisation [1] ». Cela veut dire qu'une ontologie est une représentation formelle d'une conceptualisation. Elle repose sur des concepts qui décrivent des connaissances d'un domaine et de leurs relations. Les ontologies sont aujourd'hui utilisées dans de nombreux champs d'application comme la recherche d'information, les systèmes d'aide à la décision, etc. Comme l'ontologie décrit les connaissances d'un domaine, il faut donc chercher à les acquérir. On parle de « goulet d'étranglement » pour désigner le problème de l'acquisition des connaissances d'un domaine. Dans la littérature, plusieurs travaux ont essayé de répondre à cette question : comment acquérir des connaissances d'un domaine ? Ces travaux rentrent dans le cadre de « l'ingénierie des connaissances » définie par

1. Une conceptualisation est une abstraction du monde

(Charlet *et al.*, 2000) comme « l'étude des concepts, méthodes et techniques permettant de modéliser et/ou acquérir les connaissances pour des systèmes réalisant ou aidant des humains à réaliser des tâches ne se formalisant a priori pas ou peu ».

Différentes approches ont cherché à acquérir des connaissances relatives à un domaine donné sous la forme d'ontologies en s'appuyant sur différentes sources de connaissances. La construction d'ontologies peut être établie à partir de base de données, de schémas UML, de textes, etc (Maedche & Staab, 2001). Beaucoup de travaux s'intéressent à l'acquisition des connaissances à partir de textes. Ces textes sont des sources d'information disponibles qui décrivent des connaissances partagées dans une communauté donnée. De plus, les documents sont désormais intégrés dans des systèmes de gestion électronique de documents pour des applications d'accès au contenu. Il est donc intéressant d'explorer l'ensemble des documents qui décrivent des connaissances d'un domaine de spécialité pour la construction d'ontologies. Mais les documents sont écrits en langage naturel, tout n'est pas dit, ou certaines informations sont au contraire redondantes, ou certaines éléments sont ambigus. Il est important de distinguer les données et les connaissances. Les données sont bruitées. Les données textuelles sont une séquence de caractères. Elles doivent être analysées linguistiquement et interprétées par rapport à un domaine et une tâche pour être transformées en connaissances et il faut formaliser ces connaissances pour les intégrer dans un système automatique. L'enjeu consiste à assurer le passage du non formel (le texte) vers le formel (ici l'ontologie).

Notre travail se situe dans le domaine de « l'ingénierie des connaissances textuelles ». Il s'agit de modéliser des connaissances relevant du domaine à conceptualiser pour une application à partir de textes. L'ingénierie des connaissances textuelles vise à construire des ressources sémantiques plus ou moins formelles qui décrivent des connaissances précises d'un domaine de spécialité. Beaucoup de travaux se sont intéressés à utiliser des documents comme source d'acquisition des connaissances pour la modélisation des ontologies (Velardi *et al.*, 2006). Il existe des méthodes de construction d'ontologies à partir de textes génériques (Fernandez-Lopez *et al.*, 1997) et d'autres spécifiques (Cimiano & Völker, 2005) qui vont des textes vers des modèles semi-formels (Kassel, 2002) ou formels (Drymonas *et al.*, 2010). Les travaux de l'équipe RCLN où cette thèse a été conduite s'inscrivent également depuis longtemps

dans cette perspective (Szulman *et al.*, 2002; Aussenac-Gilles *et al.*, 2008; Charlet *et al.*, 2008) de l'ingénierie des connaissances textuelles.

1.2 Contexte

Notre travail s'inscrit dans le cadre du projet européen ONTORULE http://ontorule-project.eu. Le projet ONTORULE visait à construire des systèmes d'aide à la décision pour des domaines réglementaires. Ces systèmes fonctionnent à l'aide de règles métier. Les systèmes de gestion des règles métier (SGRMs) sont des systèmes qui aident les utilisateurs à définir et maintenir des règles métier dans le but d'automatiser des processus de prise de décision. Les règles métier décrivent les aspects métier et les contraintes relatives aux connaissances manipulées dans une organisation. Pour nous l'objectif majeur du projet est d'acquérir le vocabulaire conceptuel employé pour exprimer des règles métier à partir des textes réglementaires et d'intégrer ces documents aux SGRMs. Il s'agit ensuite d'unifier et de structurer ce vocabulaire sous la forme d'une ontologie lexicalisée pour permettre aux gens du métier d'analyser et de modifier des règles métier en s'appuyant sur ce vocabulaire.

Dans ce projet, l'équipe RCLN a mis à profit son expertise dans le domaine de l'ingénierie des connaissances. Notre contribution a consisté à revisiter la méthode TERMINAE pour la construction du vocabulaire conceptuel servant à écrire des règles métier. La méthode TERMINAE (Aussenac-Gilles *et al.*, 2008) propose une approche terminologique pour la construction d'ontologies de domaine à partir de corpus d'acquisition qui décompose le processus d'acquisition en trois niveaux : terminologique, termino-conceptuel et conceptuel. Elle permet à l'ingénieur de la connaissance de s'appuyer sur le matériau linguistique pour construire des ontologies de domaine. La première étape est une étape automatique et permet d'extraire du texte des termes candidats à partir d'un corpus d'acquisition. Durant la deuxième étape, l'ingénieur de la connaissance normalise les termes pertinents pour créer un modèle du domaine plus proche du conceptuel que du linguistique (modèle semi-formel). Au cours de la dernière étape, le réseau obtenu à l'étape précédente est formalisé en une ontologie.

Nous proposons de clarifier les structures des connaissances manipulées dans chacun des niveaux de connaissance au sein de la méthode TERMINAE.

Nous définissons une méthodologie de normalisation de réseau terminologique qui accompagne l'ingénieur de la connaissance dans la sélection, la validation et la normalisation des unités et relations terminologiques. Cette méthodologie assure un pont entre l'acquisition du vocabulaire métier au niveau terminologique et sa normalisation au niveau termino-conceptuel au sein de la méthode TERMINAE.

Dans le cadre du projet ONTORULE, nous avons travaillé sur trois cas d'usage. Le corpus d'American Airlines décrit les règles et conditions d'attribution de « miles » pour des voyageurs. Le corpus Audi est un extrait d'une directive internationale qui décrit les règles et procédures que les véhicules à quatre roues ainsi que leurs équipements doivent satisfaire pour tout ce qui touche aux ceintures de sécurité. Le corpus d'Arcelor Mittal est relatif aux règles de vérification de la conformité du produit (bobine) fabriqué au cours du processus de galvanisation et aux paramètres de fabrication définis dans les commandes des clients. Les ontologies construites décrivent le vocabulaire conceptuel utilisé dans les règles métier utilisées par les systèmes d'aide à la décision développées pour les trois applications associées à chacune : un système de calcul de points de fidélité de voyageurs, un système de vérification de la conformité de caractéristiques des ceintures de sécurité avec la norme et un système de vérification des propriétés de produits fabriqués avec les commandes des clients.

1.3 Problématique : la normalisation terminologique

Automatiser le processus d'acquisition des connaissances à partir de textes est un champ de recherche qui est encore ouvert. Dans les travaux qui portent sur la construction d'ontologies, l'acquisition des connaissances consiste en l'identification des concepts et des relations les reliant. Les travaux se concentrant sur le peuplement d'ontologies, visent à l'identification d'instances de concepts et de relations. Beaucoup de méthodes et d'outils ont été développés afin d'accompagner l'ingénieur de la connaissance dans la tâche de création d'ontologies. L'ingénieur de la connaissance n'est pas obligatoirement un expert du domaine à modéliser.

Il existe deux grandes familles d'approches de construction d'ontologies de domaine à partir de corpus de textes. La première famille comporte des méthodes automatiques ou semi-automatiques qui proposent de créer automatiquement des concepts sans passer par une analyse terminologique des éléments extraits du texte (Cimiano *et al.*, 2005; Cimiano & Völker, 2005; Bisson *et al.*, 2000). Ces travaux proposent d'appliquer des algorithmes d'apprentissage ou de regroupements de mots pour la création de concepts, ce qui génère du bruit dans le résultat obtenu et oblige l'ingénieur de la connaissance à les retravailler. La deuxième famille d'approches sont des méthodes terminologiques (Alessandro *et al.*, 2007; Aussenac-Gilles *et al.*, 2008; Velardi *et al.*, 2006) qui s'appuient sur l'analyse terminologique du texte pour la détection des concepts et des relations du domaine à modéliser.

Le passage du terme au concept n'est cependant pas une étape automatisable. Le contraste entre les deux niveaux, terminologique et conceptuel, fait que la transformation doit être bien définie et outillée de façon à guider l'ingénieur de la connaissance. Nous avons besoin d'un niveau intermédiaire qui sert de pont entre les deux plans linguistique et ontologique. Ce niveau a été défini dans la méthode TERMINAE (Aussenac-Gilles *et al.*, 2008) ; c'est le niveau *termino-conceptuel*. Le modèle des connaissances sous-jacent est un réseau termino-conceptuel composé d'un ensemble de termino-concepts interconnectés par des relations termino-conceptuelles. Les termino-concepts sont des termes désambiguisés qui sont pertinents pour la modélisation du domaine. Les relations termino-conceptuelles sont des relations taxonomiques et des relations associatives.

La phase qui assure le passage du plan linguistique (une liste d'éléments textuels extraits par des outils de traitement automatique de la langue (TAL)) vers un réseau sémantique de termino-concepts qui reflète le modèle semi-formel du domaine à conceptualiser est une phase délicate dans le processus de construction d'ontologies à partir de textes. Dans la méthode TERMINAE, l'ingénieur de la connaissance consulte une liste plate d'éléments textuels qu'il doit sélectionner, valider et normaliser pour créer le réseau termino-conceptuel correspondant. Plusieurs difficultés peuvent être rencontrées durant la création du réseau termino-conceptuel. Les éléments textuels extraits par un outil de TAL ne sont pas nécessairement pertinents pour le domaine à modéliser et l'exploration d'une large liste d'éléments textuels n'est pas évidente. De plus,

les éléments textuels extraits peuvent être ambigus. A l'inverse, certains éléments décrivent le même sens et doivent être regroupés. Nous proposons d'accompagner l'ingénieur de la connaissance dans le passage du plan linguistique vers le plan semi-formel (termino-conceptuel) appelé *la phase de normalisation* en définissant une méthodologie de normalisation. Dans la méthodologie TERMINAE au départ en effet, seule la nécessité de passer par un niveau termino-conceptuel intermédiaire était posée, pas les modalités de construction de ce niveau intermédiaire à partir des éléments extraits des textes.

Ce travail de normalisation, qui consiste à construire un réseau termino-conceptuel à partir d'un réseau terminologique, soulève les questions suivantes :

- le modèle créé doit refléter le domaine : comment détecter les connaissances du domaine ?
- les connaissances doivent être organisées : comment guider l'ingénieur de la connaissance dans leur structuration ?
- les réseaux terminologique et termino-conceptuel doivent être articulés : comment assurer ce passage ?

Le travail décrit dans cette thèse se situe dans le cadre de l'acquisition des connaissances à partir de textes. Il touche aux domaines de l'ingénierie des connaissances et au traitement automatique du langage. Nous proposons une méthodologie de normalisation de réseau terminologique qui prend en compte différents éléments textuels extraits à partir d'un corpus d'acquisition et le poids des unités linguistiques dans le corpus et par rapport au domaine à modéliser. Notre méthode GRAPHONTO complète la méthode TERMINAE (Aussenac-Gilles *et al.*, 2008) pour assurer le passage du niveau terminologique au niveau termino-conceptuel.

1.4 Objectifs

Dans le cadre de cette thèse, nous cherchons à définir des moyens qui permettent de guider l'ingénieur de la connaissance dans la validation et la normalisation du réseau terminologique extrait par des outils de TAL pour aboutir à un réseau terminologique normalisé. Un réseau terminologique normalisé est un réseau non ambigu et dont les nœuds sont interconnectés à travers des relations hiérarchiques et sémantiques. Il décrit une terminologie

de domaine stabilisée et documentée. Les éléments formant les nœuds sont bien définis et non ambigus. Ils constituent le vocabulaire propre au domaine et à l'application visée.

Nous définissons un processus interactif et itératif qui vise à construire un modèle du domaine sous forme d'un *réseau terminologique normalisé* à partir de textes. Ce réseau terminologique normalisé correspond au réseau terminoconceptuel de la méthode TERMINAE. Comme nous parlons de processus interactif, nous évoquons l'intervention d'un utilisateur. Nous définissons un utilisateur comme une personne ayant la capacité de comprendre, analyser et valider l'information et de pouvoir raisonner dessus. Cet utilisateur occupe le poste d'ingénieur de la connaissance et est au centre de notre méthodologie.

Le réseau terminologique normalisé, considéré comme ressource sémantique de domaine, est exploité de différentes manières. Il peut servir de modèle intermédiaire pour la construction d'ontologies de domaine à partir de textes. Mais, il peut aussi servir plus directement de ressource pour l'annotation des documents (Amardeilh *et al.*, 2005; Novácek, 2012).

1. Le premier apport de notre travail consiste dans la clarification des structures des connaissances que la méthode TERMINAE vise à construire.
2. Le second apport correspond à l'enrichissement du niveau terminologique qui sert de point de départ au processus de normalisation du réseau terminologique. Nous montrons comment exploiter les entités nommées là où TERMINAE exploite essentiellement des termes durant l'analyse terminologique.
3. Le troisième apport est une méthodologie détaillée du processus de normalisation qui permet de guider l'ingénieur dans son travail.

1.5 Plan de la thèse

Dans ce manuscrit, nous décrivons notre méthodologie qui s'appuie sur des techniques de traitement automatique de la langue (TAL) et d'ingénierie des connaissances. Ce travail fait partie d'une réflexion sur les principes de la méthode TERMINAE et plus généralement de l'ingénierie des connaissances textuelles. Le manuscrit est organisé en sept chapitres. Cette introduction générale constitue le premier chapitre. Le deuxième et le troisième chapitre

décrivent l'état de l'art et les notions relatives à notre problématique. Les trois chapitres suivants détaillent nos contributions. Le dernier chapitre présente la conclusion et les perspectives en vue de cette thèse.

Le chapitre deux décrit les éléments linguistiques qui peuvent être exploités pour la construction d'ontologies à partir de textes. Ce chapitre expose les méthodes et les outils qui ont exploité ces éléments linguistiques. Il décrit aussi les techniques qui permettent de les détecter dans les textes.

Le chapitre trois expose les travaux sur la construction des sources terminologiques et ontologiques en s'appuyant sur des éléments linguistiques extraits à partir d'un corpus d'acquisition (chapitre 2). Nous décrivons les méthodologies et méthodes de construction d'ontologies de domaine sous trois angles différents : les techniques utilisées pour la conceptualisation, les modalités de l'intervention de l'ingénieur de la connaissance durant le processus de construction et l'introduction de méthodologie pour accompagner l'ingénieur de la connaissance dans ses choix de modélisation.

Le chapitre quatre décrit formellement les structures des connaissances manipulées dans chaque niveau : discours, niveau terminologique, niveau termino-conceptuel et niveau conceptuel. Dans ce chapitre, nous exposons les contraintes définies pour chaque niveau, les correspondances entre les niveaux des connaissances ainsi que les opérations assurant le passage d'un niveau à un autre.

Le chapitre cinq présente la méthode GraphOnto à proprement parler. Il détaille les opérations qui permettent à un ingénieur de la connaissance de manipuler un réseau terminologique et d'en dériver un réseau termino-conceptuel. Nous montrons comment ces différentes opérations s'enchaînent dans un processus interactif et itératif.

Le chapitre six décrit les corpus que nous avons utilisés pour expérimenter notre méthodologie. Les corpus d'acquisition utilisés décrivent des règles qui portent sur un domaine particulier. Nous exposons les expérimentations faites

sur ces corpus. Dans ce chapitre, nous évaluons les hypothèses de travail de notre méthodologie.

Le chapitre sept conclut ce travail de thèse et présente quelques pistes à explorer dans son prolongement.

Chapitre 2

Les indices textuels pour la construction d'ontologies

Sommaire

Avant d'aborder les méthodes de construction d'ontologies à partir de textes, nous nous penchons sur l'étude des éléments linguistiques présents dans les textes qui permettent de détecter des notions d'un domaine à modéliser. Ces « indices textuels » peuvent caractériser des connaissances du domaine et contribuent à l'identification du vocabulaire conceptuel partagé

par une communauté. Les méthodes de construction d'ontologies à partir de textes s'appuient sur ces indices textuels, d'une manière ou d'une autre, pour créer un modèle formel (généralement sous forme d'une ontologie). Dans ce chapitre, nous définissons les indices textuels utilisés durant le processus de construction d'ontologies et nous explicitons les techniques et les méthodes qui permettent de les extraire des textes. Nous montrons aussi comment ils sont utilisés pour la construction des connaissances d'un domaine spécifique.

2.1 Pourquoi le texte (corpus) ?

(Rastier, 2005) définit un corpus comme étant « un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : de manière théorique réflexive en tenant compte des discours et des genres, et de manière pratique en vue d'une gamme d'applications ». La définition de Rastier met en évidence plusieurs aspects d'un corpus. Le choix des textes formant un corpus sont rassemblés en fonction d'une ou plusieurs applications visées. Cette vision du corpus permet d'assurer que les documents formant un corpus sont homogènes et parlent d'un même thème. On parle d'homogénéité de textes constituant un corpus quand ces derniers sont situés par rapport à un même discours (par ex. juridique, médical) ou un même genre (par ex. narratif, argumentatif). Un corpus peut être enrichi par des étiquettes comme par exemple des balises qui mettent l'accent sur la structure du texte (par ex. $< titre >< /titre >$) ou d'autres qui soulignent des éléments spécifiques du corpus (par ex. l'annotation des entités nommées de type organisation $< organisation >< /organisation >$). Un corpus peut être caractérisé par une structure spécifique, autre que la typographie, comme par exemple les textes législatifs (structurés en des parties amendements, annexes, etc) ou les fiches de consultations dans le domaine médical (structurés suivant des sections de symptômes, traitements, etc).

Des outils de traitement automatique de la langue (TAL) permettent d'extraire des indices textuels à partir d'un corpus de textes. Les indices extraits à partir du corpus peuvent être bruités (des mots comportant des caractères incorrects, des déformations).Cela est dû à des problèmes liés à l'encodage du texte ou à l'utilisation des figures et des tableaux dans un document. Il faut donc réfléchir au préalable au « mode de nettoyage » (par ex. encodage,

ponctuation) du corpus lors de son exploitation par des outils de TAL dans le but de réduire le bruit.

Partir d'un corpus pour créer un modèle formel présente divers avantages. Tout d'abord, si on souhaite collecter des informations lors des entretiens faits avec l'expert du domaine, la verbalisation de ses connaissances est généralement difficile. L'expert du domaine n'a pas toujours une vision claire des connaissances du domaine et il n'est pas souvent disponible pour de longs entretiens. En contre partie, les documents sont des sources disponibles. Ils sont intégrés dans des systèmes d'information pour servir à la recherche de l'information, la gestion électronique des documents, etc. Généralement, les documents textuels véhiculent des informations propres à un domaine ce qui favorise la collecte des connaissances stabilisées et partagées dans une communauté.

Bilan

L'intérêt porté aux textes est relatif aux informations qu'ils décrivent et qui pointent vers des éléments du domaine à modéliser. Produire un modèle formel à partir d'un corpus d'acquisition permet de le documenter et de renseigner les différents éléments du modèle en les liant aux textes qui ont servi à leur modélisation. Un modèle formel est associé à une terminologie de domaine ce qui permet la création d'une « ontologie lexicalisée » [1].

Mais parfois, les documents ne sont pas disponibles ou accessibles et l'ingénieur de la connaissance doit chercher d'autres sources d'acquisition des connaissances (par ex. des sources disponibles sur le web). De plus, les textes peuvent comporter des « sous-entendus » de la langue et cela est dû à l'emploi de la métonymie et aux ellipses. Les textes passent sous silence ce qui est du « sens commun ». Par exemple dans le corpus de *American Airlines* (voir section 6.2.1), le fait que pour tout vol d'avion il y a un point de départ et un point d'arrivée n'est pas décrit dans le texte car c'est évident pour tout vol. Il y a parfois un besoin d'expliciter des informations qui sont implicites dans le texte (i.e une relation de synonymie entre des termes : itinerary est synonyme de segment, décrivant la distance minimale entre deux points de

1. Une ontologie lexicalisée est une ontologie dont les concepts et les relations sont associés à des éléments lexicaux apparaissant dans un corpus.

vol dans le corpus de *American Airlines*[2]). Les textes ne permettent pas de donner une idée sur des choix de modélisation (par ex. pas de distinction entre concept et instance au niveau linguistique) et parfois ce qui semble pertinent au niveau linguistique ne l'est pas au niveau conceptuel. Les textes peuvent aussi comporter des erreurs ou des incohérences ce qui influe sur la modélisation de l'ontologie. Prenons l'exemple du cas d'usage de Audi (voir section 6.2.2) : nous trouvons dans deux endroits différents du texte la mention d'une contrainte de domaine et son opposée. Dans ces passages, un paramètre d'un test sur les ceintures de sécurité relatif au déplacement vers l'avant de la ceinture de sécurité (forward displacement) est mentionné une fois avec une valeur précise (« s100 and 300 mm ») et une deuxième fois en indiquant dans le texte qu'il peut excéder la valeur définie dans le premier passage (« may exceed that specified in paragraph 6.4.1.3.2 »). Les deux occurences du paramètre de test sont décrits dans l'extrait tiré du corpus de Audi suivant :

6.4.1.3.2 « In the case of other types of belts, the forward displacement shall be between 80 and 200 mm at pelvic level and between 100 and 300 mm at chest level ».

6.4.1.3.3 « In the case of a safety-belt intended to be used in a seating position protected by an airbag in front of it, the displacement of the the chest reference point may exceed that specified in paragraph 6.4.1.3.2 above if its speed at this value dœs not exceed 24 km/h.»

Dans le traitement automatique de la langue, la taille et la structuration du corpus (par ex. les textes balisés) jouent un rôle important (Rastier, 2005). Cela suppose, de constituer un corpus d'acquisition qui soit « représentatif » du modèle formel que l'on souhaite créer. La tâche s'annonce difficile car on ne peut pas mesurer « la représentativité » du corpus.

Le critère du choix du corpus joue un rôle important et même primordial dans la qualité de structuration de l'ontologie. Si un corpus est porteur de données pertinentes et non ambiguës, alors la probabilité de réussir la conception d'une ontologie augmente. En effet, les documents circulant dans une organisation véhiculent de l'information sur son domaine d'activité. Plus les documents sont structurés et porteurs de connaissance de l'organisation, plus

2. Cet exemple est tiré du cas d'usage de American Airlines.

l'ontologie reflète le domaine étudié. Les documents ne sont pas une « juxtaposition » de données mais ils sont produits dans un contexte donné et sont interprétés selon un contexte applicatif. Parfois les informations extraites sont soit très génériques, soit dépendantes du corpus d'acquisition. L'ingénieur de la connaissance se trouve, aussi, face à beaucoup de données à traiter et à l'ambiguïté de certaines. Nous avons en plus besoin de l'expert pour la validation des informations extraites comme elles ne sont pas toutes pertinentes pour le domaine à modéliser.

Malgré ces inconvénients, l'ingénieur de la connaissance reste « familier » avec le langage naturel. Le modèle formel construit à partir du corpus d'acquisition peut être ensuite enrichi par d'autres ressources (par ex. ressources existantes relatives au domaine à modéliser). Nous nous intéressons à des corpus de spécialité qui portent sur des réglementations de différents domaines (par ex. programme de fidélité d'une compagnie aérienne, réglementation européenne sur les ceintures de sécurité). Les corpus d'acquisition que nous avons utilisés ne sont pas volumineux (à l'ordre de 6000 mots). Dans la section suivante, nous décrivons les indices textuels qui sont mentionnés dans le texte et qui peuvent contribuer à la construction d'ontologies de domaine.

2.2 Une multitude d'indices textuels

Dans cette section, nous définissons les différents types d'indices textuels utilisés pour la construction d'ontologies. Pour chaque indice, nous expliquons pourquoi ils sont intéressants à exploiter. Ensuite, nous mentionnons les principaux travaux qui les exploitent pour la construction d'ontologies. Enfin nous spécifions pour chacun de ces indices ses limites d'utilisation.

2.2.1 Termes

(Lerat, 2009) définit la notion de terme comme : « le nom donné dans une langue à une entité conceptualisée par une communauté de travail. Cette dénomination est souvent un nom ou un groupe nominal, mais elle peut aussi appartenir à une nomenclature alphanumérique, une unité définie dans les textes de spécialité ». Nous définissons le terme comme une unité syntaxique qui peut être composée d'un ou plusieurs mots et qui décrit une notion parta-

gée dans une communauté et un domaine spécifique. Les termes ont généralement un sens précis dans le domaine auquel ils appartiennent, ils sont moins ambigus que les mots courants et présentent moins de variation de forme. Prenons comme exemple les termes AAdvantage member (membre adhérant) et miles (mesure de kilométrage des vols) extraits du cas d'usage de *American Airlines* (voir section 6.2.1) qui décrivent des notions relatives à l'attribution de bonus aux voyageurs adhérants dans un programme de fidélité.

La détection des termes dans un corpus permet d'identifier le vocabulaire conceptuel propre à un domaine et donc le repérage des mentions linguistiques de concepts[3]. Par exemple, dans la méthode *DODDLE* (Morita *et al.*, 2008), c'est à l'ingénieur de la connaissance de sélectionner les termes qui lui semblent pertinents en s'appuyant sur des mesures de fréquence et de tf-idf. Un autre exemple, dans la méthode *OntoGain* (Drymonas *et al.*, 2010), les auteurs utilisent la mesure *CNC value* pour détecter le potentiel terminologique des termes composés (des syntagmes nominaux). La mesure *CNC value* privilégie les termes composés plutôt que les termes simples et donne de l'importance aux termes qui co-occurrent dans un même contexte. Plusieurs travaux s'appuient sur les termes pour la création des concepts au niveau formel (Cimiano & Völker, 2005; Aussenac-Gilles *et al.*, 2008; Charlet *et al.*, 2008; Alessandro *et al.*, 2007). D'autres travaux s'intéressent à l'organisation des termes extraits à partir d'un corpus sous la forme d'un réseau terminologique (Ibekwe-SanJuan, 2005; Xu *et al.*, 2002) ou d'une taxonomie liée à un domaine (Navigli *et al.*, 2011).

Bilan

La détection des termes à partir d'un corpus d'acquisition permet essentiellement d'identifier des mentions linguistiques de concepts. Mais, il ne faut pas considérer qu'un terme est équivalent à un concept car le premier fait encore partie de la langue et le deuxième fait partie d'un modèle conceptuel. Les termes peuvent dénoter plusieurs sens dans le texte. De ce fait, un terme peut correspondre à plusieurs concepts s'il décrit plusieurs sens pertinents pour la modélisation d'un domaine. Prenons l'exemple du cas d'usage de *American*

3. (Charlet *et al.*, 2008) définit le concept comme une description d'un objet du monde, qui a une signification propre dans un contexte donné.

Airlines où le terme member dénote deux sens dans le texte : (i) un voyageur adhérant dans le programme de fidélité (synonyme de Elite member) , (ii) une compagnie aérienne participant dans le même programme (synonyme de AAdvantage participant), et qui sont pertinents pour la création de deux concepts au niveau formel. De même un concept peut être exprimé dans le texte à travers plusieurs termes. Par exemple, un concept *Member* est décrit dans le texte par les termes *AAdvantage member* et *customer*. La relation entre un terme et un concept est donc une relation de cardinalité *nxn*.

2.2.2 Entités nommées

Plusieurs travaux ont cherché à définir la notion des entités nommées (EN). Prenons par exemple les énoncés suivants : « Named entities are phrases that contain the names of persons, organizations and locations » (Tjong Kim Sang & De Meulder, 2003), « On appelle traditionnellement entités nommées (named entity) l'ensemble des noms de personnes, d'entreprises et de lieux présents dans un texte donné. On associe souvent à ces éléments d'autres syntagmes comme les dates, les unités monétaires ou les pourcentages repérables par les mêmes techniques à base de grammaires locales. » (Poibeau, 2005).

Les entités nommées sont des unités textuelles particulières. Création du domaine de traitement automatique de la langue comme le souligne (Ehrmann, 2008), elles ont une histoire moins ancienne que les termes. Les entités nommées sont décrites par des « mentions d'entités nommées » qui sont des unités textuelles qui renvoient à des « entités » du domaine qui peuvent relever de différentes catégories linguistiques : des noms propres ("Air France"), mais aussi des pronoms ("elle"), et plus largement des descriptions définies ("cette compagnie", "la principale compagnie aérienne française"). En pratique, on met l'accent sur les noms propres qui sont plus faciles à identifier dans le flux textuel, en sachant qu'on n'identifie ainsi qu'un sous-ensemble des entités mentionnées dans le texte.

Or, on s'intéresse généralement aux entités nommées du fait de leur valeur référentielle : ce sont des expressions linguistiques qui renvoient de manière autonome et non ambigue, dans un contexte donné [4], à des entités du monde

4. Les cas d'ambiguïté d'entités nommées ne sont pas rares mais, dans un contexte donné, il s'agit essentiellement de métonymie.

(Ehrmann, 2008). En nous inspirant de la reconnaissance des ENs, nous définissons une entité nommée comme une unité lexicale du texte qui désigne de manière univoque une entité référentielle. Une entité nommée est « une mention » qui renvoie à un référent (valeur sémantique référentielle et stable). L'exploitation des entités nommées est relative à un cadre applicatif et une tâche (Ehrmann, 2008) qui permettent de déterminer quels sont les types sémantiques d'entités nommées pertinents à prendre en compte. Par exemple, on s'intéresse ainsi davantage aux noms de gènes et de maladies dans le domaine biomédical qu'aux noms d'organisations ou de personnes.

Beaucoup de travaux se sont intéressés à l'exploitation des entités nommées pour des applications d'accès au contenu comme l'annotation sémantique, la recherche d'informations et le peuplement des ontologies [5]. Le peuplement d'une ontologie à partir de textes est un champ d'application de l'extraction d'information (Buitelaar *et al.*, 2005), si l'on considère que la structure d'information (ou formulaire) à remplir est en réalité une structure ontologique qui sert de grille d'interprétation dans le processus d'extraction (Bontcheva & Cunningham, 2003; Nédellec *et al.*, 2009). Une tâche très classique d'extraction d'information consiste à repérer des unités textuelles particulières qu'on appelle « entités nommées ». L'idée de l'exploitation des entités nommées pour le peuplement des ontologies (Magnini *et al.*, 2006; Fort *et al.*, 2009; Manzano-Macho *et al.*, 2008) repose sur l'hypothèse que la mention d'un nom d'entité de type T dans un texte révèle l'existence de cette entité et permet de créer une instance du concept T dans l'ontologie. Des problèmes d'ambiguïté dans la détermination du type T et de rapprochement des noms correspondant à la même entité peuvent se poser mais le principe de base reste simple. (Tanev & Magnini, 2008) cherche ainsi à collecter des instances de lieux et de personnes à partir de Wikipedia en les catégorisant précisément [6]. *OntoGen* (Blaz *et al.*, 2006) extrait automatiquement les entités nommées qui vont servir pour la construction des concepts et leurs co-occurrences pour l'apprentissage des relations entre concepts.

5. Le peuplement de l'ontologie est le fait d'ajouter des instances à la base de connaissance. Il faut faire la distinction entre cette notion et celle de l'enrichissement de l'ontologie. L'enrichissement de l'ontologie est l'ajout de concepts dans la hiérarchie d'un modèle formel.

6. Cinq sous-types de lieux et de personnes sont pris en considération : MOUNTAIN, LAKE, RIVER, CITY, COUNTRY ; STATESMAN, WRITER, ATHLETE, ACTOR, INVENTOR.

Plusieurs travaux ne font pas de différence entre terme et entité nommée (Nobata *et al.*, 2000). Par conséquent, une entité nommée peut être vue comme un concept ou une instance, tout dépend de l'interprétation faite par l'ingénieur de la connaissance (choix de modélisation). Les entités nommées sont généralement interprétées comme des instances des concepts présents dans une ontologie (niveau A-Box en logique de description). La détection des mentions d'entités nommées dans le texte permet aussi d'enrichir les attributs (par ex. relier des valeurs numériques à un attribut) des concepts décrits dans une ontologie et les relations sémantiques entre concepts (par ex. Nicolas Sarkozy *est le président* de la France) (Amardeilh *et al.*, 2005).

Bilan

Si beaucoup de travaux se sont intéressés à l'extraction des ENs c'est qu'elles offrent certains avantages dans la compréhension des documents. Les ENs apportent une certaine clarification sur le contexte du domaine. En effet, l'identification de ce type d'unité lexicale spécifique fournit une information sur le contenu des documents (par exemple il s'agit du domaine médical ou juridique). Il est donc astucieux d'utiliser ces informations pour la création des concepts.

Le traitement d'entités nommées : l'identification des noms de personnes, de lieux, de dates et de leurs catégorisations favorise la désambiguïsation du texte. En effet, les unités lexicales peuvent exprimer divers sens ce qui est bien naturel. La présence d'unités nommées dans le texte facilite l'identification du contexte dans lequel se trouvent les termes et donc du sens de ces unités lexicales. Prenons par exemple un corpus qui parle de maladies, les noms propres extraits tels que les noms de maladies donnent une idée sur le fait que le corpus est relatif au domaine médical. Les EN permettent de comprendre de quoi il s'agit : ce sont des « ancres référentiels ». Elles sont importantes pour l'accès au contenu textuel.

Dans la suite du manuscrit, nous allons nous intéresser à la forme « nom propre » des entités nommées. L'aspect référentiel d'une entité nommée est le point qui a suscité notre attention lors de notre quête à extraire les connaissances afin de modéliser l'ontologie de domaine. Ce qui nous intéresse est le fait que les EN apportent un coté informationnel au contexte où elles sont

mentionnées.

2.2.3 Motifs linguistiques

Les motifs linguistiques représentent des formes de phrases connues et repérables dans les documents. Ce sont des configurations d'éléments textuels (des expressions régulières) souvent lexicaux et syntaxiques. L'un des éléments est appelé « marqueur ». Il dépend, dans sa définition, du type de l'élément à extraire du texte (par ex.terme, relation terminologique). Par exemple, (Aussenac-Gilles & Séguéla, 2000) définit un marqueur d'extraction de relations terminologiques entre termes comme « une formule linguistique dont l'interprétation définit régulièrement le même rapport de sens entre des termes ». Les marqueurs varient suivant des catégories grammaticales (sujet, objet), lexicales (lemme), syntaxiques (verbe, nom), etc. Par exemple pour une relation d'hyperonymie, un patron est défini par le marqueur « est un » ainsi que la structure syntaxique des unités lexicales potentiellement liées entre elles à travers la relation « SN [7] tel que SN » qui permet d'extraire du texte tous les couples (*t1*, *t2*) tels que *t1* est père de *t2*.

La création des motifs linguistiques pour l'extraction des relations entre termes a été initiée par les travaux de (Hearst, 1992) puis (Morin, 1999). Les patrons sont utilisés dans le but d'extraire des éléments linguistiques pour la construction d'ontologies comme par exemple des termes (Xu *et al.*, 2002), des relations sémantiques (Aussenac-Gilles & Séguéla, 2000), des entités nommées (Amardeilh *et al.*, 2005), etc. Ils sont aussi exploités pour repérer des passages particuliers dans le texte comme par exemple l'identification des énoncés définitoires (Rebeyrolle & Tanguy, 2000). Prenons comme exemple un patron qui permet de repérer des passages définitoires ayant la forme de « SN s'appelle SN » et qui identifie le contexte suivant « un type de ceinture s'appelle une catégorie de ceintures ».

Ces motifs linguistiques dépendent fortement du domaine et du corpus à partir duquel ils sont construits. En effet, ces motifs se basent pour leur définition sur des marqueurs qui sont des éléments linguistiques propres à un domaine donné (par ex. « SN a pour symptôme SN » dans le domaine médical). Ils dépendent aussi du corpus dans le mesure où les marqueurs qui

7. SN pour syntagme nominal

sont employés pour leurs définitions sont des mentions linguistiques pouvant apparaître ou pas dans les documents constituant un corpus.

Les patrons trop génériques extraient du bruit. En effet, prenons comme exemple un patron générique qui permet d'extraire une relation d'hyperonymie entre termes « SN est-un SN, SN,.. SN ». Ce patron extrait toutes les mentions linguistiques partageant cette même structure lexicale et qui ne dénotent pas forcement une relation d'hyperonymie (par ex. x *est-un* antécédent de y). L'emploi des patrons génériques nécessite une étape de nettoyage par l'ingénieur de la connaissance afin d'enlever ce qui n'est pas pertinent pour la construction d'ontologies. Par ailleurs, les patrons qui sont spécifiques génèrent du silence. En effet, l'emploi de patrons spécifiques qui sont fortement liés à un corpus (par ex. « SN accumule SN ») ne permet pas d'extraire toutes les mentions linguistiques relatives aux patrons employés (par ex. « tout membre du programme AA doit accumuler en moyenne 500 miles par an » n'est pas extrait).

Bilan

Les motifs linguistiques ne sont pas utilisés en tant que tels dans des ontologies mais permettent d'extraire à partir du corpus d'acquisition des éléments (par ex. terme, entité nommée) qui sont utiles pour leur conceptualisation. Nous proposons de considérer les patrons linguistiques qui ont permis de détecter des éléments linguistiques à partir du texte comme des propriétés linguistiques relatives aux éléments qui composent une ressource sémantique et identifiés dans le texte grâce à l'application de ces patrons. Ce qui a pour conséquence de garder un lien entre des patrons linguistiques et les éléments linguistiques détectés grâce à l'utilisation de ces derniers.

2.2.4 Classes sémantiques

Les classes sémantiques sont des groupements de mots sémantiquement proches. Elles sont généralement construites sur des critères distributionnels, ce qui signifie que les mots d'une classe partagent les mêmes contextes. Les classes sémantiques sont considérées comme des ébauches de concepts pour la construction d'ontologies (Fuchs *et al.*, 2010). La méthode ASIUM utilise des techniques d'apprentissage pour construire des classes sémantiques (Faure &

Nédellec, 1999). Plus précisément, c'est une méthode qui aide l'ingénieur de la connaissance dans la sélection des ébauches de concepts pertinents pour le domaine à modéliser. ASIUM utilise un corpus non bruité (nettoyé au préalable) afin de réduire l'ambiguïté lors de l'extraction de relations syntaxiques. ASIUM s'appuie sur un calcul du degré de similarité entre mots pour créer des classes sémantiques. Les classes sont ré-organisées progressivement dans un arbre de généralisation. La méthode ASIUM prend en considération la taille du corpus en éliminant les termes peu fréquents (bruit) et ceux qui sont très fréquents (inutiles).

Bilan

Tout un travail de choix de modélisation est à faire par l'ingénieur de la connaissance. Il doit décider si les classes formées sont des ébauches de concepts ou si elles sont pertinentes pour la création de relations sémantiques. De plus, il faut valider manuellement les classes sémantiques créées. le bon niveau de granularité. C'est un travail d'autant plus difficile et fastidieux que cette validation n'est pas contrôlée par des moyens ou des procédés qui permettent d'avoir une idée sur la progression du travail de filtrage.

Souvent les classes sémantiques sont des groupements de mots. Pourtant, les unités linguistiques qui décrivent des connaissances sont généralement des termes composés. Il est important alors de définir une méthode distributionnelle qui prend en compte la distribution des termes plutôt que les mots. (Ban, 2011) propose une approche distributionnelle qui s'appuie sur le rapprochement de termes polylexicaux qui partagent des contextes similaires. Le contexte d'un terme est défini par l'ensemble des noms et des verbes figurant dans la même phrase. L'approche proposée repose sur une pondération des mots formant les contextes afin de mettre l'accent sur ceux qui sont pertinents à prendre en compte dans la détermination des contextes. Par exemple pour le terme *ticket*, tiré du cas d'usage *American Airlines*, qui occure dans la phrase « Accrued mileage credit and tickets do not constitute property of the member », a comme contexte « mileage, credit, constitute, member ». Des mesures de similarité sont ensuite exploitées pour le calcul des classes sémantiques en s'appuyant sur ces contextes pondérés. Un exemple d'une classe sémantique *Ticket* est formée par les termes simples et composés suivant : *award ticket,*

fare ticket rate, agency, flight ticket, industry, cost, flight number.

Par rapport au domaine de la construction d'ontologies, parfois l'ingénieur de la connaissance a besoin de créer – dans une première itération du processus de construction d'ontologies– une ébauche d'ontologie afin d'avoir une idée globale sur le domaine à modéliser. Il peut ensuite enrichir l'ontologie dans une deuxième itération. Mais l'exploitation des classes sémantiques, dans ce cas, n'est pas évidente car l'interprétation du résultat obtenu par l'ingénieur de la connaissance est une tâche nécessitant déjà une bonne compréhension du domaine.

A ce stade de notre travail, nous nous intéressons moins à la construction de classes sémantiques qu'à l'identification des unités qui les composent, et qui correspondent au vocabulaire conceptuel d'un domaine. Dans la section 2.3, nous décrivons les approches qui permettent l'extraction d'indices textuels pour la construction d'ontologies à partir de textes.

2.3 Méthodes d'extraction d'unités textuelles

Dans cette section, nous décrivons les principes sur lesquels s'appuient les méthodes d'extraction d'indices textuels. Il existe deux grandes familles d'approches : celles qui sont guidées par le but et d'autres qui sont guidées par les données. La première famille d'approches vise à extraire des connaissances explicitement mentionnées dans le corpus d'acquisition. La deuxième famille cherche plutôt à extraire des connaissances implicites. Les deux types d'approches s'appuient sur différents critères de pertinence pour l'extraction des connaissances.

2.3.1 Critères de pertinence

Les méthodes d'extraction d'indices textuels s'appuient sur la structure du texte pour la détection des éléments potentiellement pertinents. Le repérage de passages marquants dans un texte favorise la détection d'unités textuelles potentiellement pertinentes pour un domaine spécifique (Rastier, 2005). Prenons un exemple tiré du cas d'usage d'Audi. Le texte est composé d'un ensemble d'articles (*Article 1*, *Article 2*, etc). Les titres de ces derniers mentionnent des éléments clés du domaine qui sont donc faciles à repérer pour l'ingénieur de

la connaissance.

Indices de pertinence

La pertinence d'une unité textuelle peut être marquée de différentes manières. Dans certains cas, c'est la structure syntaxique de la phrase qui compte parce que le terme est introduit dans une tournure marquante (par ex. les tournures de présentation « c'est..qui »). Dans d'autres cas, c'est la fréquence du terme qui en marque l'importance. Dans d'autres cas encore, c'est la typologie qui compte (caractères gras ou italiques, majuscules ou guillemets). Même si elles sont souvent ambiguës, ces marques sont souvent utilisées comme indice de pertinence.

Zones d'intérêt dans le texte

Un texte peut avoir une certaine structure logique qui permet de mettre l'accent sur des passages où sont mentionnées des unités textuelles pertinentes pour la modélisation. En effet, quand un texte a du relief, certains passages textuels ayant une structure spécifique (par ex. texte divisé en parties d'articles dans un corpus juridique). Un autre exemple de structure peut se formaliser à travers un balisage du texte (par ex. HTML). Par exemple les balises $< titre >< /titre >$ qui entourent un titre donnent un statut spécifique à la valeur entourée. Le repérage aussi du résumé, une introduction ou la présentation de résultats guide l'analyse du texte. Prenons comme exemple des documents juridiques ; ils sont structurés en articles, annexes, etc. Des comptes rendus hospitaliers sont composés en parties décrivant des informations sur le patient, ses symptômes, ses traitements, etc.

Nous nous intéressons principalement aux passages définitoires. L'étude des énoncés définitoires a pour objectif l'identification des unités linguistiques pertinentes et des relations entre termes (par ex. hyperonymie, méronymie, etc). Cette question a fait l'objet de plusieurs travaux en terminologie (Seppala, 2012) et en ingénierie des connaissances. Certains travaux se basent sur les patrons lexico-syntaxiques (Rebeyrolle & Tanguy, 2000), d'autres sur un calcul statistique (Hearst, 1992) pour l'identification des énoncés définitoires.

2.3.2 Approches guidées par le but

Les approches guidées par le but sont une famille d'approches qui extraient du texte des éléments linguistiques qui peuvent être préalablement identifiés. Utiliser ces techniques suppose qu'on connaît au départ ce que l'on cherche dans les textes (recherche guidée par le but). Ce sont des approches qui s'appuient sur le contexte et la structure (interne ou externe) de l'élément à extraire du texte. Ces approches s'appuient sur des patrons linguistiques (voir 2.2.3) pour l'extraction des éléments linguistiques. Ces patrons tendent à extraire l'information qui est explicitement mentionnée dans le texte en s'appuyant sur le contexte. Il s'agit de définir *a priori* des types d'objets à extraire (par ex. une entité nommée, une relation d'hyperonymie) et les patrons lexico-syntaxiques qui permettent de détecter dans le texte les fragments qui expriment ces types de relations. Par exemple, si on souhaite extraire des entités nommées décrivant des personnes, un patron linguistique peut être « Monsieur + Nom Propre ».

La qualité des résultats obtenus dépend de la qualité des patrons employés et du corpus utilisé (Aussenac-Gilles & Jacques, 2008). En effet, les éléments extraits ne sont généralement pas toujours corrects (par ex. un patron peut générer du bruit) et parfois les marqueurs utilisés dans des patrons ne correspondent pas exactement aux phrases décrites dans les documents.

2.3.3 Approches guidées par les données

Les approches guidées par les données visent à extraire des informations implicitement décrites dans le texte. Elles analysent des phénomènes qui peuvent être détectés entre éléments linguistiques partageant des contextes similaires. On trouve principalement les approches distributionnelles initiées par les travaux de (Harris, 1954). L'analyse distributionnelle s'appuie sur l'étude de contextes distributionnels des unités lexicales afin de grouper des mots qui peuvent partager les mêmes caractéristiques. Sous l'hypothèse que les termes qui apparaissent dans les mêmes contextes sont sémantiquement similaires, les approches distributionnelles cherchent à grouper des mots en classes sémantiques. Ces approches définissent tout d'abord une « fenêtre » qui délimite le contexte pour le calcul de la distribution d'une unité lexicale. Ensuite, elles appliquent une mesure de similarité afin de calculer la similarité entre deux

mots représentés par leurs vecteurs de contexte. Un algorithme de classification (regroupement ascendant, descendant, etc.) prend en entrée la liste des mots obtenus et donne en sortie des groupements de mots. Certains outils permettent de configurer l'environnement en choisissant la fenêtre, la mesure de similarité et l'algorithme de clustering à appliquer. Prenons l'exemple de l'outil Mo'K (Bisson & Nedellec, 2001) qui est un atelier configurable d'aide à la conception d'algorithmes de classification. Il offre des services permettant de comparer, d'évaluer, de caractériser les méthodes de classification et d'implémenter des mesures de similarité. Certains extracteurs distributionnels ne tiennent pas compte de la syntaxe du contexte (Church & Hanks, 1989; Basili *et al.*, 1996). Ils proposent de définir une fenêtre (en nombre de mots) et de choisir si le mot apparaît ou non au centre de la fenêtre. Une fois les contextes extraits, ils sont représentés dans un modèle vectoriel accompagné du nombre d'occurrences des contextes dans le texte pour chaque unité extraite. Puis les mots similaires sont extraits suivant une mesure de similarité appliquée sur les contextes similaires.

Une autre branche d'approches guidées par les données est celle de l'Analyse des Concepts Formels (ACF) ; elle s'appuie sur l'analyse syntaxique des phrases du texte pour la création d'un « contexte formel ». Un contexte formel est un ensemble d'objets et d'attributs tel que les objets sont les formes lemmatisées des noms qui sont des *sujets* ou des *compléments d'objet* dans les phrases analysées et les attributs sont les verbes associés. En se basent sur le contexte formel, l'ACF permet de former des concepts formels, c'est-à-dire des classes de noms partageant les mêmes attributs verbaux (Cimiano *et al.*, 2011; Bendaoud *et al.*, 2007).

La limite majeure des approches guidées par les données est que ces dernières ne proposent pas d'aide à l'ingénieur de la connaissance durant la validation des classes sémantiques formées ni des moyens de contrôle de cohérence des classes formées. Les groupements de mots formés peuvent être incorrects car la liste des unités textuelles utilisée initialement pour le calcul distributionnel n'est pas nettoyée au préalable ce que induit l'ajout d'éléments et de contextes peu pertinents dans le processus de construction de classes sémantiques (par ex. *industry* et *agency* qui sont groupées avec *ticket*).

Dans la suite de ce chapitre, nous nous intéressons aux méthodes qui permettent l'extraction des termes, des entités nommées et des relations termi-

nologiques pour la construction des ontologies à partir des textes.

2.4 Extraction terminologique

Dans cette section, nous nous intéressons plus spécifiquement aux travaux qui portent sur l'identification des termes. Tout d'abord, nous décrivons les indices qui peuvent aider à la détection des termes et en particulier les indices terminologiques (relatifs à la structure interne ou externe des termes et à leur distribution dans les textes). Ensuite, nous explicitons quelques travaux qui s'intéressent à la détection des termes candidats.

2.4.1 Les indices terminologiques

Il existe, dans la littérature, plusieurs critères permettant d'estimer *a priori* « le potentiel terminologique » des termes candidats qui sont généralement des syntagmes nominaux composés de plusieurs mots susceptibles de désigner par leurs propriétés syntaxiques et contextuelles une *notion du domaine*. Les indices terminologiques sont généralement liés à des caractéristiques structurelles, soit des termes, soit du corpus d'acquisition. Ils sont aussi liés à la distribution des termes dans le texte. L'association des termes à des noms toponomiques ou à des verbes de mouvement, la structure des documents et la typographie, la fréquence, la distribution et la densité sémantique du réseau terminologique jouent aussi un rôle dans la détermination du « le potentiel terminologique » des termes candidats. Les méthodes de détection des termes s'appuient sur ces caractéristiques et peuvent les combiner.

2.4.2 Méthodes de détection des termes

Les outils de TAL extraient des mots ou des groupes de mots. L'objectif de l'extraction automatique des termes est de déterminer si un mot ou un ensemble de mots est un terme qui est relatif au domaine étudié. Nous distinguons trois grandes familles d'approches.

Les approches linguistiques

s'appuient sur la structure linguistique des termes candidats et/ou de leurs contextes. Les termes sont codés sous forme de couples « tête-modifieur » (*i.e.* *Traval award* a pour tête *award*). (Frantzi & Ananiadou, 1997) ne s'intéresse qu'aux termes ayant comme voisins dans le texte des noms ou des verbes. (Maynard & Ananiadou, 1999) enrichit cette approche en prenant en compte les relations sémantiques (taxonomiques) entre termes qui ont été détectées à partir d'une ressource externe (taxonomie) afin de détecter les termes qui sont similaires (ayant le même père dans la taxonomie utilisée comme ressource sémantique).

Les approches statistiques

reposent sur le calcul de la fréquence et de la distribution des termes dans le corpus. Les termes sont extraits à partir de leurs occurrences dans les textes. Les termes les plus fréquents sont considérés comme les plus pertinents à identifier. La mesure *tf.idf*[8] permet d'identifier des termes qui apparaissent fréquemment dans un corpus mais spécifiquement dans quelques documents constituant ce corpus. Une autre mesure utilisée pour l'extraction des termes est *l'entropie*[9]. Cette mesure permet d'identifier les termes suivant leur répartition dans les documents constituant un corpus.

(Wilson *et al.*, 2007) propose un nouvel indicateur, appelé *TermHood*[10], qui décrit la distribution d'un terme candidat dans un corpus d'un domaine par contraste avec des corpus de langue générale ou d'autres domaines. D'autres travaux se sont inspirés de la mesure *tf.tdf* pour pondérer les termes en fonction de leur distribution dans les documents constituant le corpus. (Drouin, 2003) propose une méthode d'extraction de termes pertinents pour des corpus de spécialité. La méthode s'appuie sur deux mesures statistiques appliquées à deux corpus l'un de spécialité et l'autre général : le calcul de la fréquence des éléments lexicaux et la probabilité d'observer une fréquence d'un élément égale ou supérieure à celle du corpus de spécialité. L'outil qui supporte la méthode s'appelle *TermoStat*[11].

8. http://www.tfidf.com
9. http://fr.wikipedia.org/wiki/Enropie_de_Shannon
10. Le potentiel terminologique d'un terme en français.
11. http ://olst.ling.umontreal.ca/ drouinp/termostat_web/

Ces approches pondèrent des groupes de mots dans le but de détecter des termes potentiellement pertinents pour un domaine donné. Un seuil numérique est défini afin de filtrer les listes produites.

Les approches mixtes

combinent des critères linguistiques et statistiques. (Daille, 1994) extrait les termes pertinents à partir d'une analyse statistique du texte et d'un filtrage linguistique des termes candidats. *Acabit*[12] (Daille, 2003) produit une liste ordonnée des termes les plus représentatifs d'un domaine spécifique ainsi que les variations morphologiques sous la forme de groupes de termes candidats en s'appuyant sur les co-occurrences des termes candidats.

Bilan

Dans (Drouin & Langlais, 2006), les auteurs ont évalué les approches linguistiques et statistiques en comparant les résultats obtenus aux termes occurrant dans un corpus annoté à la main par un expert et en utilisant les mesures de précision et rappel. Les auteurs en concluent que la fréquence est un bon indice de pertinence des termes par rapport à d'autres indices statistiques. Ils notent que certaines mesures statistiques privilégient les termes courts et d'autres les termes composés et que leur combinaison permet d'augmenter le rappel.

Généralement, les approches d'extraction de la terminologie extraient des candidats termes dont certains constituent du bruit. Les outils qui implémentent ces approches sont des outils génériques et qui ne sont pas liés à un domaine donné. Il existe beaucoup de méthodes d'extraction des termes mais on ne peut pas les évaluer ou préciser leur domaine d'application car la qualité des résultats obtenus dépendent de la taille de corpus, du domaine, du genre du corpus, etc (Mondary *et al.*, 2012). Les indices utilisés pour l'extraction des termes ne donnent pas une idée sur le domaine directement, à l'exception des travaux de (Drouin, 2003).

Pour notre part, nous nous situons en aval du processus d'extraction terminologique. Nous utilisons le résultat de ce type d'outil. Nous ne cherchons pas le « bon » outil d'extraction des termes et préférons laisser l'utilisateur choisir

12. Automatic Corpus Based Acquisition of Binary Terms

l'outil le mieux adapté au domaine qu'il étudie. Nous proposons de pondérer des termes dans le but de filtrer le résultat obtenu par un extracteur de termes et de mettre en évidence des termes du domaine. Nous définissons une mesure sémantique de pondération de termes qui tient compte de plusieurs indices reliés au domaine et aux caractéristiques du corpus.

2.5 Extraction des entités nommées

Dans cette section nous nous intéressons à la reconnaissance des entités nommées. La reconnaissance d'entités nommées (REN) était vue au départ comme une sous-tâche de l'extraction d'information. L'extraction d'information est considérée comme l'une des tâches du traitement automatique de la langue. « Elle consiste dans un domaine restreint, à extraire des éléments d'information précis à partir d'un ensemble de textes homogènes et à remplir des formulaires prédéfinis avec ces éléments d'information » (Ehrmann, 2008). Dans la suite, nous définissons les méthodes qui permettent leur reconnaissance dans le texte.

2.5.1 Méthodes de reconnaissance d'entités nommées

L'intérêt porté à ce type spécifique d'unités lexicales est apparu depuis une vingtaine d'années pendant les conférences américaines sur la compréhension de messages MUC (Message Understanding Conferences). La reconnaissance d'entités nommées est une phase importante pour l'identification des entités pertinentes d'un domaine. Les entités nommées portent une connaissance riche sur le domaine plus que d'autres unités lexicales dans le texte comme elles font référence à des entités du monde.

Les systèmes d'extraction des entités nommées reposent sur diverses techniques. L'article de (Poibeau, 2005) évalue les performances des systèmes reconnaissant les entités nommées. (Poibeau, 2005) considère trois types de systèmes :

– Systèmes fondés sur une base de règles : les règles[13] utilisées dans ces systèmes sont écrites à la main. Elles sont traduites par des patrons

13. Une règle d'extraction s'écrit à l'aide de patrons qui reposent sur des expressions régulières.

constituant des catégories syntaxiques qui caractérisent la forme grammaticale des entités nommées.

- Systèmes à base d'algorithmes d'apprentissage : le système analyse sur un corpus où les entités nommées ont été annotées pour apprendre des règles. Il génère un modèle d'étiquetage du corpus.
- Approche mixte : l'identification des EN est assurée de deux manières. Soit les règles d'extraction sont écrites à la main par un expert et ensuite soumises à un système d'inférence pour la déduction de nouvelles règles. Soit le système apprend des règles automatiquement que l'expert valide ultérieurement.

Le processus d'extraction des systèmes d'extraction des entités nommées qui repose sur des règles d'extraction (écrites sous la forme de prédicats, d'expression rationnelle) couplées avec des dictionnaires se déroule comme suit :

- Prétraitement du corpus suivant un découpage en mots et d'étiquetage des tokens.
- Analyse syntaxique et étiquetage morphosyntaxique en repérant des relations de dépendance (sujet, complément d'objet) entre éléments du texte.
- Analyse sémantique en appliquant des règles d'extraction pour extraire des EN ainsi que leurs types. Il s'agit de reconnaître les noms propres (personnes, lieux, organisations), les expressions temporelles (date, durée) ou les expressions de quantités (monétaires, pourcentages). Ils sont repérés à partir de dictionnaires par exemple.

Il existe des logiciels de reconnaissance d'entités nommées comme par exemple : TagEN [14] (Taggeur d'Entités Nommées), Stanford Named Entity Recognizer (NERClassifier) [15], la chaîne de traitement de la plateforme Gate Annie [16] et l'outil NERD [17].

2.5.2 Bilan

L'inconvénient majeur des outils de REN est qu'ils ne sont pas génériques vu qu'ils reposent généralement sur des patrons d'extraction fortement liées

14. http ://josdblog.blogspot.com/search/label/OutilsTAL
15. http ://nlp.stanford.edu/software/CRF-NER.shtml
16. ANNIE est une chaîne de traitement de la plateforme Gate, http ://gate.ac.uk/
17. http ://nerd.eurecom.fr/

au domaine. L'extraction de l'information possède des limites parmi lesquelles l'aspect polysémique des entités. Il est parfois difficile de catégoriser le nom d'une organisation comme étant un lieu, une communauté ou une personne. Par exemple l'entité nommée « CNRS » peut avoir un type sémantique ORGANISATION (par ex. Le CNRS est un organisme public à caractère scientifique et technologique.), LOCALISATION (par ex. Le débat sur l'innovation ouverte aura lieu dans les locaux du CNRS à Paris.) ou encore PERSONNE (par ex. Inria et le CNRS collaborent deouis de nombreuses années). De même, le nom d'une personne peut référer à un lieu. Par exemple l'entité nommée Roland Garros est relative au type sémantique LOCALISATION dans cette phrase « Nadal a créé l'exploit au premier tour à Roland Garros ». Les contextes où sont mentionnées ces entités nommées peuvent lever l'ambiguïté (une entité nommée précédée par *Mr* indique qu'il s'agit d'une personne).

Malgré ces inconvénients qui sont liés aux caractéristiques de la langue naturelle, les entités nommées sont considérées comme des unités textuelles particulières faisant référence à des entités du monde. Leur détection dans les textes permettent d'identifier des éléments de domaine qui vont servir à la construction d'un modèle formel du domaine. Nous nous intéressons à ce type d'unités textuelles du fait de leur aspect référentiel. Nous utilisons un extracteur d'entités nommées existant et nous filtrons le résultat obtenu suivant les types sémantiques pertinents au domaine en question.

2.6 Extraction des relations

L'acquisition des relations à partir du texte fait appel à la linguistique et au traitement automatique de la langue. L'extraction des relations terminologiques aide à la construction des relations conceptuelles au niveau ontologique. Si l'on considère que les termes sont des mentions linguistiques de futurs concepts, les relations qu'ils entretiennement sont potentiellement pertinentes à exploiter pour la création de relations conceptuelles.

Il existe différents types de relations terminologiques à savoir :

- les relations taxonomiques : l'hyperonymie et la méronymie
- les relations d'équivalence : la synonymie et l'antonymie
- les relations spécialisées qui sont relatives à un domaine spécifique

Les méthodes d'extraction de relations terminologiques sont aussi diverses

suivant les techniques sur lesquelles elle s'appuient. Nous identifions trois grandes familles :

- les approches structurelles qui s'appuient sur la composition syntaxique des termes (Didier & Frérot, 2005). Ces approches exploitent la structure «tête-modifieur » des termes afin d'identifier des relations hiérarchiques (en général, la relation d'hyperonymie). Cette approche ne permet pas non plus de donner un nom à la relation. Prenons par exemple les deux termes tirés du cas d'usage de *American Airlines AAdvantage platinum member* et *member* où *member* forme la tête du terme *AAdvantage platinum* et qui sont reliés par une relation d'hyperonymie.
- les approches contextuelles qui, à l'opposé des approches structurelles, prennent en considération les contextes des termes (Aussenac-Gilles & Séguéla, 2000). Ces approches s'appuient sur des règles qui, classées par type de relation, extraient directement du corpus les arguments de la relation. Les patrons utilisés sont soit généralement basés sur des verbes, soit décrivent la structure interne des termes.

  L'identification des relations entre termes est souvent assurée par des patrons linguistiques. Ces derniers sont construits sur la base des marqueurs identifiés préalablement.
- les approches distributionnelles qui se fondent sur l'exploration des contextes globaux des mots (Faure & Nédellec, 1999). Ces approches classent des mots suivant des mesures de similarité et ne permettent pas de typer les relations. Il s'agit moins de détecter des relations que de construire des classes sémantiques. Les classes sémantiques obtenues peuvent mettre l'accent sur des relations candidates entre les mots qui forment ces classes. Mais ces relations ne sont pas typées. L'ingénieur de la connaissance doit les interpréter en ayant comme information seulement les classes sémantiques formées.

Bilan

Les méthodes d'extraction de relations terminologiques sont soit génériques ou bien spécifiques. Les méthodes génériques génèrent du silence. Elles doivent être adaptées au type de corpus utilisé pour l'extraction des relations terminologiques. Les méthodes spécifiques sont fortement liées au domaine. Les

résultats obtenus sont précis mais avec un faible rappel. Ces méthodes sont généralement difficiles à adapter d'un corpus à un autre.

Ces approches diffèrent par leurs méthodes et le type de corpus qu'elles exploitent. Il est difficile d'évaluer l'apport de chaque méthode d'extraction de relations. Il y a souvent besoin d'adapter et d'ajuster ces approches suivant le type de corpus.

Dans cette thèse, nous ne nous sommes pas intéressés aux méthodes de repérage de relations terminologiques. Pour notre part, nous nous appuyons sur le repérage des verbes dans le corpus d'acquisition afin d'identifier des relations de domaine.

2.7 Conclusion

Dans ce chapitre, nous avons présenté les différents indices textuels qui sont exploités durant le processus de construction d'ontologies à partir de corpus de textes. Ces indices textuels sont extraits en utilisant des outils dédiés à leur identification dans les textes. Durant l'analyse linguistique du processus de construction d'ontologies, les outils de TAL extraient des termes qui sont exploités pour la modélisation sous forme de concepts. Les outils de REN extraient des ENs ainsi que leur type sémantique. Les deux catégories d'information servent à identifier la connaissance du domaine et participent à la modélisation et à l'enrichissement de l'ontologie en spécifiant les concepts, leurs instances et leurs rôles. Intuitivement les ENs forment les instances des concepts et les termes donnent une idée sur la modélisation des concepts. Cette hypothèse apparaît comme un point de vue simplifié mais opératoire. Concernant le choix de modélisation des ENs, les auteurs optent souvent pour une modélisation sous forme d'instances de concepts puisque à la base les ENs renvoient à des objets du modèle formel.

Dans cette thèse, nous exploitons les outils de TAL existants pour l'extraction des termes qui vont servir à la construction des concepts d'une ontologies de domaine. Nous proposons une méthode de filtrage pour réduire le taux de bruit des résultats des outils de TAL. Nous exploitons les outils de REN pour l'identification des entités nommées dans le but d'intégrer ce type d'unité textuelle dans la construction d'ontologies à partir de textes et non pas seulement pour le peuplement d'ontologies. L'identification des relations terminologiques

permet de créer des relations conceptuelles dans une ontologie. Nous n'avons pas cherché à choisir une méthode d'extraction de relations terminologiques. Nous avons exploré comment exploiter le résultat de l'extraction terminologique pour la construction d'ontologies à partir de textes. Nous proposons de combiner termes, entités nommées et relations terminologiques durant le processus de construction d'ontologies. D'autres indices textuels peuvent être utilisés pour la création de concepts comme par exemple les classes sémantiques. Ces indices peuvent être intégrés avec les résultats des outils de TAL et de REN mais nous n'avons pas exploité cette piste dans notre travail.

Le choix de modélisation est fortement lié au domaine où l'ontologie traduit le modèle formel, aux données, aux spécificités du domaine, de l'application cible et de la pratique de l'ingénieur de la connaissance. Plusieurs méthodologies et méthodes de construction d'ontologies à partir de textes combinent différents indices textuels durant l'analyse linguistique dans le but d'identifier à partir de ces éléments linguistiques des connaissances du domaine qui sont formalisées dans une ontologie. Dans le chapitre suivant, nous présentons les différentes méthodologies et méthodes de construction d'ontologies à partir de corpus de textes.

Chapitre 3

Construction d'ontologies à partir de textes

Sommaire

3.1 Introduction

Les méthodologies et méthodes de construction d'ontologies à partir de textes ont pour objectif de réduire l'effort et le temps de développement des ressources ontologiques et d'assurer un passage des textes aux ontologies tout en articulant les termes avec les concepts.

Dans ce chapitre, nous nous intéressons aux méthodologies et méthodes de construction d'ontologies à partir de textes. Nous ne faisons pas une description exhaustive de l'état de l'art (Maedche & Staab, 2001) mais nous nous intéressons aux approches les plus connues dans le domaine de l'ingénierie des connaissances.

Nous essayons, dans ce chapitre, d'exposer, depuis l'introduction de la notion d'ontologies en informatique, les principales méthodologies et méthodes de construction d'ontologies à partir de corpus. Nous mettons l'accent sur l'importance du rôle de l'ingénieur de la connaissance dans ce processus. Pour guider le travail humain de modélisation, les méthodologies (génériques et spécifiques) définissent des procédés permettant à l'ingénieur de la connaissance d'aboutir à la construction d'ontologies reflétant les connaissances d'un domaine précis et répondant aux besoins d'une tâche donnée. Les méthodologies génériques s'intéressent à tout le processus de construction d'ontologies tandis que les méthodologies spécifiques ne se focalisent que sur une partie du processus de construction d'ontologies ou bien elles exploitent des particularités du domaine non transposables à d'autres domaines.

Ce chapitre est organisé en six sections. La première section introduit les différents notions utilisées dans le domaine de l'ingénierie des connaissances en matière de construction d'ontologies. Dans la deuxième section, nous définissons les types d'ontologies à prendre en compte pour la construction d'ontologies. Dans la troisième section, nous dressons un état de l'art sur la construction d'ontologies à partir de textes. La quatrième section présente la manière dont on fait intervenir l'ingénieur de la connaissance dans le pro-

cessus de construction d'ontologies. Dans la cinquième section, nous exposons les principales méthodologies générales et spécifiques qui accompagnent l'ingénieur de la connaissance dans la construction de ressources terminologiques et ontologiques. Avant de conclure, nous décrivons les différents niveaux de connaissance présentés dans la méthode TERMINAE que nous souhaitons enrichir pour améliorer le processus de construction d'ontologies de domaine à partir de textes.

3.2 Quelques notions

Dans cette section, nous définissons d'abord la notion d'ontologie. Ensuite nous expliquons ce qu'est un concept. Enfin, nous exposons les types de relations qu'entretiennent des concepts au sein d'une ontologie.

3.2.1 Ontologie

Les ontologies étaient connues dans le domaine de l'intelligence artificielle (IA) comme un modèle de représentation des connaissances formalisé. Puis elles ont émergé dans presque tous les domaines informatiques nécessitant la manipulation de la connaissance.

L'ontologie était au préalable introduite en philosophie : « la science qui s'occupe de ce qui est, des genres et des structures des objets, des propriétés, des événements, des relations dans tous les secteurs de la réalité» (Smith, 2003). Beaucoup de chercheurs se sont intéressés à cette notion et ont essayé de la définir dans un contexte informatique. Une des définitions les plus utilisée dans la communauté de l'intelligence artificielle est celle de (Gruber, 1993) : « Une ontologie est une spécification explicite d'une conceptualisation». Selon Gruber, la conceptualisation est une abstraction simplifiée du monde que nous souhaitons représenter par rapport à un objectif donné. Cela veut dire qu'une ontologie est une modélisation conceptuelle de la réalité.

(Studer *et al.*, 1998) a essayé de préciser cette définition en présentant l'ontologie comme la : « spécification formelle et explicite d'une conceptualisation partagée», mettant ainsi l'accent sur l'aspect consensuel d'une ontologie. En effet, il est important qu'elle le soit parce qu'une ontologie structure la connaissance du domaine suivant un modèle formel permettant aux utilisateurs et aux

systèmes le partage des connaissances. La conceptualisation d'une ontologie est la construction d'un modèle formel en prenant en considération les aspects liés à l'application visée et en faisant abstraction d'autres aspects.

Dans (Smith, 2001), l'auteur élargit la définition de (Gruber, 1993) à d'autres modèles de représentations de connaissances comme par exemple les glossaires, les thésaurus et les taxonomies. L'interprétation d'une ontologie est liée à une application et n'est pas considérée comme une modélisation du monde (ontologie en philosophie). Une ontologie est composée d'un ensemble de concepts inter-connectés par un ensemble de relations conceptuelles de types taxonomique et associatives. Une ontologie est explicite puisqu'elle définit les concepts, leurs propriétés et leurs relations d'une manière non ambigue dans un langage formel.

Les ontologies apportent une sémantique aux données pouvant être interprétée par les machines et permettent de raisonner et d'inférer de nouvelles connaissances (dans le cas où elles sont formalisées dans un langage de logique). C'est cet aspect qui a suscité l'utilisation des ontologies dans le web sémantique. En effet, la modélisation des connaissances partagées et leur mise à disposition dans le web sémantique a fait que les ontologies sont largement utilisées dans le web. L'ontologie est utilisée pour la recherche d'informations et le traitement automatique de la langue. Elles permettent d'annoter, par exemple, des ressources du web, ce qui permet d'améliorer la qualité des résultats de la recherche d'informations.

3.2.2 Concept

(Uschold & King, 1995) propose une définition du concept qui nous paraît claire : un concept a une intention (sa description), une extension (l'ensemble des instances) et un ensemble d'étiquettes. L'intention d'un concept correspond à la sémantique dénotée par ce dernier. L'extension d'un concept est composé des objets définis par ce concept. Les étiquettes d'un concept, appelées aussi labels, représentent l'ensemble des termes qui permettent de le désigner. (Charlet *et al.*, 2008) définit le concept comme une description d'un objet du monde, qui a une signification propre dans un contexte donné. (Bachimont, 2004) définit un concept suivant trois points de vue. Un concept a une signification qui permet de le distinguer et de le différencier par rapport

aux autres concepts. Un concept est défini dans une ontologie de manière consensuelle à travers des propriétés précises et différentielles permettant de le distinguer des autres concepts dans une ontologie. Un concept n'a de sens que dans une ontologie donnée pour un domaine précis. Si le domaine change, la sémantique du concept change aussi.

3.2.3 Relations entre concepts

Les relations sémantiques définissent des relations binaires entre concepts. Les relations les plus utilisées sont les relations de subsomption (relations taxonomiques), les relations causales, associatives et fonctionnelles. Nous distinguons deux grandes familles de types de relations entre concepts :

- les relations de subsomption « est un » sont définies par (Guarino, 1998) de la manière suivante : « un concept $c1$ subsume un concept $c2$ si toute relation sémantique de $c1$ est aussi une relation sémantique de $c2$ ». Cela signifie que le concept $c2$ est plus spécifique que le concept $c1$.
- les relations associatives se traduisent par des propriétés reliant deux concepts ou un concept à un type de donnée primitif. Elles se traduisent sous la forme de labels grâce auxquels un humain peut avoir une idée de leur sémantique ainsi que d'autres propriétés comme la transitivité et la symétrie.

3.3 Types d'ontologies

Pour construire une ontologie, il est indispensable de préciser son type. La catégorisation des ontologies peut être élaborée sur la base du langage utilisé pour leur modélisation ou par rapport à la nature des connaissances représentées.

3.3.1 Degré de formalisation

Le niveau de spécification formelle des ontologies est un critère qui entre en jeu au moment de la construction d'ontologies Le degré de formalisation dépend du langage utilisé pour décrire une ontologie. Les langages vont du non-formel (par ex. langage naturel) au formel (par ex. OWL). L'évolution

du niveau d'expressivité des langages utilisés donne naissance à une classification des ontologies allant d'une liste de concepts jusqu'à une ontologie avec contraintes. Plus la structure interne d'une ontologie est riche et plus les concepts sont associés à une sémantique formelle plus le niveau d'expressivité est formel, contraint et permet de faire des inférences.

3.3.2 Nature des connaissances représentées

(Guarino, 1998) propose une typologie des ontologies basée sur la nature des connaissances représentées. Nous citons les principaux types :

- Les ontologies primitives qui sont utilisées pour la représentation d'autres connaissances (par ex. l'ontologie qui représente les concepts du langage OWL),
- Les ontologies de niveau supérieur (top-level), appelées aussi ontologies fondationnelles, représentent des concepts considérés comme génériques (les plus abstraits) à plusieurs domaines (par ex. le temps, l'espace) qui sont utilisés comme le haut des racines d'autres ontologies,
- Les ontologies noyaux (core-ontologies) représentent des concepts propres à une discipline (par ex. la médecine),
- Les ontologies de domaine représentent des concepts spécifiques à des domaines particuliers (par ex. pneumologie, industrie automobile) et qui sont une spécialisation des concepts génériques (appartenant aux core-ontologies),
- Les ontologies d'application modélisent des concepts correspondant à une activité spécifique (par ex. achat, test).

Dans cette thèse, nous nous intéressons à l'acquisition d'ontologies de domaine qui décrivent les vocabulaires conceptuels utilisés pour décrire des règles métiers auxquels le projet ONTORULE s'intéresse. L'ontologie construite doit représenter le domaine et les concepts nécessaires pour l'explicitation des règles métier. Cette ontologie présente des concepts de domaine qui possèdent des propriétés soumises à des restrictions de valeur et qui sont reliés par des relations conceptuelles. Dans la suite du mémoire, nous considérons les travaux qui s'intéressent à la construction d'ontologies de domaine. Ces ontologies décrivent un vocabulaire structuré et spécifique à une application particulière. Cela veut dire que les concepts sont associés aux termes liées à un domaine

spécifique.

3.4 Des textes vers des ontologies

La construction d'ontologies à partir de textes est une tâche difficile qui nécessite la mise en place de méthodes et de procédés accompagnant l'ingénieur de la connaissance. (Velardi *et al.*, 2006) décrivent plusieurs méthodologies et méthodes de construction d'ontologies. Ces procédés de conceptualisation sont couplés à des techniques de traitement automatique de la langue (TAL) et d'autres techniques statistiques et d'apprentissage pour la construction de ressources terminologiques et ontologiques à partir de texte. Dans cette section, nous présentons un état de l'art portant sur les principales méthodologies et méthodes de construction d'ontologies à partir de zéro et d'enrichissement d'ontologies existantes.

Depuis le milieu des années 90, beaucoup de travaux se sont intéressés à la définition de méthodes et de techniques permettant d'acquérir et de formaliser les connaissances d'un domaine pour répondre à des besoins de gestion de connaissances partagées dans un système d'information. Initialement, des méthodologies ont été présentées pour guider un travail manuel de construction d'ontologies. Ces méthodologies reposent sur des principes qui tirent leur origine des travaux de développement logiciel. Principalement, nous citons les méthodologies *Methondology* (Fernandez-Lopez *et al.*, 1997), *KACTUS* (Bob *et al.*, 1995) et *On-To-Knowledge* (Gómez-Pérez *et al.*, 2004) (qui sont détaillées dans la section 3.6). Ces méthodologies définissent un cadre général pour la construction d'ontologies et nécessitent la définition de procédés pour leur mise en œuvre. Des outils de construction d'ontologies sont définis pour supporter ces méthodologies comme par exemple les outils *Protégé* (Noy *et al.*, 2000) et *OntoEdit* (Sure *et al.*, 2002) pour la méthodologie *Methondology* et l'outil *OntoStudio* pour la méthodologie *On-To-Knowledge*.

D'autres travaux ont vu le jour suite au progrès technologique atteint dans différents domaines comme le traitement automatique de la langue, la représentation des connaissances et l'apprentissage. Ces approches se distinguent par la nature des procédés qu'elles utilisent pour automatiser la construction d'ontologies. Plus particulièrement, ces approches s'intéressent à la détection des connaissances du domaine et à leur structuration. Nous exposons

une typologie des approches de construction de ressources terminologiques et ontologiques à partir de textes suivant les techniques utilisées pour cette construction.

3.4.1 Approches fondées sur l'analyse terminologique

L'acquisition de la terminologie liée à un domaine permet d'identifier les connaissances du domaine à modéliser. Les méthodologies et méthodes de construction d'ontologies à partir de textes qui sont fondées sur l'analyse terminologique du texte s'appuient sur un pré traitement du texte pour la détection des termes candidats. Il s'agit généralement d'une segmentation en phrases du texte, d'un étiquetage morpho-syntaxique des mots constituant chaque phrase et d'une analyse des dépendances syntaxiques des mots constituant une phrase.

Souvent, l'analyse terminologique du texte s'accompagne d'une analyse statistique qui peuvent extraire les termes sur la base de leur fréquence ou d'autres mesures statistiques (*i.e.* tf-idf, Dice) afin de réduire le bruit généré par les outils de TAL [1]. Plusieurs travaux se fondent sur ce principe. Nous citons principalement les travaux de (Cimiano & Völker, 2005), (Alessandro *et al.*, 2007) et (Aussenac-Gilles *et al.*, 2008).

T2K (Alessandro *et al.*, 2007) est une méthodologie hybride de construction d'ontologies de domaine à partir de textes réglementaires qui se fonde sur une analyse terminologique du texte à la fois linguistique et statistique. Cette méthodologie est composée de deux étapes essentielles : l'extraction de la terminologie de domaine et la structuration des termes candidats. L'extraction des termes candidats s'appuie sur un texte étiqueté syntaxiquement au préalable. Les termes extraient sont des termes simples (composés d'un seul mot) et des termes composés (des syntagmes nominaux composés de tête et de modifieur) et ils sont extraient de deux manières différentes. L'extraction des termes simples s'appuie sur leur fréquence dans le texte. L'extraction des termes composés s'appuie sur l'application des patrons syntaxiques sur le texte. Ces patrons couvrent toutes les variations de modifieurs apparaissant dans les termes composés. Le résultat de l'extraction terminologique est une

1. Le bruit étant le nombre de termes incorrects divisé par le nombre total des termes extraits.

liste de termes candidats simples qui sont ordonnés suivant leur fréquence dans le texte et une liste de termes candidats composés qui sont ordonnés suivant leur score d'association. *T2K* propose deux manières pour structurer des termes candidats : construction d'une taxonomie ou de classes sémantiques. Dans la première approche, les termes qui partagent la même tête sont structurés sous la forme d'une taxonomie telle que les têtes d'un terme candidat (*i.e.* environment) hyponymes (*i.e.* urban environment, marine environment). De la deuxième méthode, *T2K* assure la création de classes lexico-sémantiques appelées « structures proto-conceptuelles ». Le regroupement de ces classes est assuré par l'exploitation d'une approche distributionnelle. Les termes sont regroupés dans une même classe et considérés comme similaires s'ils sont statistiquement permutables dans un même contexte syntaxique.

La méthode *OntoLearn* (Navigli *et al.*, 2004) se fonde sur une analyse terminologique du texte pour la détection des termes du domaine. Cette méthode propose de suivre les démarches ci-après :

- Extraction des termes à partir d'un corpus textuel
- Interprétation sémantique
- Création d'une ontologie de domaine

3.4.2 Approches fondées sur l'extraction à base de patrons

Cette famille d'approches s'appuie sur l'identification des relations pour la création des concepts. L'identification des relations sémantiques est assurée par l'exploitation des patrons qui peuvent être définis soit manuellement ou en utilisant des techniques d'apprentissage. Plusieurs travaux se sont intéressés à l'identification des relations sémantiques spécialisées.

La première famille d'approches pour l'identification des relations associatives repose sur la définition de patrons linguistiques (par ex. (Aussenac-Gilles & Séguéla, 2000)) appliqués sur un corpus d'acquisition. L'ensemble des instances d'un triplet (t_i, r, t_j) identifiées dans le texte permettent de définir une relation générique R pour toutes ces instances. Ces patrons linguistiques décrivent des relations prédéfinies. Les travaux les plus connus fondés sur l'extraction à base de patrons sont (Malaisé, 2005) et (Drymonas *et al.*, 2010). *OLE* (Smrz & Novácek, 2006; Novácek, 2012) est une méthode d'apprentis-

sage non supervisé de patrons de connaissance à partir d'une grande masse de données. Elle se fonde sur une extraction de relations terminologiques. Les relations sont extraites par application de patrons spécifiques sur le texte. Les patrons peuvent être appris automatiquement à partir du corpus d'acquisition. *OLE* détecte essentiellement des relations taxonomiques (« est un ») ce qui permet la création d'une taxonomie noyau. La méthodologie *TextOntoEx* (Dahab *et al.*, 2008) s'appuie sur une méthode à base de patrons pour l'extraction des relations associatives. Ces patrons permettent de retrouver dans le texte des instances de relations préalablement identifiées et décrites à travers ces patrons. Les patrons utilisées par la méthodologie sont des représentations générales de fragments de texte telles que les arguments d'un patron donné sont annotés par des types ontologiques ou linguistiques (*i.e.* <Fleure Plante> <Possède Verbe> <Couleur>). La particularité de l'utilisation de patrons sémantiques est la possibilité de remplacer les arguments de ce dernier par d'autres éléments appartenant à la même classe (*i.e.* les concepts fils du concept Plante ou les verbes décrivant le même sens dans WordNet) afin d'enrichir l'extraction des relations à partir du texte.

La deuxième manière consiste en l'application de règles d'association permettant la détection des triplets les plus fréquents avec une meilleure confiance. Dans ce type d'extraction, les relations ne sont pas définies au préalable (*i.e.* (Maedche & Staab, 2001)). Dans *OntoGain* (Drymonas *et al.*, 2010), les auteurs identifient des relations associatives à partir de l'application des règles génériques d'association (Srikant & Agrawal, 1995). Les règles associatives sont sous la forme de triplets (par ex. « sujet, verbe, objet »).

3.4.3 Approches fondées sur la création de groupes de concepts

L'approche distributionnelle, de manière générale, est héritée de (Harris, 1954) et vise à extraire des concepts à partir de l'analyse des mots et de leur distribution dans le texte. Elle repose sur l'hypothèse que les mots similaires apparaissent dans des contextes similaires et elle a été proposée au départ sur des corpus de domaines spécialisés sur lesquels elle est supposée être plus fiable. Les classes sémantiques sont créées à partir des groupements de mots ayant des distributions proches. L'ontologie est construite à partir de

ces classes sémantiques qui forment des concepts candidats.

L'analyse formelle des concepts (AFC) se fonde sur le même principe que les approches distributionnelles. Elle repose sur le regroupement d'objets partageant les mêmes attributs binaires dans un même contexte formel et les structure sous la forme d'un treillis de concepts. Un contexte formel est défini par un triplet (O, A, R) tel que O est un ensemble d'objets, A est un ensemble d'attributs d'objets et R est un ensemble de relations binaires où $(o, a) \in R, o \in O, a \in A$ qui pose que l'objet o possède l'attribut a. Un treillis de concepts est un ensemble ordonné $(C, \preceq)$ où C est un ensemble de concepts et $\preceq$ est une relation d'ordre sur C. Le treillis possède un concept qui subsume tous les autres concepts ($\top$) et un autre qui est subsumé par tous les autres concepts ($\bot$).

Les approches qui utilisent l'analyse formelle de concepts sont principalement (Cimiano *et al.*, 2011) et (Bendaoud *et al.*, 2008). *Text2Onto* (Cimiano *et al.*, 2011) vise à construire automatiquement une taxonomie de concepts à partir d'un corpus de textes en s'appuyant sur l'AFC. Le contexte formel des termes est représenté par un vecteur décrivant les dépendances syntaxiques. L'analyseur syntaxique extrait, pour chaque phrase du texte, un arbre de dépendances syntaxiques dans le but de détecter les dépendances syntaxiques entre des verbes et les têtes de leurs sujets, compléments d'objets et prépositions. Une fois que ces couples sont extraits par l'analyseur syntaxique, ils sont remplacés par leurs formes lemmatisées. Les couples <verbe, argument> sont filtrés suivant un seuil de confiance décrit par une mesure statistique dans (Cimiano *et al.*, 2004). Seuls les couples dont le poids est supérieur ou égal au seuil de confiance sont transformés en contextes formels pour l'application de l'AFC. Puis, le treillis de concepts est transformé en taxonomie de concepts.

Dans la méthodologie *Pactole* (Bendaoud *et al.*, 2008), l'ontologie est construite en plusieurs étapes. La première étape est la construction du noyau de l'ontologie sous la forme d'une taxonomie. L'AFC est une technique utilisée pour la classification des objets avec les propriétés qu'ils partagent pour la construction d'un treillis de concepts. Les termes sont regroupés dans des classes suivant leur présence dans des syntagmes syntaxiques avec le même groupe de verbes (inspiré de (Faure & Nédellec, 1999)) afin d'extraire les termes qui sont le sujet du même groupe de verbes et le complément d'un autre groupe de verbes.

OntoGain (Drymonas *et al.*, 2010) cherche à construire des ontologies formelles à partir de corpus textuel à travers l'exploitation des outils de TAL, d'approche distributionnelle et de l'AFC. Les termes similaires forment des concepts formels ayant comme attribut les verbes détectés dans leurs contextes. Les auteurs appliquent l'AFC sur l'ensemble des concepts formels pour la création d'un treillis de concepts.

3.4.4 Bilan

Cette réflexion autour des méthodologies et méthodes de construction d'ontologies à partir de corpus de textes a connu peu d'évolution. Pour nous, cela tient au fait que la construction d'ontologies ne peut pas être totalement automatisable. Il n'existe, à ce jour, aucune méthodologie ou méthode qui bénéficie d'un consensus général dans le domaine de l'ingénierie des connaissances. Le choix d'une méthodologie ou d'une méthode de construction de ressources terminologiques et ontologiques dépend du type d'ontologies à créer, des objectifs de leur construction et d'autres besoins relatifs aux caractéristiques d'une méthode.

Le choix d'une méthode dépend généralement de plusieurs critères. Par exemple, une méthode est choisie si elle propose l'utilisation de ressources existantes pour un domaine précis. En effet, il existe des méthodes qui permettent d'enrichir des ontologies existantes et d'autres qui utilisent des ressources sémantiques pour l'analyse terminologique comme par exemple l'utilisation de *WordNet*. Une méthode est choisie aussi suivant le niveau de formalisation visé. En effet, suivant l'objectif de l'utilisation d'ontologies, leur construction peut aboutir à la création d'un thésaurus décrivant un vocabulaire structuré du domaine ou jusqu'à la création d'une ontologie formelle pour pouvoir exécuter des raisonnements dessus. Le choix d'une méthode peut être lié, aussi, au fait qu'elle définit des procédés pour assurer l'évolution des ontologies et leur maintenance. Enfin, une méthode est choisie suivant le rôle qu'occupe l'ingénieur de la connaissance durant les étapes de construction d'ontologies pour faire des choix de paramétrage de l'outil, de sélection ou de validation. Dans la section suivante, nous décrivons les différents niveaux de construction d'ontologies auxquels l'ingénieur de la connaissance peut intervenir.

3.5 L'expertise humaine dans la construction d'ontologies

Les méthodes de construction d'ontologies différent par les techniques utilisées pour la construction d'ontologies mais aussi par le degré d'automatisation du processus de construction. Suivant les méthodes, l'ingénieur de la connaissance est sollicité durant toutes les étapes de la méthode ou dans quelques unes. Dans cette section, nous décrivons ces deux types d'approches.

3.5.1 Processus de construction d'ontologies : automatique ou interactif

Il y a deux grandes familles d'approches pour la construction d'ontologies à partir de textes : automatiques ou semi-automatiques.

Dans la première famille d'approches, les ontologies à créer doivent être validées par l'ingénieur de la connaissance (Noy & Mcguinness, 2001). Prenons comme exemple l'outil *OntoGain* (Drymonas *et al.*, 2010) qui cherche à construire des ontologies formelles à partir de corpus textuel. L'ingénieur de la connaissance intervient à la fin du processus de construction d'ontologies pour valider les concepts et les relations créées. La méthode *GAleOn* (Manzano-Macho *et al.*, 2008) combine plusieurs techniques de construction d'ontologies pour la construction ou l'enrichissement d'une ontologie de domaine de manière à réduire l'intervention d'un ingénieur de la connaissance. Mais le résultat obtenu doit être validé par l'ingénieur de la connaissance.

La deuxième famille d'approches repose sur un processus semi-automatique. La plupart des méthodes optent pour la réutilisation et l'enrichissement d'ontologies existantes. Ce second type d'approches a l'inconvénient de consommer du temps mais garantit que les ontologies créées correspondent aux objectifs de leur utilisation. La méthode *DODDLE* **??** permet l'adaptation d'une ontologie de domaine en l'enrichissant avec de nouveaux concepts correspondant à des termes extraits à partir d'un document spécialisé par des outils de TAL. Cette méthode propose une construction interactive comportant trois phases principales : une phase de spécification d'une ontologie existante à enrichir et un document de domaine, une phase d'identification des concepts centraux et une phase d'amélioration de l'ontologie créée. C'est

à l'ingénieur de la connaissance de sélectionner les termes qui lui semblent pertinents en s'appuyant sur des mesures de fréquence et de tf-idf. Ensuite, *DODDLE* propose une liste des concepts correspondant à chacun des termes sélectionnés sur la base d'une correspondance entre les labels des concepts et les termes. L'ingénieur de la connaissance décide de garder ou pas les concepts trouvés, ajoute de nouveaux concepts représentant les termes sélectionnés et crée les propriétés et les relations associatives des concepts créés.

Dans la méthode TERMINAE (Aussenac-Gilles *et al.*, 2008), l'ingénieur de la connaissance intervient dans toutes les étapes de la construction d'ontologies. Il choisit un corpus d'acquisition qui décrit les connaissances du domaine à modéliser. Il filtre et valide les termes qui lui semblent pertinents à normaliser. C'est à l'ingénieur de la connaissance de créer les concepts et les relations conceptuelles à partir d'unités terminologiques validées.

TextOntoEx (Dahab *et al.*, 2008) est une méthodologie de construction d'ontologies de domaine à partir de textes s'appuyant sur une méthode à base de patrons pour l'extraction des relations associatives. Ces patrons permettent de retrouver dans le texte des instances de relations préalablement identifiées et décrites à travers ces patrons.

3.5.2 Rôle de l'ingénieur de la connaissance dans le processus de construction d'ontologies

La plupart des méthodes et outils de construction d'ontologies à partir de textes sollicitent l'intervention de l'ingénieur de la connaissance. Dans cette section, nous identifions les différents niveaux du processus de construction d'ontologies durant lesquels l'ingénieur de la connaissance peut intervenir. Nous citons quelques exemples de méthodes qui définissent le rôle que joue l'ingénieur de la connaissance à ces différents niveaux.

Choix des ressources

Dans certaines méthodes, l'ingénieur de la connaissance constitue un corpus d'acquisition à partie de documents afin de s'assurer que ces documents couvrent les notions pertinentes au domaine à modéliser. De plus, dans certains cas, sélectionne des ressources sémantiques (*i.e.* des patrons sémantiques,

un thésaurus ou une ontologie de domaine) pour amorcer la construction d'ontologies. Par exemple, la méthodologie *TextOntoEx* (Dahab *et al.*, 2008) suggère à l'ingénieur de la connaissance de définir ses propres patrons sémantiques et de choisir un document d'acquisition pour créer des relations conceptuelles d'une ontologie.

Validation des informations lexicales

Les outils de TAL extraient des connaissances pertinentes au domaine mais aussi du bruit. L'ingénieur de la connaissance doit filtrer et sélectionner les éléments linguistiques qui seront conceptualisés. Plusieurs méthodes semi-automatiques sollicitent l'intervention de l'ingénieur de la connaissance pour la validation d'une liste de termes candidats ou de relations terminologiques afin de préparer les données qui seront utilisées pour la construction de concepts et de relations conceptuelles.

La méthode *GAleOn* (Manzano-Macho *et al.*, 2008) combine plusieurs techniques de construction d'ontologies. Durant l'extraction terminologique, chacune des informations relatives à un élément candidat possède un degré de confiance. Parfois, pour un même élément candidat, deux informations extraites par deux techniques différentes peuvent être contradictoires. Les auteurs citent un exemple d'élément qui peut être à la fois un concept fils et un concept parent pour un même autre élément candidat. Dans ce cas, c'est à l'ingénieur de la connaissance de filtrer et de choisir les informations correctes pour cet élément.

La méthodologie *T2K* (Alessandro *et al.*, 2007) est composée de deux étapes essentielles : l'extraction de la terminologie de domaine et la structuration des termes candidats. L'ingénieur de la connaissance valide les termes extraits dans la première étape pour ensuite créer automatiquement une taxonomie de termes validés. Dans la méthode TERMINAE (Aussenac-Gilles *et al.*, 2008), l'ingénieur de la connaissance filtre et valide les termes candidats extraits par un outil de TAL en s'appuyant sur leur fréquence et leurs occurrences dans le corpus d'acquisition pour détecter ceux qui sont pertinents à conceptualiser.

L'outil *OntoLP* (Lopes *et al.*, 2010) est un outil de construction d'ontologies à partir de textes en s'appuyant sur une extraction terminologique

automatique. L'outil crée des classes sémantiques décrivant des termes synonymes. C'est à l'ingénieur de la connaissance de sélectionner les termes qui sont les plus représentatifs de ces classes. La méthode *OntoLearn* (Navigli *et al.*, 2004) se fonde sur une analyse terminologique du texte pour la détection des termes du domaine. Dans cette méthode, les relations entre les composants d'un terme composé sont étudiées et validées par l'ingénieur de la connaissance pour la construction d'une taxonomie de domaine qui sera intégrée dans une ontologie générique. Durant l'étape d'annotation des classes sémantiques, la méthode *OLE* (Smrz & Novácek, 2006; Novácek, 2012) propose à l'ingénieur de la connaissance d'utiliser la ressource *WordNet* pour l'annotation de ces classes. Dans *OntoLearn* (Navigli *et al.*, 2004), les relations entre les composants d'un terme composé sont étudiées et validées par l'ingénieur de la connaissance.

Choix de modélisation

Les unités terminologiques extraites par des outils de TAL peuvent être ambigües et ainsi dénoter plusieurs sens qui ne sont pas tous nécessairement pertinents à modéliser. Il faut donc sélectionner les termes qui sont relatifs au domaine et regrouper ceux qui sont synonymes. La formalisation de ces unités terminologiques en des unités conceptuelles nécessite d'abord de décontextualiser ces dernières afin de normaliser leurs sens (Bachimont, 2000). De plus, au niveau terminologique, on ne peut pas faire de différence entre les unités dénotant des concepts, des relations conceptuelles et celles décrivant des propriétés de concepts. L'ingénieur de la connaissance doit donc faire des choix de modélisation que le processus de construction d'ontologies ne peut pas faire automatiquement.

OntoGen (Blaz *et al.*, 2006) est une méthodologie de construction d'ontologies de thèmes (topic ontologies) à partir d'un corpus textuel. Elle exploite des techniques de fouille de données et procède de manière descendante. L'ingénieur de la connaissance choisit à chaque fois un thème (une classe), *OntoGen* lui propose automatiquement des sous-thèmes (des sous-classes). *OntoGen* propose aussi d'annoter les classes sélectionnées par l'ingénieur de la connaissance par extraction de mots clés (en appliquant la technique d'apprentissage SVM). Dans la méthodologie *Pactole* (Bendaoud *et al.*, 2008), l'ingénieur de la

connaissance étiquette les concepts à partir de l'étude des propriétés partagées par ses instances.

Validation des ressources sémantiques

Dans les méthodes automatiques, l'ingénieur de la connaissance intervient à la fin du processus de construction pour la validation et l'enrichissement de l'ontologie créée. Par exemple, il affine la structure arborescente d'une ontologie en ajoutant de nouveaux concepts plus spécifiques que d'autres. Il définit aussi des relations non taxonomiques entre ces concepts.

Dans la méthode *Text2Onto* (Cimiano *et al.*, 2011), l'ingénieur de la connaissance étudie le treillis de concepts créé en validant les concepts formels créés. Dans la méthode *Pactole* (Bendaoud *et al.*, 2008), l'ingénieur de la connaissance valide les relations associatives créées par application de l'ARC entre les concepts formels. Dans la méthode *OntoGain* (Drymonas *et al.*, 2010), l'ingénieur de la connaissance valide l'ensemble des concepts formels qui forment le treillis construit.

3.6 Des méthodologies pour guider le travail humain

Nous avons vu dans la section précédente que le rôle de l'ingénieur de la connaissance est important et que le travail de la construction d'ontologies ne peut pas être totalement automatique.

La définition d'une méthodologie permet d'intégrer les choix de sélection, de validation et de modélisation pris par l'ingénieur de la connaissance pour la construction de ressources terminologiques et ontologiques. Il existe deux familles de méthodologies. La première famille regroupe des méthodologies générales qui touchent à toutes les étapes de construction d'ontologies. La deuxième famille décrit des méthodologies qui ne s'intéressent qu'à une phase (la conceptualisation) précise du processus de construction d'ontologies.

3.6.1 Les méthodologies générales

La méthodologie de Uschold et King (Uschold & King, 1995) a été la première méthodologie proposée dans le domaine de l 'ingénierie des connaissances et constitue la base d'autres méthodologies. Elle définit un guide de construction d'ontologies. La méthodologie *KACTUS* (Bob *et al.*, 1995) (modelling Knowledge About Complex Technical systems for multiple USe) propose des procédés pour guider la réutilisation des ontologies existantes. Il existe d'autres méthodologies générales, la plus connue est *Methondology* (Fernandez-Lopez *et al.*, 1997; Gómez-Pérez *et al.*, 2007). Cette méthodologie s'inspire des travaux faits dans la gestion de projets. Il existe des outils qui permettent de construire des ontologies en utilisant la méthodologie *Methondology* comme par exemple l'outil *OntoEdit* (Sure *et al.*, 2002). *Methondology*définit un ensemble de procédés pour chacune des étapes suivantes :

- la spécification de la construction d'ontologies en termes d'objectif et de futurs utilisateurs,
- l'acquisition des connaissances à travers plusieurs techniques comme par exemple les textes ou les interviews d'experts,
- la structuration des connaissances en un modèle conceptuel,
- l'enrichissement du modèle conceptuel par l'intégration d'ontologies existantes pour structurer les concepts du domaine ou les spécifier,
- la formalisation du modèle conceptuel en un modèle formel à travers un langage formel,
- l'évaluation d'ontologie, notamment pour vérifier si l'ontologie répond aux spécifications.

La méthodologie *On-To-Knowledge* (Gómez-Pérez *et al.*, 2004) est plus générale que la méthodologie *Methondology*. Elle permet l'enrichissement d'ontologies et leur spécialisation pour répondre à des tâches et des applications. Elle définit la construction d'ontologies en quatre phases : une phase d'étude de faisabilité, une phase d'enrichissement, une phase d'évaluation et une phase de maintenance. L'outil *OntoStudio* se fonde sur la méthodologie *On-To-Knowledge* pour la spécialisation d'ontologies.

3.6.2 Les méthodologies spécialisées

Une autre famille d'approches plus précises se focalise sur l'étape de formalisation d'ontologies. L'étape de formalisation est une étape durant laquelle l'ingénieur de la connaissance traduit d'une manière formelle les entités pertinentes du domaine extraites à partir d'un corpus.

La méthodologie *OntoSpec* est une méthodologie de construction d'ontologies formelles (Kassel, 2002, 2009). La création d'ontologies formelles repose sur un ensemble de primitives et de principes de modélisation. *OntoSpec* vise à utiliser des termes définis dans un langage naturel et produit des concepts définis par des caractéristiques appelées des méta-propriétés. Les concepts sont définis par des propriétés nécessaires et suffisantes. L'ontologie semi-informelle peut être ensuite représentée dans un langage formel choisi par l'ingénieur de la connaissance.

D'autres approches de construction d'ontologies formelles se fondent sur la notion de patrons conceptuels (connus sous le nom de « Ontology Design Patterns» en anglais) pour assurer une formalisation du modèle conceptuel. Les patrons conceptuels tirent leur origine des travaux de la conception logicielle. L'objectif de l'utilisation de patrons conceptuels est de promouvoir des bonnes pratiques de conception, en capturant les expériences existantes qui fournissent des solutions à des problèmes récurrents de modélisation. L'idée de l'utilisation des patrons de conception (ODPs) est apparue pour éviter de construire une ontologie à partir de zéro et encourager la réutilisation des « bouts » d'ontologies existants appelés « patrons conceptuels ». Ces approches proposent d'outiller l'étape de formalisation d'ontologies en définissant des patrons conceptuels afin d'accompagner l'ingénieur de la connaissance durant la construction d'ontologies formelles.

Les patrons conceptuels[2] sont définis comme des blocs conceptuels (des classes et leurs propriétés). Ils sont caractérisés par :

- un langage de représentation (souvent ils sont représentés en OWL),
- une taille modeste (de 2 à 10 classes ainsi que les relations définies entre elles),
- la possibilité de définir : des classes disjointes, une propriété possédant

2. Un portail présente des ODPs se trouve sur le lien suivant : www.ontologydesignpatterns.org.

un domaine et un range ou une restriction de rôle,
- ils doivent modéliser une solution à un problème de conception tout en ayant une forme compacte et pertinente.

Les patrons conceptuels peuvent être modélisés à partir de différentes ressources sémantiques, comme par exemple des core ontologies (des ontologies de haut niveau), des thesaurus ou bien à partir d'une spécialisation d'autres ODPs. Les patrons conceptuels sont relatifs à des « cas d'usage» de taille modeste. Un patron conceptuel peut être relatif à plusieurs cas d'usage et inversement.

Les principaux travaux qui s'intéressent à définir une méthodologie pour accompagner l'ingénieur de la connaissance durant l'étape de formalisation sont (Cea *et al.*, 2008; Presutti, 2008; Gangemi & Presutti, 2009) et (Fuchs *et al.*, 2010). Un outil qui supporte cette méthodologie est *S.O.S* (System for Ontology design pattern Support) [3]. *S.O.S* propose les étapes suivantes :

- la spécification des contraintes du domaine par l'ingénieur de la connaissance sous la forme de définitions en langage naturel. L'ingénieur de la connaissance rédige des phrases tout en employant des verbes au présent simple. Une telle phrase est considérée comme l'instance d'un patron lexico-syntaxique (LSP) (ou de plusieurs en cas d'ambiguïté). Un patron lexico-syntaxique représente une relation terminologique entre deux termes ou une propriété du terme considéré dans la phrase le décrivant,
- la recherche de l'ensemble des LSPs qui correspondent à cette spécification. Dans le cas où le système ne trouve aucun patron lexico-syntaxique qui correspond à cette spécification, il propose à l'ingénieur de la connaissance un système de questions/réponses afin de détailler sa spécialisation,
- la proposition d'un ensemble de ODPs correspondant à chacun des LSPs identifiés. Les LSPs sont reliés aux ODPs qui sont leur réalisation formelle en OWL. Un LSP peut donc être attribué à un ODP, à n ODPs ou bien à une combinaison de n ODPs.
- l'ingénieur de la connaissance combine plusieurs ODPs ou spécialise un ODP existant pour construire une ontologie.

3. C'est un outil qui s'intègre comme plug-in dans la plateforme *NeonToolkit*

LSP Identifier : LSP -SC-Di-EC- ES	
NeOn ODPs Identifier : LP-SC-01+LP-Di-01+LP-EC-01	
Formalization	
1	Los/las NP<superclass> se clasifican en|como | se dividen en [CD] [los/las siguientes] [CN] [PARA] [(NP<subclass>,)* and] NP<subclass>
2	Se distinguen CD CN de NP<superclass> : [(NP<subclass>,)* y] NP<subclass> CD CN de NP<superclass> se distinguen : [(NP<subclass>,)* y] NP<subclass>
Examples	
1	*Los hongos se clasifican en cuatro grandes grupos: Ficomicetos, Ascomicetos, Basidiomicetos y Deuteromicetos.* *Las grasas se dividen en saturadas e insaturadas.*
2	*Se distinguen dos tipos de tilacoides: los tilacoides de las granas y los tilacoides del estroma.*

FIGURE 3.1 – Un patron lexico-syntaxique correspondant à une relation de subsomption (Cea *et al.*, 2008).

3.6.3 Bilan

Les méthodologies générales définissent un cadre général pour la construction d'ontologies. Notre objectif est de définir une méthodologie qui supporte plus spécifiquement la conceptualisation des ontologies.

L'intérêt de la méthodologie qui s'appuie sur les patrons conceptuels est de permettre à des utilisateurs « novices » de définir une ontologie formelle grâce à l'utilisation de patrons conceptuels existants. L'inconvénient majeur de cette technique est que l'étude de nouveaux patrons est coûteuse en temps et en effort. Un autre inconvénient de l'utilisation de patrons conceptuels pour la construction d'ontologies est que l'outil qui supporte la méthodologie ne propose pas d'aide à l'ingénieur de la connaissance pour ce qui concerne le choix des patrons conceptuels à utiliser.

Cette famille d'approches s'intéresse à l'étape de formalisation d'ontologies à partir de la spécialisation des patrons conceptuels. Or, nous nous intéressons à l'exploitation des textes pour la construction d'ontologies (*cf.* indices). La réflexion autour de l'acquisition des connaissances à partir de textes a été initiée par le groupe TIA (Terminologie Inteligence Artificielle) en France. Beaucoup d'approches s'appuient sur les textes comme source d'acquisition de connaissances parmi les quelles nous citons la méthode TERMINAE (Aussenac-Gilles *et al.*, 2008). Dans la section suivante, nous décrivons la méthode TERMINAE et nous justifions notre choix de définir une méthodologie qui se fonde sur cette méthode.

3.7 La méthode TERMINAE

Aucune des méthodes citées précédemment ne permet de construire automatiquement une ontologie à partir de textes. Or, nous avons vu que l'intervention humaine est nécessaire : l'ingénieur de la connaissance extrait les connaissances pertinentes des textes et construit un modèle du domaine adapté à l'application visée. Il faut donc, là-aussi, proposer une méthodologie pour guider le travail humain.

La méthode TERMINAE permet à l'ingénieur de la connaissance de s'appuyer sur le matériau linguistique pour construire un tel modèle. Elle propose une approche terminologique pour la construction d'ontologies de domaine à partir de corpus d'acquisition et décompose le processus d'acquisition en trois niveaux : terminologique, termino-conceptuel et conceptuel (voir figure 3.2). La transition du texte à l'ontologie n'est pas automatique, c'est pourquoi l'acquisition est un processus interactif dans la méthode TERMINAE. Dans cette section, nous décrivons chacun de ces niveaux ainsi que les enjeux qui en ressortent.

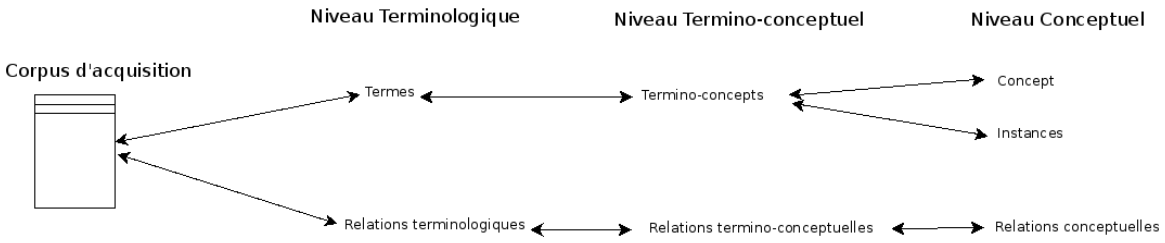

FIGURE 3.2 – Les trois niveaux de la méthode TERMINAE.

3.7.1 Le niveau terminologique

La première étape du processus de construction d'ontologie est une étape d'analyse terminologique qui permet d'étudier des termes candidats extraits automatiquement par des outils de TAL à partir d'un corpus d'acquisition. Le résultat des outils de TAL étant généralement bruité, l'ingénieur de la connaissance doit sélectionner les termes qui lui paraissent les plus pertinents. Dans l'outil TERMINAE, les résultats de l'analyse terminologique se visualisent sous la forme d'une liste de termes candidats ordonnée par fréquence

ou par ordre alphabétique. La figure 3.3 décrit un extrait de cette liste tirée d'un cas d'usage *American Airlines*. Nous présentons dans le chapitre 6 le cas d'usage correspondant qui décrit les règles et conditions d'attribution de « miles » pour des voyageurs.

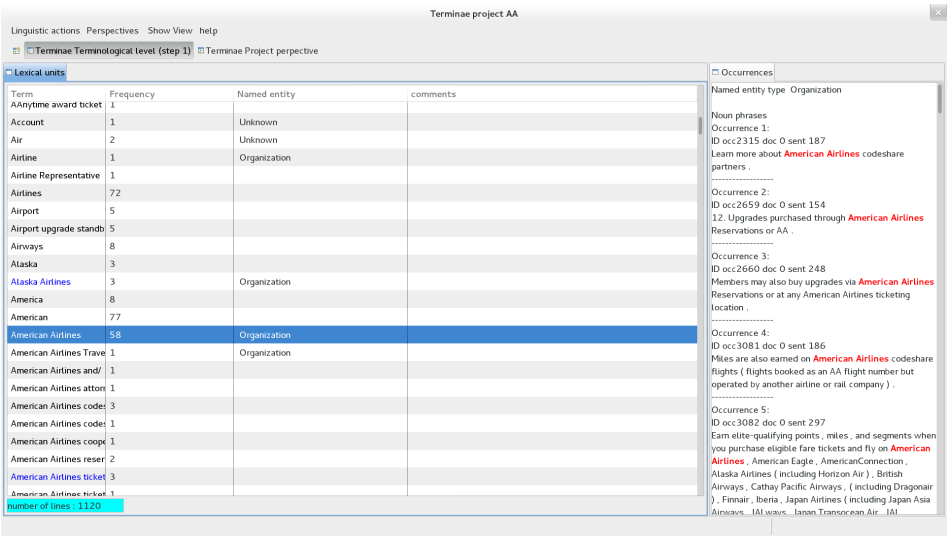

FIGURE 3.3 – Un extrait de la liste de termes candidats relatifs au cas d'usage de American Airlines.

L'ingénieur de la connaissance doit filtrer la liste obtenue et regrouper certaines unités. Parmi les unités retenues, l'ingénieur de la connaissance peut ensuite identifier des unités synonymes (qui décrivent un même sens) et les regrouper sous une forme canonique. Il peut s'agir aussi de regrouper des termes qui sont des variantes d'autres termes (*i.e.* une variante morphologique). Les termes extraits peuvent être mal formés car les documents du domaine contiennent des particularités orthographiques, syntaxiques et lexicales que les outils de TAL génériques ne peuvent prendre en considération. D'autres termes ne sont pas pertinents pour le domaine et ne seront pas pris en considération pour la construction du modèle conceptuel. Durant l'étape terminologique, l'analyse de la liste obtenue n'est généralement pas exhaustive, mais à l'issue de cette étape, la liste obtenue contient les termes validés et considérés comme pertinents par l'ingénieur de la connaissance.

Le travail précis d'analyse des termes puis des termino-concepts se fait dans l'outil TERMINAE à l'aide de fiches terminologiques qui regroupent dans une même vue toutes les informations obtenues automatiquement ou manuel-

lement pour un terme et les termino-concepts qui lui sont associés. La fiche terminologique (voir figure 3.4) décrit toutes les caractéristiques d'un terme et l'associe à un ou plusieurs termino-concepts suivant que le terme décrit un ou plusieurs sens pertinents. Pour le terme *airline participant* sélectionné dans la liste affichée à gauche, la fiche terminologique de droite indique si l'élément sélectionné est un terme, une entité nommée ou bien les deux (onglet « Lexical information »). L'ingénieur de la connaissance intervient à ce stade pour valider la fiche créée : en parcourant la liste des occurrences de l'unité linguistique en question (onglet « Occurrences »), il peut vérifier si elle possède un ou plusieurs sens, choisir le(s) pertinent(s) voire en ajouter.

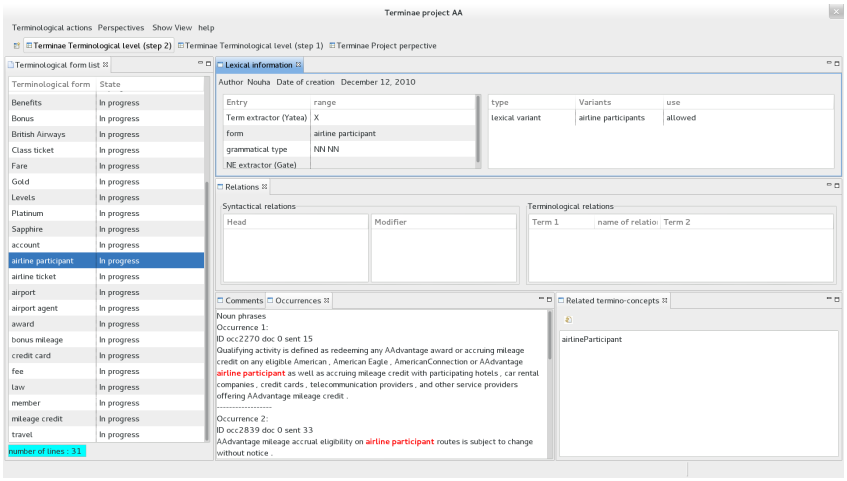

FIGURE 3.4 – La fiche terminologique du terme *airline participant*.

3.7.2 Le niveau termino-conceptuel

La deuxième étape de la méthode TERMINAE est la création de termino-concepts. Un termino-concept est un terme désambiguïsé dont le sens est défini par son usage dans le corpus. La liste des unités lexicales pertinentes est transformée en un réseau de termino-concepts structuré par des liens taxonomiques. Durant cette étape, l'ingénieur de la connaissance commence à normaliser le réseau sémantique en vue de la création d'un modèle du domaine plus proche du conceptuel que du linguistique (modèle semi-formel). Ce processus de normalisation sémantique permet à nouveau de regrouper les termes synonymes en les associant à un unique termino-concept et de désambiguïser

les termes ambigus car l'ingénieur de la connaissance crée un termino-concept pour chaque sens pertinent de terme.

Le résultat de la phase termino-conceptuelle est un réseau terminologique normalisé qui décrit le domaine considéré et s'apparente à un thésaurus. Le modèle obtenu peut être exporté à partir de l'outil TERMINAE sous la forme d'un fichier SKOS [4], le standard W3C utilisé pour représenter ce type de structure de connaissances. A chaque termino-concept est attribué un ou plusieurs termes, des synonymes, des définitions, des notes et des relations de généricité-spécificité et d'association.

La figure 3.5 présente la fiche termino-conceptuelle décrivant le termino-concept ***AAdvantage participant*** créé à partir du terme *airlineparticipant*. L'ingénieur de la connaissance saisit une définition du termino-concept en langue naturelle (onglet « Definition »).

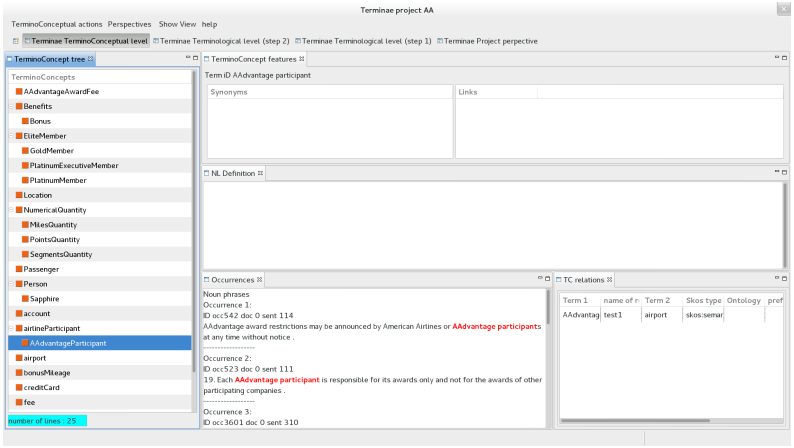

FIGURE 3.5 – La fiche termino-conceptuelle du termino-concept ***AAdvantage participant***.

3.7.3 Le niveau conceptuel

La dernière étape est la phase conceptuelle. Durant cette phase, le réseau obtenu à l'étape précédente est formalisé en une ontologie. Là non plus, la transformation d'un termino-concept en un concept décrit en OWL [5] n'est

4. SKOS (Simple Knowledge Organisation System) est un standard pour la représentation de thésaurus ou de taxonomies pour le web sémantique. http ://www.w3.org/2004/02/skos/

5. http://www.w3.org/2001/sw/wiki/OWL

pas automatique. Dans la pratique, l'ingénieur de la connaissance est amené à faire des choix de formalisation. Plusieurs questions importantes doivent être traitées :

- la distinction entre concepts généraux et concepts individuels (classes et instances) n'est pas établie au niveau termino-conceptuel, aussi entre concepts primitifs et définis,
- il faut organiser les concepts dans une structure hiérarchique, pour assurer les traitements ultérieurs comme les mécanismes d'inférence et de subsomption,
- l'ontologie de domaine doit souvent être rattachée à une ontologie générique ou à une ou plusieurs ontologie(s) existantes.

Prenons comme exemple la formalisation d'un termino-concept en un concept de l'ontologie à créer. Le termino-concept ***airline participant***, relié à plusieurs termes (AAdvantage airline participant, member airline) considérés comme synonymes durant l'analyse terminologique, est formalisé en un concept **Airline participant**. Un termino-concept peut être lié à plusieurs concepts au niveau conceptuel. Dans ce cas de figure, l'ingénieur de la connaissance considère que le sens du termino-concept n'est pas spécifique pour dénoter un sens précis d'un concept, il crée donc plusieurs concepts décrivant chacun un sens spécifique du termino-concept correspondant. Par exemple, le termino-concept ***Benefit*** décrit tous les avantages dont un membre adhérant au programme de fidélité *AA* peut bénéficier. Nous avons remarqué que dans le corpus *American Airlines*, des types d'avantages spécifiques sont mentionnés comme par exemple des espaces privilèges dans les aéroports et des offres de chambres d'hôtels à prix réduit. Dans ce cas, l'ingénieur de la connaissance crée un concept **Benefit** et deux concepts fils **Espace_Privilege** et **Promotion_Hotel**.

Un autre cas de formalisation des termino-concepts est relatif à la création d'instances de concepts. Un termino-concept peut correspondre à une instance au niveau conceptuel. Prenons comme exemple le termino-concept ***Japan airlines*** qui décrit une compagnie aérienne. Ce dernier est formalisé en une instance **Japan airlines** du concept **AAdvantage participant**.

Au niveau conceptuel, l'ingénieur de la connaissance crée aussi des relations conceptuelles à partir de relations termino-conceptuelles. Prenons comme exemple une relation termino-conceptuelle ***earns*** qui relit les termino-

concepts ***AAdvantage member*** et ***Benefit***. L'ingénieur de la connaissance crée une relation relations conceptuelle **earns** entre les concepts **AAdvantage member** et **Benefit**. L'éditeur d'ontologies qui permet ce travail est celui du plugin NEON TOOLKIT ONTOLOGY qui est intégré dans l'outil TERMINAE.

Dans la méthode TERMINAE, chaque concept créé doit être relié à un concept existant dans l'ontologie par une relation de subsomption. TERMINAE permet d'utiliser une core-ontologie afin d'éviter de créer de zéro les concepts génériques liés au domaine à modéliser. Par exemple, l'ingénieur de la connaissance peut enrichir l'ontologie à créer avec des concepts génériques de la core ontologie *LKif*[6]. Dans le cas d'usage *American Airlines*, le termino-concept ***AAdvantage_member*** corerspond au concept **AAdvantage Member** qui est relié par une relation de subsomption au concept **Legal Person** de la core ontologie *LKif*.

A la fin de cette étape, une ontologie décrite en langage OWL est créée et est reliée au lexique normalisé dans l'étape termino-conceptuelle.

3.7.4 Bilan

La méthodologie proposée dans *Terminae* a l'avantage de définir un niveau intermédiaire qui répond à certaines questions autour des choix de modélisation des termes extraits d'un corpus d'acquisition. Elle définit la création des termino-concepts comme objets qui permettent de faire le pont des termes vers des concepts. Elle repose sur l'utilisation d'outils de TAL pour analyser une liste de termes d'une manière globale (en s'appuyant sur leur fréquence) Puis, la méthode propose de regrouper les termes validés par l'ingénieur de la connaissance suivant leur contexte et de définir des termino-concepts sur la base des termes extraits en créant des fiches terminologiques permettant de spécifier l'ensemble des termino-concepts dénotant chacun un sens précis et pertinent pour le domaine. L'outil *Terminae* associé propose des interfaces et quelques recommandations pour l'analyse terminologique et la conceptualisation des termes extraits. Cependant, il n'existe pas une description précise des structures des connaissances manipulées ni des opérations de normalisation nécessaires et de leur enchaînement. La réflexion autour d'une méthodologie qui guide l'ingénieur de la connaissance dans le travail de normalisation

6. http ://www.estrellaproject.org/lkif-core

dans la méthode TERMINAE (Aussenac-Gilles *et al.*, 2008) reste une réflexion méthodologique à gros grain. Pour notre part, nous nous sommes intéressés plus particulièrement aux approches terminologiques car elles concernent la construction d'ontologies à partir du matériau linguistique extrait de textes. Nous proposons dans cette thèse de partir de la méthodologie *Terminae* et de l'enrichir afin de guider au mieux l'ingénieur de la connaissance dans la construction d'ontologies.

Durant l'étape terminologique de la méthode TERMINAE, l'ingénieur de la connaissance se retrouve devant une masse d'information importante et il doit sélectionner les unités textuelles pertinentes au domaine. Les deux principaux inconvénients sont : la quantité d'information à manipuler et le bruit que génèrent les outils de TAL. En effet, les candidats termes extraits par ces logiciels ne sont pas forcément tous des termes du domaine. L'outil TERMINAE propose de trier la liste de candidats termes par fréquence ou suivant un ordre alphabétique. Or l'ingénieur de la connaissance a besoin d'autres indices plus liés au domaine à modéliser. De plus, pour démarrer la phase de conceptualisation, l'ingénieur de la connaissance a besoin de commencer par certaines unités textuelles qui sont marqueurs du domaine surtout lorsqu'il est en face d'une longue liste de termes.

Notons aussi qu'il n'existe pas de dérivation possible entre les relations terminologiques et celles termino-conceptuelles. Pour cela, nous souhaitons définir des liens de correspondance possibles qui permettent de créer une structure de connaissances à partir d'une autre et de garder une « une traçabilité » entre les différents objets formant ces structures de connaissances y compris les relations.

3.8 Conclusion

Dans ce chapitre, nous avons exposé les notions sous-jacentes autour de la construction d'ontologies. Nous avons présenté les méthodes de construction d'ontologies à partir de corpus de textes proposées dans la littérature. Ensuite, nous avons mis l'accent sur l'importance de l'intervention de l'expertise humaine dans le processus de construction d'ontologies à partir de textes. Nous avons souligné l'effort consistant à définir une méthodologie pour guider ce travail Nous avons enfin présenté la méthode TERMINAE qui permet de guider

l'ingénieur de la connaissance dans la construction d'ontologies à partir de textes. C'est le cadre dans lequel nous avons travaillé en essayant d'enrichir et de formaliser la méthode initialement proposée par (Aussenac-Gilles *et al.*, 2008).

Nous avons notamment cherché à proposer une méthodologie précise pour guider le travail de normalisation qui permet de construire un réseau termino-conceptuel à partir du réseau terminologique produit par des outils de TAL. Nous présentons notre méthodologie de normalisation nommée GraphOnto dans le chapitre 5.

Chapitre 4

Structures des connaissances

Sommaire

4.1 Introduction

Ce chapitre présente les niveaux et les structures des connaissances sur lesquels notre méthode de construction de réseau sémantique normalisé s'appuie. Le but de cette formalisation est de clarifier les liens existant entre les structures des connaissances des différents niveaux afin de décrire une traçabilité entre les différentes structures des connaissances. Nous définissons dans ce chapitre quatre niveaux des connaissances : *le niveau discours*, *le niveau terminologique*, *le niveau termino-conceptuel* et *le niveau conceptuel*. Au *niveau discours*, les connaissances sont présentes d'une manière implicite à travers des unités textuelles mentionnées dans un corpus d'acquisition. Au *niveau*

terminologique, les connaissances sont représentées à travers un réseau terminologique qui décrit un vocabulaire spécialisé tel qu'il a été employé dans le texte. Au niveau *termino-conceptuel*, les connaissances du domaine sont décrites à travers un réseau terminologique normalisé appelé *réseau termino-conceptuel* qui décrit un vocabulaire normé défini dans un contexte précis [1] pour un domaine spécifique. Au niveau *conceptuel*, les connaissances sont décrites par un réseau sémantique appelé réseau conceptuel.

Nous précisons des contraintes qui caractérisent les structures des connaissances de chacun de ces niveaux. Nous définissons aussi les correspondances qui peuvent être établies entre ces structures des connaissances, d'une part, pour garder des liens vers le corpus et la représentation conceptuelle, et d'autre part, pour construire une structure des connaissances à partir d'une autre. La dérivation d'une structure des connaissances à partir d'une autre est assurée grâce à des opérations qui permettent de passer d'un niveau des connaissances à un autre.

Nous avons choisi de représenter les connaissances dans chacun des niveaux terminologique, termino-conceptuel et conceptuel à travers les réseaux sémantiques pour plusieurs raisons. D'abord, les unités mentionnées dans un corpus d'acquisition entretiennent en réalité des relations. Il est donc naturel de représenter ces unités et ces relations à travers des réseaux sémantiques. En effet, la représentation des connaissances sous la forme de réseaux sémantiques donne une vision globale des éléments d'un domaine spécifique. De plus, l'interprétation de ces unités ne se fait pas d'une manière isolée ou indépendante. On a besoin d'analyser les relations qu'elles entretiennent pour les interpréter. La modélisation des réseaux sémantiques sous la forme de graphes permet d'exploiter les propriétés de ces réseaux : par exemple la connectivité du réseau et le degré des nœuds. De plus, nous pouvons utiliser des opérations propres aux graphes afin d'établir de nouveaux liens entre les objets constituant ces graphes et d'extraire des sous-graphes remarquables (par ex. extraction de sous-graphes formant des composantes connexes).

Tout au long de ce chapitre, nous mentionnons des exemples pour illustrer et expliquer les structures des connaissances manipulées dans chacun des niveaux et les contraintes associées. Les exemples sont tirés des cas d'usage

1. Les connaissances au niveau termino-conceptuel sont définies par rapport à un cadre applicatif spécifique.

qui sont utilisés pour la construction et l'évaluation du réseau terminologique normalisé dans le chapitre 6. Les cas d'usage sont relatifs à des textes réglementaires qui décrivent des règles métiers associées à un domaine spécifique : le corpus d'American Airlines qui décrit les règles et conditions d'attribution de « miles » pour des voyageurs et le corpus Audi qui est extrait d'une directive internationale et qui décrit les règles et procédures que les véhicules à quatre roues ainsi que leurs équipements doivent satisfaire pour tout ce qui touche aux ceintures de sécurité.

Comme nous manipulons des réseaux sémantiques comme moyen de représentation des connaissances, dans la première section nous définissons les caractéristiques et les opérations effectuées sur ces réseaux. Dans la deuxième partie, nous présentons chacun des niveaux des connaissances (discours, terminologique, termino-conceptuel et conceptuel) ainsi que les différentes structures de connaissances manipulées. La troisième section décrit les correspondances entre les niveaux des connaissances et donc les différents liens existant entre les objets manipulés dans les différents niveaux.

4.2 Préliminaire sur les réseaux sémantiques

L'utilisation des réseaux sémantiques pour la représentation des connaissances est largement exploitée en sciences cognitives (Quillian, 1968), en lexicographie pour la représentation du sens en langue (Fellbaum, 1998) (par ex. WordNet), en terminologie (Nazarenko, 2004), ou encore en visualisation des connaissances (Crampes *et al.*, 2008). Un type particulier de réseau sémantique est celui des graphes conceptuels (Kayser, 1997) qui sont utilisés pour la modélisation des connaissances.

Nous adoptons une définition simple des réseaux sémantiques. Un réseau sémantique est un ensemble d'unités sémantiques qui sont reliées entre elles à travers des relations. Les relations sont représentées par des triplets $R(Source, Label, Destination)$ où les arguments *Source* et *Destination* sont des unités sémantiques et *Label* est le type de la relation reliant ces deux unités sémantiques.

Le réseau sémantique est représenté par un graphe $G(N, A, L_A)$ étiqueté

et orienté où les nœuds de N sont interconnectés par des arcs [2] de A et les arcs portent des étiquettes L_A tels que $\{A : N * N * L_A = \{(x, y, l), x \in N, y \in N, l \in L_A\}\}$. Les nœuds représentent des unités sémantiques et les arcs étiquetés représentent des types de relations. Un arc relie un nœud *Source* s'agissant de l'extrémité initiale de l'arc et un nœud *Destination* représentant l'extrémité finale de l'arc. Un nœud peut être lié à un ou plusieurs nœuds.

Certaines informations peuvent être extraites des réseaux sémantiques décrits sous la forme de graphe :

- l'ensemble des arcs tel que un nœud x joue le rôle de source ou de destination : ce sont respectivement les ensembles des successeurs $Succ_x$ et prédécesseurs $Pred_x$. L'ingénieur de la connaissance peut explorer les relations que le nœud x entretient avec d'autres nœuds.
- le degré d'un nœud $d^{\circ}(x) = \{y \in N, (x, y) \in A\}$ est le nombre d'arcs partant et arrivant d'un nœud x. Cette information donne une idée sur l'importance du nœud au sein du réseau. L'ingénieur de la connaissance peut extraire le sous-graphe $G_x(x, A_x)$ tel que le nœud x est relié à ses voisins s'agissant de l'ensemble de ses successeurs et prédécesseurs.

4.3 Structures des connaissances

Dans cette section, nous décrivons les structures qui supportent les connaissances associées dans chacun des niveaux discours, terminologique, terminoconceptuel et conceptuel. Nous introduisons aussi les contraintes appliquées à ces structures des connaissances. Enfin, nous définissons les opérations qui permettent de manipuler ces différentes structures.

4.3.1 Le niveau discours

Au niveau discours, les connaissances sont implicites dans les documents. Les documents sont rassemblés en tenant compte du domaine et de l'application visée. Ces documents forment un corpus. Généralement, ces documents véhiculent des connaissances propres à un domaine à travers le vocabulaire partagé au sein d'une communauté.

2. On parle d'arête lorsque le graphe n'est pas orienté.

La figure 4.1 décrit un extrait du corpus du cas d'usage de *American Airlines*.

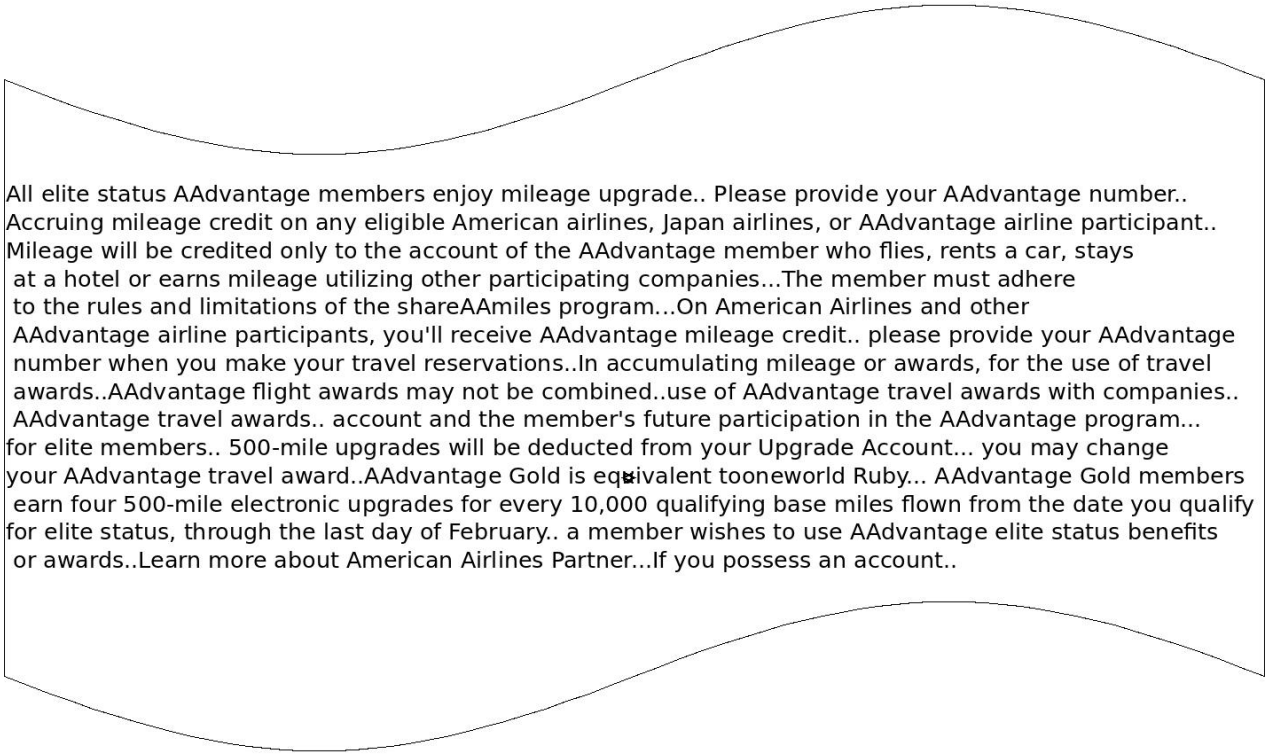

FIGURE 4.1 – Exemple d'un extrait de corpus *American Airlines*.

Formellement un corpus d'acquisition C est constitué d'une séquence de phrases $ph_i, ph_j, ...$ telle que chaque phrase est composée d'une séquence d'unités textuelles $utxt_i, utxt_j, ...$ Chaque unité est repérée par sa position de début (numéro du premier caractère de l'unité dans la phrase) et de fin (numéro du dernier caractère de l'unité dans la phrase) dans une phrase du texte. Comme le texte peut être préalablement étiqueté et annoté, les unités textuelles peuvent être associées à des propriétés syntaxiques comme leurs formes lemmatisées ou encore des types syntaxiques. Le tableau 4.1 décrit les informations qui sont associées à des unités textuelles. Il n'y a pas de restriction sur les types syntaxiques associés aux unités textuelles. Ces types dépendent de l'analyse syntaxique utilisée.

Le niveau discours fait partie de la langue. Il est peu contraint Les unités

Propriété	Définition
Forme	séquence de caractères
Lemme	forme canonique
Type syntaxique	Nom, Verbe, etc
PositionInitiale	offset du premier caractère de l'unité dans la phrase
PositionFinale	offset du dernier caractère de l'unité dans la phrase
IdPh	numéro de la phrase correspondante

TABLE 4.1 – Les propriétés des unités textuelles.

textuelles extraites peuvent correspondre à plusieurs catégories syntaxiques (cas d'ambiguité). Prenons l'exemple où suivant les phrases l'unité textuelle *credit* est rattachée soit au type syntaxique *Nom*, soit au type syntaxique *Verbe* :

1. *You receive credit from your AAdvantage transaction.*
2. *You may credit accrued mileage for every flight on American airlines.*

Le niveau discours est caractérisé par une redondance d'informations, de répétitions et de variations de formes. Les mêmes unités textuelles peuvent apparaître plusieurs fois dans le texte sous différentes formes (par ex. *member*, *members*). Les unités textuelles mentionnées dans le texte peuvent être polysémiques et ambigues. Prenons l'exemple de l'unité textuelle *member* qui décrit, suivant les phrases où elle est mentionnée, un membre adhérant dans un programme de fidélité ou une compagnie aérienne :

1. *Members must have mileage earning or redeeming activity once every 18 months in order to retain their miles.*
2. *The minimum mileage amount earned may be less than 500 miles for travel on oneworld member airlines and AAdvantage participating airlines.*

4.3.2 Le niveau terminologique

Le niveau terminologique est une représentation linguistique du vocabulaire spécialisé utilisé dans le texte pour le domaine spécifique que l'on cherche

à modéliser. Ce vocabulaire est constitué d'unités terminologiques, essentiellement des termes et des entités nommées, ainsi que des relations terminologiques. L'interprétation de ces unités terminologiques par l'ingénieur de la connaissance permet d'identifier des notions partagées dans une communauté et un domaine spécifiques. Le plus souvent, les termes sont composés de syntagmes nominaux mais des verbes ou des syntagmes verbaux peuvent aussi avoir une valeur terminologique. Les entités nommées sont un autre type d'unités terminologiques. Le plus souvent, elles renvoient à des « entités » du domaine et peuvent relever de différentes catégories linguistiques (par ex. des noms propres «Air France», des pronoms «elle», etc). Ce sont des unités terminologiques particulières car elles ont une valeur sémantique référentielle et stable : elles désignent de manière stable des référents dans un domaine spécifique. Les relations terminologiques sont souvent décrites par l'intermédiaire de verbes mais aussi de syntagmes nominaux (par exemple le syntagme nominal *mileage credit* décrit le fait de créditer un compte avec des miles). Le fait que la valeur d'une relation terminologique s'exprime elle-même par une unité terminologique fait que les étiquettes et les arguments des relations sont de même nature. Nous pensons que c'est une particularité du niveau terminologique. D'autres types d'informations peuvent être associées aux unités terminologiques ou aux relations extraites, comme par exemple les types sémantiques pour les entités nommées ou bien des classes sémantiques formées d'unités terminologiques voisines.

Au niveau terminologique, les connaissances terminologiques sont représentées par un réseau terminologique. Le réseau terminologique est constitué par un ensemble d'unités terminologiques qui sont reliées par des relations terminologiques. Le réseau terminologique est généralement peu connexe. Il peut y exister des unités terminologiques isolées (c'est-à-dire qui n'entretiennent pas de relations avec d'autres unités). Cela est dû au fait que les extracteurs de relations n'identifient pas toutes les relations terminologiques pouvant relier des unités terminologiques. Souvent le réseau terminologique est plus riche en relations syntaxiques qu'en relations sémantiques car l'analyse syntaxique est assez générique alors que les relations sémantiques sont détectées par application de patrons linguistiques définis au préalable.

Formellement, le réseau terminologique est un réseau étiqueté et orienté $G_T(UT, RT)$ défini par la donnée d'un ensemble d'unités terminologiques,

UT, contenant des termes et des entités nommées, et d'un ensemble de relations terminologiques, RT. Une relation terminologique est représentée par un triplet $rt(ut_i, ut_k, ut_j)$ tel que ut_i, ut_j sont des unités terminologiques de UT correspondant aux arguments *Source* et *Destination* de la relation et ut_k est une unité terminologique de UT correspondant au type de la relation $ut_k = typeRT$.

La figure 4.2 décrit un exemple d'un réseau terminologique. Pour faciliter la lecture des graphes et éviter de représenter des relations ternaires, ici, nous dupliquons les unités terminologiques qui servent à étiqueter des relations sous la forme d'étiquettes de relations. Afin de simplifier la visualisation des éléments du réseau G_T, nous ne représentons pas systématiquement les nœuds qui décrivent les unités terminologiques associées aux types des relations terminologiques, dans la suite de ce chapitre. Les types des relations terminologiques sont portés par les arcs reliant des nœuds au niveau du graphe terminologique G_T.

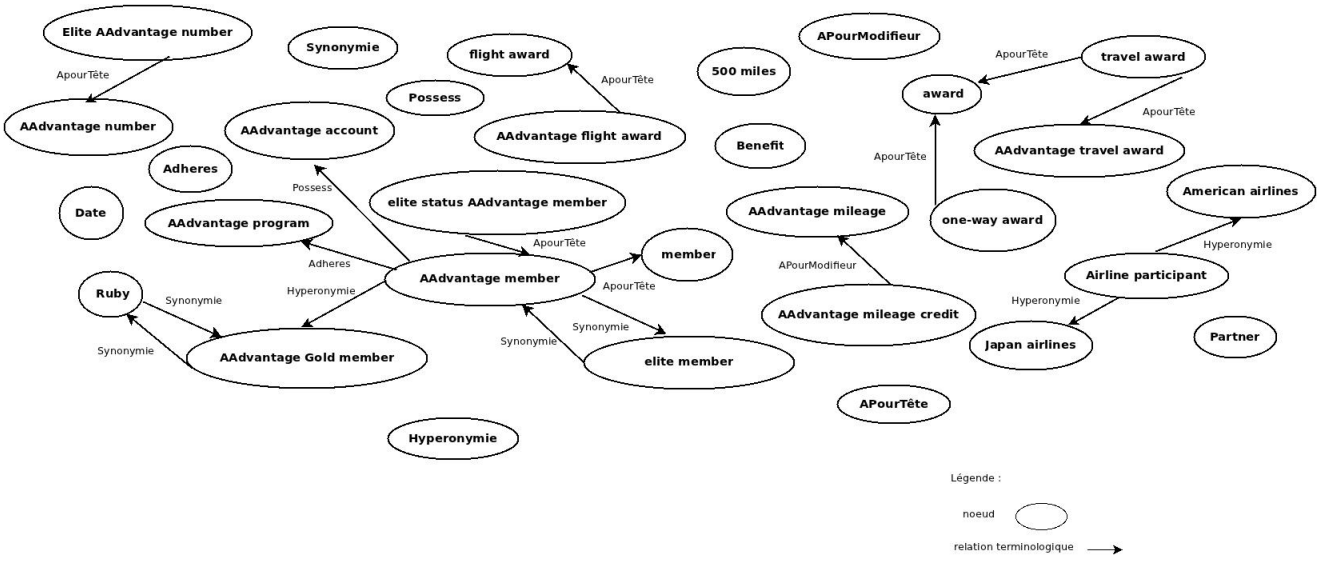

FIGURE 4.2 – Exemple d'un sous réseau terminologique du cas d'usage de *American Airlines*.

Réseau terminologique

Dans cette section, nous décrivons chacun des éléments constituant le réseau terminologique, à savoir les unités terminologiques, les types des relations terminologiques et les relations terminologiques, ainsi que les contraintes appliquées sur ces connaissances.

Unités terminologiques et contraintes associées Une unité terminologique $ut \in UT$ est représentée par un nœud au niveau du réseau terminologique. Ces unités sont caractérisées par la structure définie dans le tableau 4.2. Parmi les propriétés relatives à une unité terminologique, nous définissons la propriété $Marqueurs(UT)$ qui liste l'ensemble des formes que peut prendre une unité terminologique dans le texte. Il peut s'agir de termes (*Airline participant*), d'entités nommées (*Japan airlines*) ou de patrons linguistiques (*airline company/ies*) [3]. Une unité terminologique peut être associée à un ou plusieurs types sémantiques. Cette information est plus généralement disponible lorsque l'un des marqueurs est une entité nommée.

Propriété	Définition
Label	étiquette lexicale
Type sémantique	le(s) type(s) sémantique(s) associé(s)
Marqueurs	ensemble des marqueurs relatifs à une unité terminologique

TABLE 4.2 – Les propriétés des unités terminologiques au niveau terminologique.

Généralement, le label est l'un des marqueurs choisi comme forme canonique de l'unité terminologique. Le tableau 4.3 tiré du cas d'usage *AA* décrit par exemple les propriétés de l'unité terminologique *Airline participant*.

Les nœuds représentant les unités terminologiques au niveau du réseau terminologique sont soumis à des contraintes qui délimitent leur niveau d'expressivité. Nous définissons des contraintes associées aux unités terminologiques :

– Toute unité terminologique est décrite par au moins un marqueur qui est son label :

3. Au niveau terminologique, il y a une redondance d'informations comme le plus souvent les résultats des outils de TAL se chevauchent. Par exemple, un des marqueurs d'une unité terminologique peut aussi être extrait comme synonyme de la même unité.

Propriété	Valeur
Label	Airline participant
Type sémantique	Organisation
Marqueurs	AAdvantage Airline participant, Airline participating, Airline participants

TABLE 4.3 – Exemple des propriétés relatives à l'unité terminologique *Airline participant*.

$$\forall ut \in UT, Label(ut) \in Marqueurs(ut)$$

– Deux unités terminologiques distinctes ne partagent pas le même label :

$$\forall ut_i, ut_j \in UT, ut_i \neq ut_j \sqsupset Label(ut_i) \neq Label(ut_j)$$

– Les relations définies entre deux unités terminologiques sont distinctes :

$$\forall ut_i, ut_j \in UT, rt_1, rt_2 \in RT, rt_1(ut_i, type_k, ut_j) \sqcap rt_2(ut_i, type_s, ut_j)$$
$$\sqcap rt_1 \neq rt_2 \sqsupset type_k \neq type_s$$

Types des relations terminologiques et contraintes associées Les types des relations terminologiques sont des unités terminologiques. Certains types n'ont pas de marqueurs qui les dénotent dans le texte comme par exemple les relations syntaxiques. Comme nous prévoyons d'utiliser des opérations de normalisation qui sont paramétrées suivant le type de la relation en question et de définir des requêtes suivant un type précis de relation terminologique afin d'extraire par exemple le sous graphe correspondant, nous avons défini une typologie des types de relations terminologiques. La hiérarchie des types de relations terminologiques est décrite dans la figure 4.3 :

– les relations syntaxiques permettent d'identifier les unités terminologiques qui composent d'autres unités. Les relations syntaxiques sont importantes à prendre en considération car les termes spécialisés sont composés les uns des autres. Il est donc intéressant d'étudier ce type de relation afin de détecter d'éventuelles relations propres à un domaine spécifique. Prenons par exemple la relation *APourTête* qui relit l'unité terminologique *AAdvantage Airline participant* à l'unité terminologique *AAdvantage Airline*.

- les relations lexicales sont des relations identifiées entre des unités terminologiques et reconnues par des patrons lexico-syntaxiques définis suivant le type de la relation lexicale à identifier dans le texte, à savoir l'hyperonymie, l'antonymie, la synonymie et la méronymie. Prenons par exemple la relation de *synonymie* définie entre les unités terminologiques *oneworld Ruby* et *AAdvantage Gold* grâce à l'application du patron *SN is equivalent to SN*.
- les relations spécialisées sont des relations sémantiques que l'ingénieur de la connaissance définit entre les termes du domaine ou bien qui sont détectées à travers l'application des patrons lexico-syntaxiques sur le texte. Nous avons par exemple défini une relation spécialisée *creditedBy* entre les les unités terminologiques *AAdvantage account* et *Miles*.
- La relation de type *voir aussi* est un type que l'ingénieur de la connaissance choisit pour marquer l'existence d'une relation terminologique entre deux unités quand il ne sait pas identifier son type de manière claire.

.

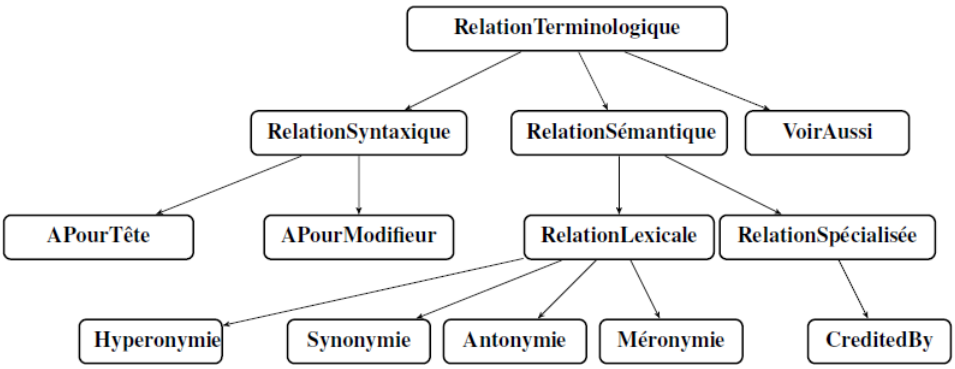

FIGURE 4.3 – Typologie des types de relations terminologiques. Ces types de relations terminologiques sont observables dans les réseaux terminologiques extraits de corpus.

Les types des relations terminologiques sont associés à des contraintes :

- la relation de *Synonymie* est une relation symétrique. Il existe donc une relation inverse entre les deux arguments de la relation ;
- la relation de *Hyperonymie* n'est pas une relation symétrique ;
- la relation de *Antonymie* est une relation symétrique ;

- la relation *Syntaxique* est une relation entre composé et composant ;
- la relation *Spécialisée* est une relation sémantique de domaine ;
- la relation *Voir Aussi* est un type de relation terminologique qui ne porte pas de sémantique ;
- un type de relation terminologique ne peut avoir qu'un et un seul père ;
- deux types de relations terminologiques sont synonymes si leurs labels (donc les labels des unités terminologiques correspondantes) sont synonymes ;
- tout type de relation terminologique a un label (le type de la relation) unique.

Relations terminologiques et contraintes associées Au niveau du réseau terminologique, les relations terminologiques sont définies par des triplets $rt(ut_i, ut_k, ut_j)$ tels que ut_i, ut_j sont des unités terminologiques jouant le rôle d'arguments et ut_k décrit le type de la relation terminologique.

Prenons l'exemple suivant qui décrit la relation terminologique *creditedBy* entre les les unités terminologiques *AAdvantage account* et *Miles* (4.4).

Propriété	**Valeur**
Type	creditedBy
Source	AAdvantage account
Destination	Miles

TABLE 4.4 – Exemple d'une relation terminologique *creditedBy*.

Il y a peu de contraintes sur les relations au niveau terminologique. Il peut y avoir plusieurs relations de type différent entre deux unités terminologiques. Il peut par exemple y avoir une relation de synonymie et une autre d'hyperonymie entre deux unités terminologiques. Prenons l'exemple suivant tiré du cas d'usage de *American Airlines* où les unités terminologiques *privilege* et *benefit* dénotent une relation d'hyperonymie dans le première phrase et une relation de synonymie dans la deuxième :

Elite membership has its privilege such as benefits earned during the year.
An elite member will receive the privilege of exclusive service as a benefit.

Opérations

Au niveau terminologique, l'ingénieur de la connaissance peut effectuer des changements au sein du réseau terminologique. Lui seul peut vérifier ou juger de la cohérence du réseau obtenu suite aux changements appliqués en l'absence de contrainte formelle définie au niveau terminologique. Dans la suite, nous définissons un ensemble d'opérations permettant d'agir sur le réseau terminologique :

1. *Fusion d'unités terminologiques* : cette opération permet de fusionner deux ou plusieurs unités terminologiques en une seule unité. L'unité créée porte un label et se voit associer l'ensemble des marqueurs des unités fusionnées. Les successeurs (respectivement les prédécesseurs) des unités à fusionner deviennent les successeurs (respectivement les prédécesseurs) de l'unité fusionnée.

   La figure 4.5 décrit le résultat de la fusion des unités terminologiques *AAdvantage travel award* et *AAdvantage flight award*.

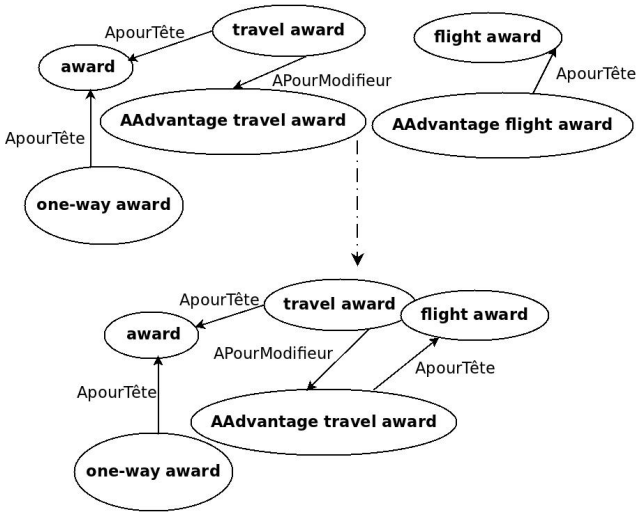

FIGURE 4.4 – Opération de fusion.

2. *Éclatement des unités terminologiques* : cette opération permet l'éclatement d'une unité donnée en deux ou plusieurs unités. Chacune des unités créées, suite à un éclatement, est reliée aux successeurs (respectivement aux prédécesseurs) de l'unité éclatée. Les nœuds créés suite à l'éclatement se voit associer aux marqueurs du nœud éclaté.

La figure 4.5 décrit le sous réseau initial et la figure 4.6 décrit le résultat de l'éclatement de l'unité terminologique *AAdvantage member* en deux unités terminologiques *AAdvantage member* et *Member airline*.

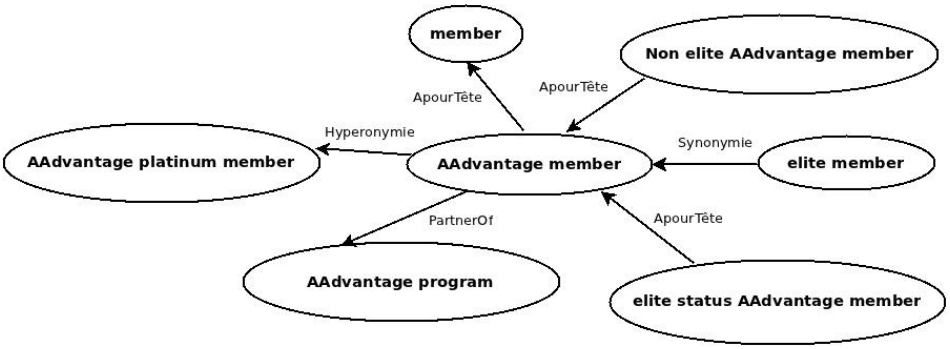

FIGURE 4.5 – Réseau avant éclatement de l'unité terminologique *AAdvantage member*.

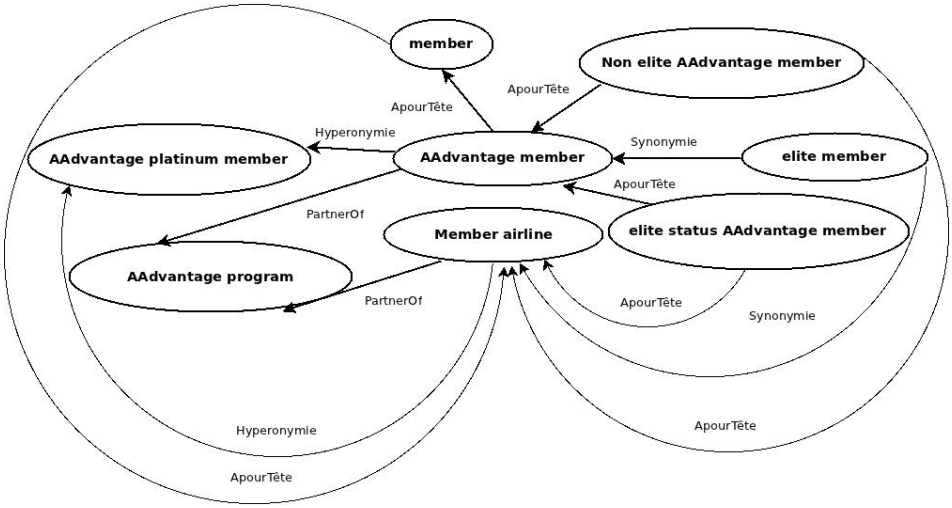

FIGURE 4.6 – Réseau après éclatement de l'unité terminologique *AAdvantage member*.

3. *Ajout/ Suppression d'éléments* : l'ingénieur de la connaissance peut ajouter ou supprimer des unités (respectivement des relations). La suppression des unités entraîne la suppression des relations entrant et sortant de l'unité supprimée. Par contre, la suppression d'une relation n'entraîne pas la suppression des unités source et destination.
4. *Opérations élémentaires* : ces opérations permettent de modifier les propriétés des unités et des relations terminologiques. L'ingénieur de la

connaissance peut par exemple ajouter des marqueurs à une unité terminologique ou modifier le type d'une relation terminologique.

Au niveau terminologique, la sémantique des connaissances manipulées reste floue et cela est dû à plusieurs aspects linguistiques caractérisant ce niveau. D'abord, il y a des variations de formes des unités terminologiques (*N.Y, New York*) qui sont interchangeables. Parfois, elles peuvent être ambiguës (l'unité terminologique *member* décrit à la fois un passager adhérant dans un programme de fidélité et une compagnie aérienne participant dans le même programme).

Dans le réseau terminologique certaines unités terminologiques qui décrivent des types de relations de domaine d'où un mélange entre des unités décrivant des notions du domaine et d'autres unités qui dénotent des relations sémantiques (le terme *mileage credit* décrit le fait de créditer un compte appartenant à un membre à travers des miles ou des points de fidélité). Les unités terminologiques et les relations terminologiques ne sont pas toutes pertinentes pour le domaine. L'ingénieur de la connaissance doit faire le tri. Par exemple pour le cas d'usage de AA, l'unité terminologique *airport agent* n'a pas été considérée comme pertinente. Par ailleurs, au niveau terminologique les connaissances relatives à un domaine spécifique ne sont pas toutes explicites. Des unités ou des relations terminologiques peuvent manquer à ce niveau comme par exemple la relation de synonymie qui relit l'unité terminologique *itinerary* à *segment* décrite implicitement dans l'extrait du texte suivant :

Each upgrade is valid for 500 miles of travel. Each flight segment requires at least one upgrade....AAdvantage members checking-in for an itinerary that requires 500-mile upgrades.

4.3.3 Le niveau termino-conceptuel

Le niveau termino-conceptuel décrit le vocabulaire normalisé d'un domaine spécifique d'une manière non ambiguë dans le but de faciliter la communication dans une communauté spécifique. La normalisation suppose qu'on a fixé au préalable le contexte pour interpréter le vocabulaire du domaine. Il s'agit de définir la manière avec laquelle les termes sont utilisés dans un usage particulier, de fixer leurs sens et de les différencier les uns des autres. La normalisation terminologique vise à expliciter la dénomination des notions d'un

domaine spécifique dans un usage particulier. Cet usage permet de fixer la signification des termes employés au sein d'un domaine spécifique.

Le réseau termino-conceptuel est un réseau terminologique normalisé tel que sa structure permet de contraindre l'interprétation des unités définies. En effet, la position d'un termino-concept dans le réseau est définie suivant ses différences avec l'unité parente (le termino-concept jouant le rôle de *Destination* de la relation de type *Généricité/Spécificité*) et de ses frères (les termino-concepts qui entretiennent des relations de type *Généricité/Spécificité* avec l'unité parente).

La structure des connaissances définie à ce niveau est un réseau termino-conceptuel composé d'un ensemble de termino-concepts interconnectés par des relations termino-conceptuelles. Les termino-concepts sont des termes non ambigus qui sont pertinents pour un domaine spécifique. Les relations termino-conceptuelles sont des relations sémantiques de domaine décrivant des liens de généricité/spécificité ou des liens associatifs entre termino-concepts. Le réseau termino-conceptuel est représenté par un graphe étiqueté, orienté et pas forcement connexe [4] $G_{TC}(TC, RTC, Type_{RTC})$ défini par la donnée d'un ensemble de termino-concepts TC, d'un ensemble de relations termino-conceptuelles RTC et un ensemble de types de relations termino-conceptuelles $Type_{RTC}$.

Réseau termino-conceptuel

Dans cette section, nous définissons les éléments constituant le réseau termino-conceptuel à savoir les termino-concepts, les types et les relations termino-conceptuelles. Nous spécifions pour chacun d'eux les contraintes qui lui sont associées. La figure 4.7 décrit un exemple de réseau termino-conceptuel tiré du cas d'usage de *American Airlines*.

Unités termino-conceptuelles et contraintes associées Un termino-concept décrit un sens pertinent d'une manière unique par rapport à un contexte précis. Il représente un terme normalisé et désambiguisé. Chaque termino-concept $tc \in TC$ est étiqueté par un terme vedette tv. Les termino-concepts appartenant à TC sont caractérisés par un ensemble de propriétés

4. Le réseau termino-conceptuel forme un thésaurus dont les nœuds décrivent des sens pertinents pour le domaine et sont distingués de leurs pères et de leurs frères.

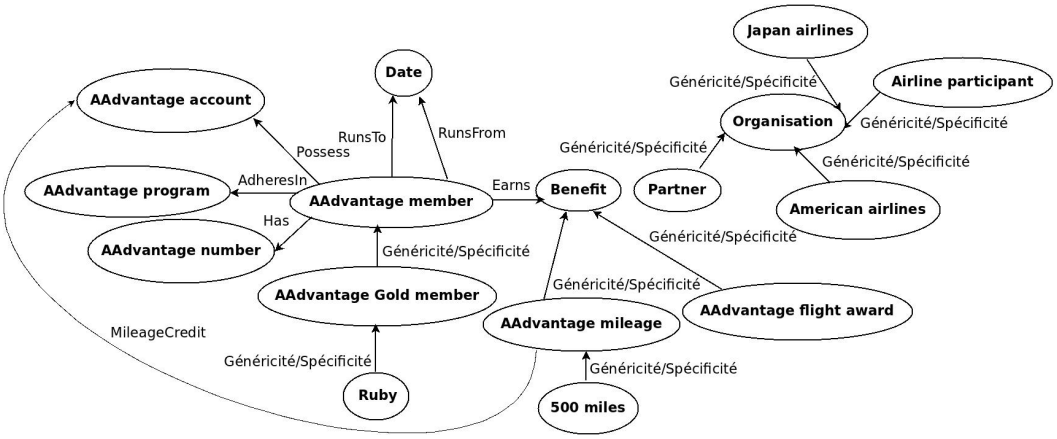

FIGURE 4.7 – Exemple d'un sous réseau termino-conceptuel tiré du cas d'usage de *American Airlines*.

définies dans le tableau 4.5.

Propriété	**Définition**
Label préféré	terme vedette correspondant
Labels alternatifs	liste des termes alternatifs
Définition	une définition de la notion dénommée par le TC écrite en langage naturel

TABLE 4.5 – Les propriétés des termino-concepts de TC.

La définition d'un termino-concept (Charlet *et al.*, 2008) est établie suivant les propriétés partagées avec l'unité parente et celles le différenciant de l'unité parente et des unités frères. Cette définition permet de fixer le sens du termino-concept dans un contexte précis. Prenons l'exemple du termino-concept *AAdvantage Gold member* qui décrit une catégorie particulière de membres adhérants au programme de fidélité et bénéficiant de certains avantages comme par exemple des bonus gagnés à chaque vol. L'exemple est décrit dans le tableau 4.6.

L'ensemble des termino-concepts de TC est soumis à des contraintes :

– un termino-concept n'est relié qu'à un et un seul termino-concept père ;
– un termino-concept ne possède qu'un et un seul label préféré ;

Propriété	Valeur
Label préféré	AAdvantage Gold member
Labels alternatifs	One-world Ruby, Gold member,
Définition	une catégorie particulière de membres adhérants au programme de fidélité et qui gagnent 25 % de bonus pour chaque vol

TABLE 4.6 – Les propriétés du termino-concept *AAdvantage Gold member*.

- un termino-concept doit avoir une définition en langage naturel exprimant les caractéristiques en commun avec le termino-concept père et celles qui le différencient des termino-concepts frères et du termino-concept père ;
- deux termino-concepts ne peuvent pas avoir le même label préféré.

Types des relations termino-conceptuels et contraintes associées Les relations termino-conceptuelles sont de deux types distincts décrits dans la figure 4.8 et tels que :

- les relations taxonomiques décrivent des liens de type *Généricité/Spécificité* entre des termino-concepts ;
- les relations associatives dénotent des liens non taxonomiques entre des termino-concepts.

Les types termino-conceptuels sont représentés par des étiquettes lexicales associées aux arcs au niveau du graphe sous-jacent au réseau termino-conceptuel G_{TC}. Prenons l'exemple du termino-concept *AAdvantage Gold member* relié par une relation termino-conceptuelle de type *Généricité/Spécificité* au termino-concept père *AAdvantage member*. Ce dernier étant relié au termino-concept *AAdvantage Program* par une relation de type

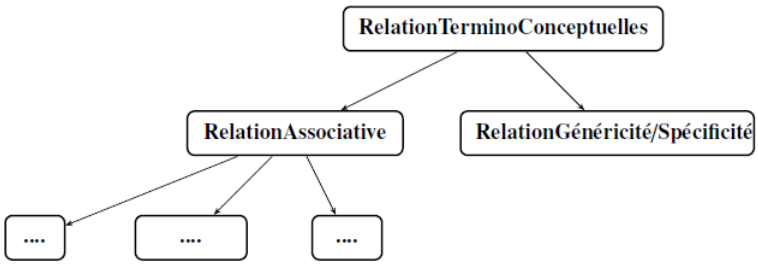

FIGURE 4.8 – Typologie des types de relations termino-conceptuelles.

RelationAssociative ayant comme étiquette *adheresIn*.

Les types termino-conceptuels de *RTC* possèdent des propriétés décrites dans le tableau 4.7.

Propriété	**Définition**
Id	identifiant unique
Label	type sémantique de la relation
Père	type sémantique de la relation parente
Transitivité	si relation est transitive
Symétrie	si relation est symétrique
Réflexivité	si relation est réflexive
Inverse	type sémantique de la relation inverse

TABLE 4.7 – Les propriétés des types termino-conceptuels *RTC*.

Prenons l'exemple du type termino-conceptuel *adheresIn* qui relit les deux termino-concepts *AAdvantage member* et *AAdvantage Programme* décrit dans le tableau 4.8 :

Un type termino-conceptuel ne peut avoir qu'un seul père. La relations de type *Généricité/Spécificité* est une relation transitive, non réflexive et non symétrique. Le type de relation associatif en revanche n'est pas contraint. Les relations associatives sont des relations spécialisées relevant d'un domaine spécifique.

Relations termino-conceptuelles et contraintes associées Au niveau termino-conceptuel, les relations termino-conceptuelles sont des relations sé-

Propriété	Valeur
Id	XX
Label	adheresIn
Père	RelationAssociative
Transitivité	Non
Symétrie	Non
Réflexivité	Non
Inverse	HasMember

TABLE 4.8 – Les propriétés du type termino-conceptuel *adheresIn*.

mantiques qui décrivent des relations spécialisées de domaine. Les relations taxonomiques décrivent des liens de type *Généricité/Spécificité* entre des termino-concepts. Les relations associatives dénotent des liens non taxonomiques entre des termino-concepts. Les relations taxonomiques permettent de structurer l'ensemble des termino-concepts sous la forme d'un arbre ce qui permet de définir des termino-concepts plus généraux ou plus spécifiques que d'autres. Le degré de granularité (la profondeur de l'arbre définie de la racine vers les feuilles) dépend de tâche de modélisation et de l'application visée. Les relations associatives permettent de définir des liens associatifs entre termino-concepts et de rendre explicite les dépendances entre ces derniers. Cela enrichit la sémantique du réseau termino-conceptuel. En effet, un termino-concept est défini par ses propriétés et par des relations qui le lient avec d'autres termino-concepts.

Au niveau du réseau termino-conceptuel, les relations termino-conceptuelles sont décrites par un triplet $rtc(tc_i, typertc, tc_j)$ avec tc_i et tc_j des termino-concepts *Source* et *Destination* et *typertc* est le type d'une relation termino-conceptuelle appartenant à l'ensemble $Type_{RTC}$. Les relations termino-conceptuelles sont associées à des contraintes :

- une relation termino-conceptuelle est associée qu'un et un seul type (par ex. la propriété *Père* ne décrit qu'une seule valeur) ;
- une relation termino-conceptuelle de type *RelationAssociative* possède un label unique.

Opérations

Nous définissons des opérations que l'ingénieur de la connaissance peut exécuter au niveau du réseau termino-conceptuel :

1. création et suppression des termino-concepts et des relations termino-conceptuelles : cette opération permet la création ou la suppression d'un termino-concept ou d'une relation termino-conceptuelle. La suppression d'un termino-concept entraîne la suppression des relations termino-conceptuelles où il joue le rôle de *Source* ou de *Destination*. La création d'une relation termino-conceptuelle nécessite que l'ingénieur de la connaissance spécifie ses arguments.
2. modification des propriétés des termino-concepts et des relations termino-conceptuelles : l'ingénieur de la connaissance peut modifier les propriétés d'un termino-concept ou d'une relation termino-conceptuelle tout en respectant les contraintes appliquées sur ces deux types de connaissances ;
3. association de labels alternatifs aux termino-concepts : cette opération permet d'associer d'autres labels à un termino-concept. Plusieurs termino-concepts peuvent partager les mêmes labels alternatifs. Par exemple les deux termino-concepts *AAdvantage member* et *Airline participant* partagent le label alternatif *member*.

Le réseau termino-conceptuel est un réseau qui n'a pas de contraintes de connectivité et non ambigu. Les termino-concepts sont des unités pertinentes pour le domaine. Chaque termino-concept décrit un sens unique. Ces termino-concepts sont structurés sous une forme taxonomique et entretiennent des liens associatifs. Le réseau termino-conceptuel peut servir comme un vocabulaire de domaine normalisé et structuré pour la construction d'ontologies de domaine à partir de textes. Il peut aussi être considéré comme une ressource sémantique pour l'annotation des documents et des applications d'accès au contenu.

4.3.4 Le niveau conceptuel

Au niveau conceptuel, il s'agit de définir les concepts, leurs propriétés, leurs instances et les relations conceptuelles les reliant dans un langage formel. Le langage formel OWL permet de définir des concepts, de leur associer des

propriétés, des instances et des relations formant une structure des connaissances appelée *ontologie*. Il assure l'héritage des propriétés entre concepts, entre concepts et instances et entre relations. OWL propose aussi la restriction des relations et leur associe des cardinalités. Une ontologie créée peut être enrichie avec des axiomes et des règles afin de contrôler les champs d'expressivité des concepts et de leurs relations. Elle décrit des connaissances partagées dans une communauté, par rapport à un usage précis.

Notre but étant de formaliser les connaissances définies pour montrer la particularité des connaissances manipulées à ce niveau et pouvoir établir des correspondances avec les niveaux des connaissances, nous proposons d'utiliser au niveau conceptuel une représentation simplifiée de l'ontologie sous la forme d'un réseau sémantique appelé *réseau conceptuel*. Une ontologie peut se décrire par un réseau comprenant un ensemble de concepts, d'instances, de valeurs, de types de relations et de relations conceptuelles. Nous ne représentons ni la cardinalité des relations, ni leur restriction au niveau du réseau sémantique. Dans la suite, nous définissons chacune de ces connaissances manipulées au sein du réseau conceptuel.

Réseau conceptuel

Le réseau conceptuel se modélise sous la forme d'un graphe orienté et étiqueté $G_C(UC, RC, TRC)$ tel que l'ensemble des unités conceptuelles UC décrit des concepts C, des instances I et des valeurs V de l'ontologie et l'ensemble de RC décrit des relations conceptuelles où chaque relation conceptuelle est définie par un triplet $r(uc_i, trc, uc_j)$ tel que uc_i et uc_j sont des unités conceptuelles et trc est le type de la relation conceptuelle. La figure 4.9 décrit un exemple d'un réseau conceptuel tiré du cas d'usage de AAdvantage.

Unités conceptuelles et contraintes associées Au sein du réseau conceptuel, l'ensemble de UC correspond à l'ensemble des concepts C, des instances I et des valeurs V et constituent des nœuds dans le graphe sous-jacent au réseau conceptuel. Les concepts de l'ontologie O décrivent des notions pertinentes par rapport à un domaine et une application visée. L'ensemble des instances de I représente les instances des concepts de l'ontologie O. L'ensemble des valeurs de V décrivent des valeurs numériques et non numériques s'agissant des propriétés des concepts au niveau de l'ontologie O.

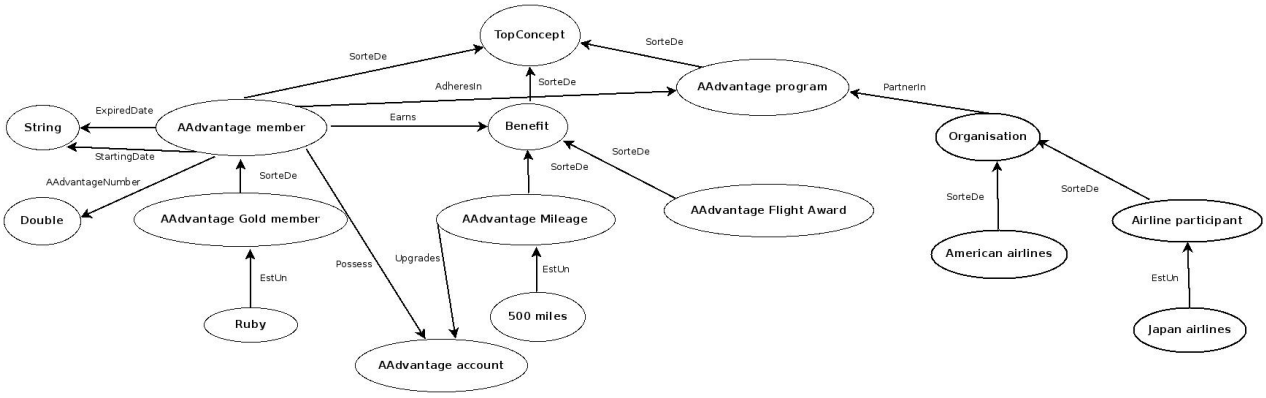

FIGURE 4.9 – Exemple d'un réseau conceptuel.

Les unités conceptuelles sont définies à travers des propriétés décrites dans le tableau 4.9, qui permettent de distinguer les types des nœuds associés à ces unités.

Propriété	**Définition**
Label	étiquette lexicale
Type	concept, instance ou valeur

TABLE 4.9 – Les propriétés des unités conceptuelles.

Au niveau conceptuel, nous définissons des contraintes relatives aux unités conceptuelles manipulées :

- parmi les unités conceptuelles, il existe une et une seule unité conceptuelle appelée *TopConcept* appartenant à l'ensemble des concepts de C représentant la racine du réseau conceptuel ;
- deux unités conceptuelles distinctes ne peuvent pas avoir le même label ;
- Une unité conceptuelle $uc_i \in UC$ est soit un concept, soit une instance soit une valeur ;
- toute unité conceptuelle uc_i jouant le rôle de nœud *Source* d'une relation donnée appartient à l'ensemble de $\{C \cup I\}$. Cet ensemble définit les concepts et les instances de l'ontologie O. Toute unité conceptuelle uc_j jouant le rôle de nœud *Destination* d'une relation donnée appartient à l'ensemble de $\{C \cup I \cup V\}$;

– toute unité conceptuelle uc_i jouant le rôle de nœud *Source* d'une relation taxonomique rt_{ij} hérite toutes les relations associatives de l'unité conceptuelle c_j jouant le rôle de nœud *Destination* de la même relation rt_{ij}.

La différenciation entre des nœuds concepts et des nœuds instances se fait suivant les types des relations conceptuelles reliant ces nœuds. En effet, les instances sont représentées par des nœuds feuilles c'est-à-dire ne jouant pas le rôle de nœud *Destination* d'une relation taxonomique.

Types de relations conceptuelles et contraintes associées Au niveau conceptuel, nous définissons deux types de relations conceptuelles reliant des unités conceptuelles :

– R_H : ensemble des relations taxonomiques ;
– R_A : ensemble des relations associatives.

L'ensemble de R_H représente la subsomption entre des concepts à travers le type *SorteDe* et entre un concept et ses instances à travers le type *EstUn*. L'ensemble de R_A représente l'ensemble des relations associatives qui décrivent des relations spécialisées de domaine entre des unités conceptuelles.

Les types de relations conceptuelles sont représentés au niveau du graphe sous-jacent au réseau conceptuel G_C par des étiquettes TRC portées au niveau des arcs.

Relations conceptuelles et contraintes associées Les relations conceptuelles explicitent différents types de liens pouvant exister entre des unités conceptuelles. Ces relations permettent, d'une part, la structuration des unités conceptuelles sous la forme d'un arbre et d'autre part la définition de liens associatifs entre ces dernières. Au niveau du graphe sous-jacent au réseau conceptuel G_C, les relations conceptuelles sont représentées par des triplets $rc(uc_i, trc, uc_j)$ tels que $uc_i \in C \cup I$ et $uc_j \in C \cup I \cup V$ sont des unités conceptuelles et $trc \in TRC$ est le type de la relation conceptuelle. Au relations conceptuelles sont associées des propriétés décrites dans le tableau 4.10.

Les relations conceptuelles sont associées à des contraintes :

– une relation de type *SorteDe* relit deux unités conceptuelles appartenant à l'ensemble des concepts de C ;

Propriété	Définition
ID	identifiant de la relation
Label	label correspondant
Type	type sémantique de la relation
Réflexivité	si relation est réflexive
Transitivité	si relation est transitive
Symétrie	si relation est symétrique

TABLE 4.10 – Les propriétés des relations conceptuelles RC.

- une relation de type *EstUn* relit deux unités conceptuelles avec le premier argument appartient à l'ensemble des instances de I et le deuxième appartient à l'ensemble des concepts de C ;
- une relation de type *RoleAssociatif* relit deux arguments avec le premier argument appartient à l'ensemble de CUI et le deuxième argument appartient à l'ensemble de $C \cup I \cup V$;
- entre deux concepts, s'il existe au moins une relation associative alors il n'existe pas de relation taxonomique.

Opérations

L'ingénieur de la connaissance peut effectuer des changements au sein du réseau conceptuel tout en respectant les contraintes appliquées à chacun des types d'objets conceptuels. Nous définissons les opérations suivantes :

- ajout/ suppression des unités conceptuelles : cette opération permet à l'ingénieur de la connaissance d'ajouter et de supprimer des unités conceptuelles à savoir des concepts, des instances et des valeurs. L'ajout d'un nouveau concept dans le réseau conceptuel déclenche la création d'une relation *SorteDe* qui relit ce dernier à un autre concept. La suppression d'un concept entraîne la suppression de ses instances mais l'inverse n'est pas vrai. La création d'une instance nécessite la création d'une relation *EstUn* qui relit cette dernière à un concept ;
- ajout/suppression des relations conceptuelles : cette opération permet d'ajouter et de supprimer des relations conceptuelles au sein du réseau conceptuel. La création d'une relation conceptuelle nécessite la création

des concepts reliés. La suppression d'une relation conceptuelle n'entraîne pas la suppression des concepts. La suppression des relations de types *SorteDe* et *EstUn* entraîne la suppression des relations héritées par le premier argument de chacune de ces deux types de relations ;
- mise à jour des propriétés : l'ingénieur de la connaissance peut modifier des propriétés relatives à des relations conceptuelles comme par exemple le label ou encore la propriété de transitivité de la relation.

Le choix du formalisme du modèle conceptuel dépend de l'application visée et du degré d'expressivité souhaité par l'ingénieur de la connaissance. Le niveau conceptuel est un niveau très contraint. Ces contraintes portent sur chaque type conceptuel (concept, instance, valeur, relation) et permettent de contrôler leur utilisation. D'un niveau des connaissances à un autre, les contraintes deviennent fortes jusqu'à l'obtention du niveau très contraint qu'est le niveau conceptuel.

Notre méthode propose un réseau terminologique normalisé qui va servir de point de départ pour la création du modèle ontologique. Dans la section suivante, nous définissons les correspondances qui peuvent exister entre les différentes structures des connaissances qui sont définies dans chacun des niveaux discours, terminologique, termino-conceptuel et conceptuel.

4.4 Correspondance entre les niveaux des connaissances

Dans la section précédente, nous avons décrit chacun des niveaux des connaissances séparément : les niveaux discours, terminologique, termino-conceptuel et conceptuel. Nous décrivons dans cette section comment naviguer, dériver les connaissances d'un niveau à partir de celles des autres niveaux et garder une trace de cette correspondance pour permettre par exemple de retrouver les différentes réalisations linguistiques des unités conceptuelles. L'ensemble de ces connaissances et les liens tissés entre elles constitue une ressource plus riche que chacun des niveaux considéré à part. Nous parlons d' « ontologie lexicalisée » parce que le niveau conceptuel est lié au niveau terminologique et discursif.

Dans cette section, nous décrivons les correspondances qui existent entre

les différents niveaux des connaissances :

- entre le niveau discours et le niveau terminologique ;
- entre le niveau terminologique et le niveau termino-conceptuel ;
- entre le niveau termino-conceptuel et le niveau conceptuel.

4.4.1 Entre les niveaux discours et terminologique

Le réseau terminologique décrit le vocabulaire mentionné dans le texte. Pour le construire, il faut passer de la description de flux de données vers une représentation en réseaux des éléments décrits dans le texte ainsi que des relations terminologiques les reliant. Au final, les marqueurs des unités terminologiques décrites au sein du réseau terminologique sont associés à leurs réalisations dans le texte. Il s'agit de l'ensemble des variantes des termes et de leurs formes fléchies, des entités nommées et des verbes ayant une valeur terminologique figurant dans le texte.

Une séquence d'unités textuelles décrite dans un corpus d'acquisition C peut correspondre à une ou plusieurs unités terminologiques et inversement (cardinalité de la correspondance est $n * n$). Ces réalisations linguistiques (des unités textuelles) correspondent aux marqueurs d'une ou de plusieurs unités terminologiques.

Nous avons décrit dans la section 4.3.2 que les relations terminologiques sont définies par des triplets $rt(ut_i, ut_k, ut_j)$ tels que ut_i, ut_k et ut_j sont des unités terminologiques. Une relation terminologique peut correspondre à plusieurs séquences d'unités textuelles et inversement. La figure 4.10 décrit les liens de correspondance entre le niveau discours et le niveau terminologique.

Prenons l'exemple de l'unité terminologique *AAdvantage platinum elite member* qui correspond à la séquence des unités textuelles *AAdvantage*, *platinum*, *elite* et *member*. L'unité textuelle *AAdvantage* correspond aussi à plusieurs unités terminologiques comme par exemple *AAdvantage member*, *AAdvantage airline* et *AAdvantage program* au niveau terminologique.

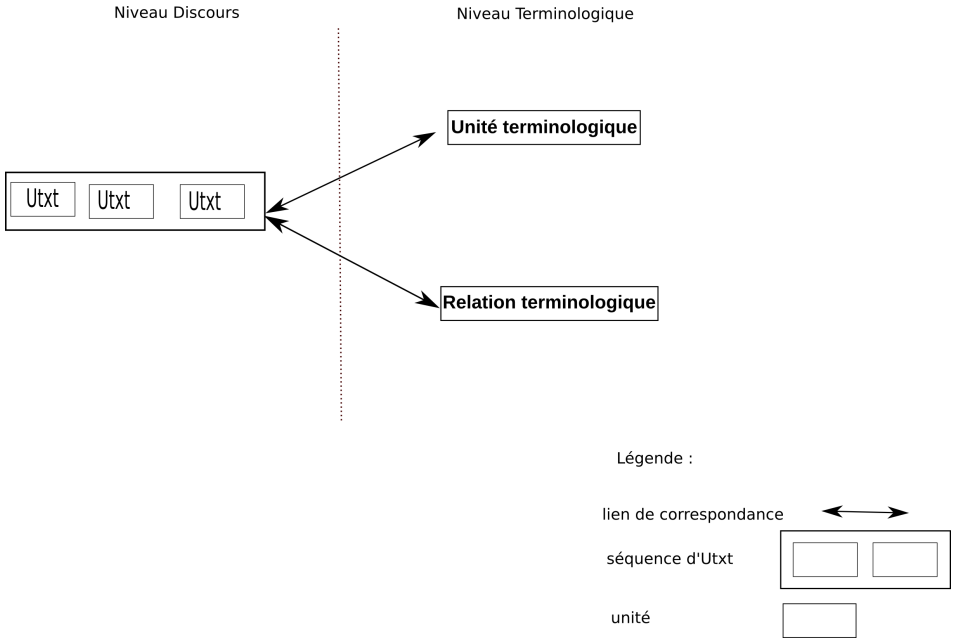

FIGURE 4.10 – Correspondance entre les niveaux dicours et terminologique.

4.4.2 Entre les niveaux terminologique et termino-conceptuel

Il existe aussi une correspondance $n * n$ entre des termino-concepts et des unités terminologiques. Un termino-concept peut correspondre à 0 ou plusieurs unités terminologiques si ces dernières sont synonymes. Par exemple, les unités terminologiques *Elite status member* et *Elite member* désignent tous les deux un AAdvantage member et sont associés au même termino-concept *AAdvantage member*. Réciproquement une unité terminologique peut correspondre à plusieurs termino-concepts si elle décrit, dans le corpus d'acquisition, plusieurs sens pertinents au domaine à modéliser tel que chaque sens pertinent correspond à un termino-concept (voir figure 4.11). Par exemple, l'unité terminologique *AAdvantage member* correspond à deux termino-concepts. Le premier termino-concept *AAdvantage member* décrit une catégorie spécifique de passagers adhérant dans le programme de fidélité. Le deuxième termino-concept *Airline participant* décrit une compagnie aérienne qui participe au même programme de fidélité. Un terme peut aussi n'avoir aucun correspondant au niveau termino-conceptuel : cela signifie qu'il n'est pas retenu comme suffisamment pertinent pour le domaine. Les liens entre les termino-concepts

et les unités terminologiques sont de cardinalité $n * n$.

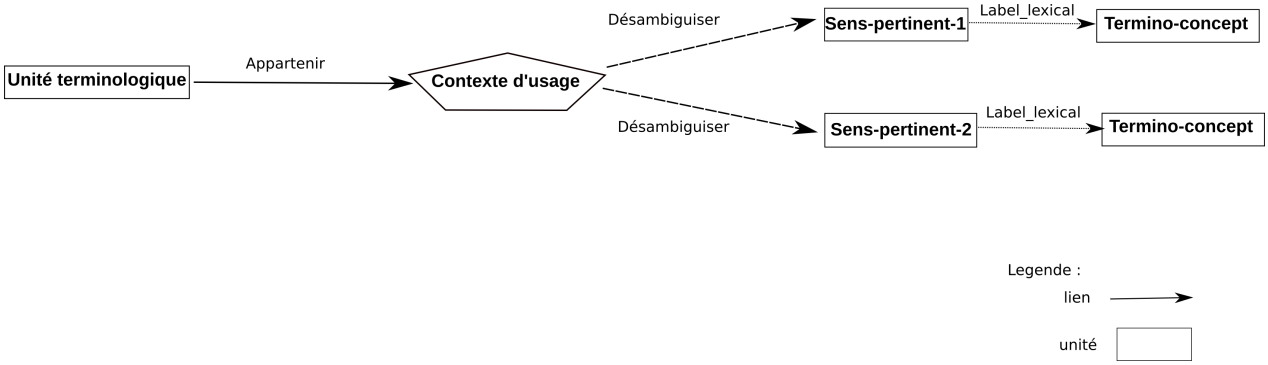

FIGURE 4.11 – Liens entre unités terminologiques et termino-concepts.

Une unité terminologique peut également correspondre à un type de relation termino-conceptuelle. Par exemple, l'unité terminologique *mileage credit* correspond au type termino-conceptuel *MileageCredit* qui relit les termino-concepts *AAdvantage mileage* et *AAdvantage account*. On a aussi des correspondances entre des relations terminologiques et des relations termino-conceptuelles. Dans ce cas, chacun des éléments de la relation terminologique, le triplet (ut_i, ut_k, ut_j) est en correspondance avec une relation termino-conceptuelle, le triplet (tc_i, trc, tc_j). Un exemple de cette correspondance est le triplet *(AAdvantage member, AdheresIn, AAdvantage program)*. Ces exemples sont décrits dans la figure 4.12.

Une relation termino-conceptuelle de type *Généricité/Spécificité* peut correspondre aux relations terminologiques de types *APourTête/APourModifieur* et *Hyperonymie*. Une relation termino-conceptuelle de type *RelationAssociative* peut correspondre à une ou plusieurs unités terminologiques si ces dernières décrivent un même sens. La figure 4.13 décrit les liens de correspondance entre le niveau terminologique et le niveau termino-conceptuel.

La correspondance entre les réseaux terminologique et termino-conceptuel enrichit ce dernier niveau par des informations linguistiques qui permettent de garder le lien entre le modèle sémantique et le corpus d'acquisition. Les liens de correspondance entre les niveaux discours et terminologique puis entre les niveaux terminologique et termino-conceptuel permettent de relier indirectement les occurrences textuelles aux unités termino-conceptuelles, mais le

FIGURE 4.12 – Exemple de correspondance entre les réseaux terminologique et termino-conceptuel.

niveau terminologique étant ambigu, il est aussi possible de relier directement les unités textuelles aux unités termino-conceptuelles de manière à avoir des liens de correspondance désambiguïsés. La figure 4.14 décrit les liens de correspondance entre le niveau discours et le niveau termino-conceptuel.

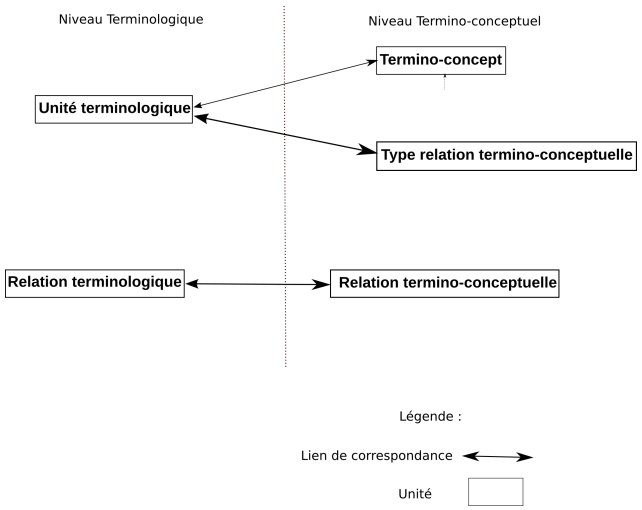

FIGURE 4.13 – Correspondance entre les niveaux terminologique et termino-conceptuel.

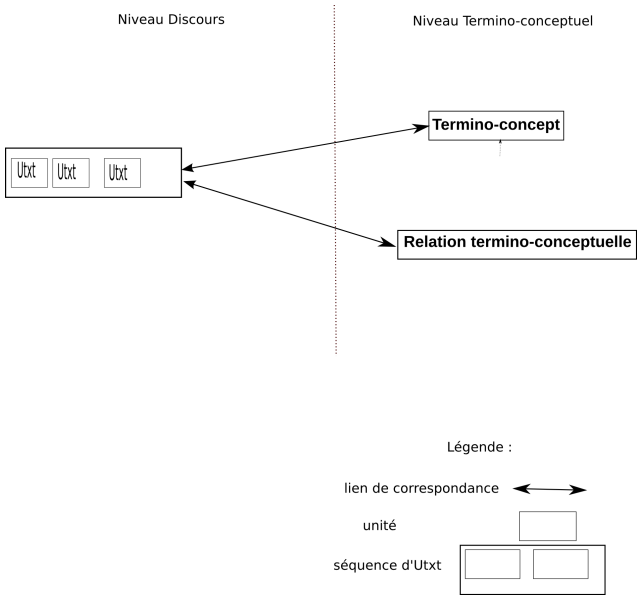

FIGURE 4.14 – Correspondance entre les niveaux discours et termino-conceptuel.

4.4.3 Entre les niveaux termino-conceptuel et conceptuel

Le niveau conceptuel est une représentation simplifiée d'une ontologie. Il spécifie l'ensemble des concepts et des instances du domaine, des valeurs ainsi que les relations les reliant. La correspondance entre le niveau termino-

conceptuel et le niveau conceptuel permet de relier une terminologie normalisée aux unités et aux relations conceptuelles du réseau conceptuel. La figure 4.15 décrit ces liens de correspondance.

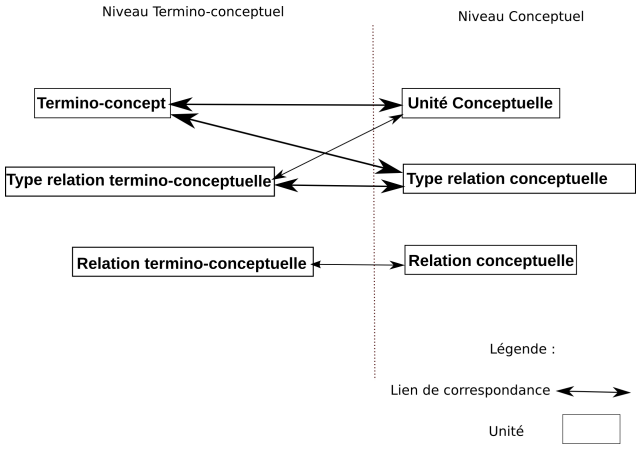

FIGURE 4.15 – Correspondance entre les niveaux termino-conceptuel et conceptuel.

Il n'y a pas de correspondance stabilisée entre des unités termino-conceptuelles et celles conceptuelles. Un termino-concept peut par exemple correspondre à un concept ou à une instance de concept. Dans notre modélisation, le termino-concept *Airline participant* correspond ainsi au concept *Airline participant* et le termino-concept *Japan airlines* correspond à l'instance *Japan airlines*. Un termino-concept peut aussi correspondre à un type de relation conceptuelle. Prenons l'exemple du termino-concept *Partner* qui correspond au type conceptuel *PartnerOf*. Un termino-concept peut encore correspondre à un type de relation conceptuelle. Par exemple, le termino-concept *AAdvantage number* correspond à un type de relation conceptuelle *AAdvantage number*, qui est défini comme type de la relation entre le concept *AAdvantage member* et la valeur *Double* dans l'ontologie que nous avons construit pour le cas d'usage de *American Airlines*.

Un type de relation termino-conceptuelle *Généricité/Spécificité* peut correspondre aux types de relations conceptuelles *EstUn* ou *SorteDe* au niveau conceptuel tels que nous les avons définis dans la section 4.3.4. Prenons l'exemple du type termino-conceptuel *Généricité/Spécificité* qui relit les termino-concepts *AAdvantage member* et *AAdvantage Gold member* et qui correspond au type de relation conceptuelle *SorteDe* reliant les concepts *AAdvantage member* et *AAdvantage Gold member*. Le type termino-conceptuel

Généricité/Spécificité qui relit les termino-concepts *AAdvantage Gold member* et *Ruby* correspond au type de relation conceptuelle *EstUn* qui relit le concept *AAdvantage Gold member* à l'instance *Ruby*.

Il s'agit de faire des choix de modélisation pour ce qui concerne les différents types de correspondance concernant les types de relations. Un type de relation termino-conceptuelle *RelationAssociative* peut correspondre à un type de relation conceptuelle appartenant à l'ensemble des relations associatives (R_A) et noté par *RoleAssociatif* ou à une unité conceptuelle. Par exemple, le type de relation termino-conceptuelle *AdheresIn* correspond au type de relation conceptuelle *AdheresIn* et le type de relation termino-conceptuelle *FlightsOn* correspond au concept *Flight*.

Une relation termino-conceptuelle de type *RelationAssociative* correspond à une relation conceptuelle de type *RoleAssociatif* au niveau conceptuel. C'est-à-dire que les arguments formant la relation termino-conceptuelle sont des unités conceptuelles au niveau conceptuel. Le triplet *(AAdvantage member, Earns, Benefit)* correspond par exemple au triplet *(AAdvantage member, Earns, Benefit)* au niveau conceptuel. Ces exemples sont décrits dans la figure 4.16.

Afin d'assurer la correspondance entre les deux niveaux, nous définissons ces contraintes :

- un concept ne peut correspondre qu'à un seul et un seul termino-concept ;
- un rôle conceptuel peut être relié à 0 ou plusieurs types de relations termino-conceptuelles ;
- une instance de concept ne peut être liée qu'à un et un seul termino-concept.

L'évolution de l'ontologie en termes de création de concepts, d'instances et de relations conceptuelles n'entraîne pas la suppression des liens de correspondance entre des unités termino-conceptuelles et celles qui sont modifiées au niveau conceptuel. Nous définissons un patron conceptuel comme un sous graphe sous-jacent à l'ontologie constitué par un sous ensemble d'unités conceptuelles $UC_1 \in UC$ reliées par un sous ensemble de relations conceptuelles $RC_1 \in RC$. Dans certain cas, quand la cible d'un lien de correspondance au niveau conceptuel se trouve décomposée en un sous-graphe conceptuel (plusieurs unités conceptuelles reliées entre elles) le lien de correspon-

FIGURE 4.16 – Exemple de correspondance entre les réseaux termino-conceptuel et conceptuel.

dance se retrouve pointé vers l'ensemble du sous-graphe. Nous désignons le nom du graphe issu de la décomposition d'un élément conceptuel par le terme de « patron conceptuel ». Prenons l'exemple du termino-concept *AAdvantage member* qui décrit toutes les différentes catégories des membres adhérants dans le programme de fidélité de AA. L'évolution de la politique d'affection de miles définie par la compagnie American airlines dans le cas d'usage de *American Airlines* a fait qu'il faut distinguer dans l'ontologie les membres qui n'ont pas encore atteint un solde de 25 000 miles de ceux qui l'ont atteint. De ce fait, le concept *AAdvantage member* a deux concepts fils *Elite member* et *Passenger*. Le premier concept décrit toutes les catégories des membres ayant 25 000 miles et plus dans leurs comptes. Le concept *Passenger* décrit la caté-

gorie des membres adhérant dans le programme de fidélité et qui ont gagné moins de 25 000 miles. La figure 4.17 décrit le résultat de cette évolution qui se traduit par un lien de correspondance entre un termino-concept *AAdvantage member* et un patron conceptuel.

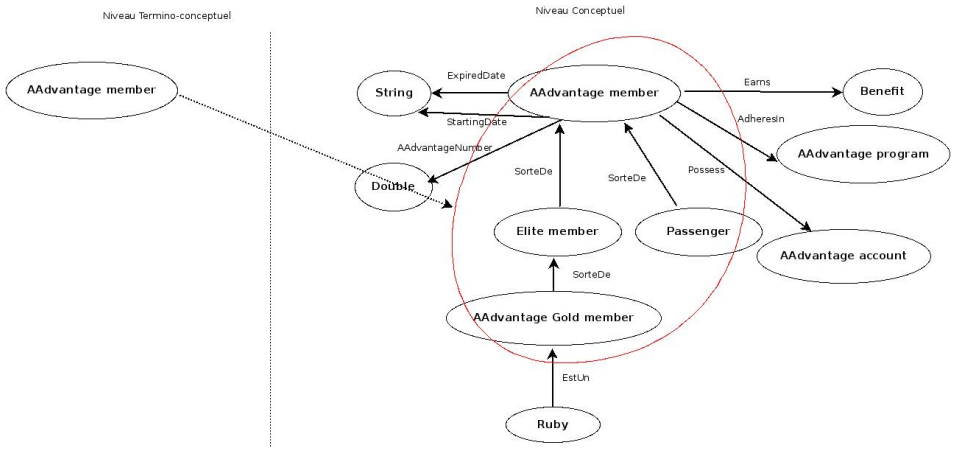

FIGURE 4.17 – Lien de correspondance entre les niveaux termino-conceptuel et conceptuel résultant de la décomposition du concept *AAdvantage member* en un patron conceptuel.

4.5 Conclusion

Dans la première section de ce chapitre, nous avons défini la notion de réseau sémantique et les opérations qui s'appliquent sur les différentes structures de connaissances. Cela nous a permis d'introduire les concepts sur lesquels se basent les structures de connaissances que nous utilisons. Dans la deuxième section, nous avons présenté les différents niveaux de connaissances, à savoir le niveau discours, terminologique, termino-conceptuel et conceptuel et les structures des connaissances existantes à chaque niveau. Dans la troisième section, nous avons décrit les différentes correspondances existant entre ces niveaux de connaissance.

Nous remarquons que les différents niveaux introduisent des structures de connaissances différentes ayant des degrés d'expressivité plus ou moins contraint. Le modèle obtenu au niveau conceptuel peut se construire par dérivation progressive du niveau terminologique à partir du texte du niveau de discours, du niveau termino-conceptuel à partir du niveau terminologique et

enfin du niveau conceptuel à partir du niveau termino-conceptuel. Le premier niveau est construit automatiquement par les outils de TAL. Le niveau intermédiaire (termino-conceptuel) est dérivé du niveau terminologique grâce aux opérations de normalisation du réseau terminologique. Le dernier niveau (conceptuel) est dérivé du niveau termino-conceptuel à travers des choix de modélisation faits par l'ingénieur de la connaissance.

Plusieurs difficultés sont rencontrées au niveau du passage entre les niveaux de connaissances terminologique, termino-conceptuel et conceptuel. Cela est dû à la différence entre les connaissances manipulées. Dans ce chapitre, nous avons essayé de formaliser les différentes structures de connaissances définies dans chacun des niveaux de connaissances de la méthode TERMINAE et nous avons défini les différents liens de correspondance pouvant exister entre ces structures de connaissance. Dans le chapitre suivant, nous présentons notre approche qui se fonde sur la méthode TERMINAE.

Chapitre 5

Méthode de normalisation d'un réseau terminologique : GraphOnto

Sommaire

5.1 Introduction

L'objectif de notre méthode est de construire un réseau terminologique normalisé appelé réseau termino-conceptuel à partir du réseau terminologique. Le réseau terminologique normalisé peut ensuite servir de ressource sémantique pour l'annotation des documents ou pour la construction d'ontologies lexicalisées de domaine. Notre méthode GraphOnto entre dans le cadre général de la construction d'ontologies de domaine à partir de textes et plus particulièrement dans le cadre de la méthode terminae. Plus spécifiquement, nous proposons une méthode de normalisation du réseau terminologique qui constitue la deuxième phase du processus de conceptualisation d'ontologies. Ce dernier est défini par trois phases : l'extraction terminologique, la normalisation et la formalisation en un modèle formel (ontologie). La conceptualisation étant l'identification et la mise en relation des concepts représentatifs du domaine à modéliser. Notre méthode s'intègre au sein de la méthode

TERMINAE pour assurer le passage du niveau terminologique vers le niveau termino-conceptuel. Les différents niveaux de structures de connaissances terminologique, termino-conceptuel et conceptuel sont définis dans la méthode TERMINAE (voir chapitre 4).

GRAPHONTO s'appuie sur les structures des connaissances introduites dans le chapitre 4. Nous avons montré que les connaissances manipulées sont de natures différentes (vocabulaire terminologique *vs.* vocabulaire normalisé), que la structure du réseau termino-conceptuel est plus contraint que la structure du réseau terminologique mais qu'il existe des « liens de correspondance » $n*n$ entre ces structures, ce qui empêche de dériver automatiquement le niveau termino-conceptuel du niveau terminologique. En effet, comme les unités et les relations terminologiques sont extraites du texte, ces dernières sont souvent ambiguës, parfois redondantes. De plus, la prise en compte de l'utilisation finale du modèle et la définition même d'un modèle conceptuel impose à l'ingénieur de la connaissance de faire des choix de modélisation. Dans notre méthode, l'ingénieur de la connaissance intervient durant tout le processus de normalisation pour sélectionner des unités et des relations à normaliser ou à mettre à jour. Notre méthode définit les opérations nécessaires pour assurer cette construction.

Notre méthode repose sur un réseau terminologique formé par le matériau linguistique extrait à partir du corpus d'acquisition par des extracteurs de termes, de reconnaissance d'entités nommées et des outils d'extraction de relations terminologiques. GRAPHONTO définit le travail de normalisation du réseau terminologique pour la construction d'un réseau termino-conceptuel qui peut être exporté en SKOS [1].

Ce chapitre est divisé en cinq sections. La première section présente la structure du réseau terminologique extrait à partir d'un corpus d'acquisition et servant de point de départ à notre méthode. Dans la deuxième section, nous exposons les difficultés que soulève la normalisation d'un réseau terminologique. La troisième section définit la structure du réseau obtenu à la fin de la normalisation et les opérations élémentaires de transformation. Dans la quatrième section, nous décrivons comment les opérations élémentaires de transformation s'enchaînent et les différents indices de contrôle qui sont mis en place pour guider et vérifier le travail de normalisation et la navigation au

1. http ://www.w3.org/TR/2009/REC-skos-reference-20090818/

sein du réseau terminologique. Enfin, avant de conclure ce chapitre, la cinquième section expose les principaux cas particuliers de la normalisation des éléments constituant le réseau terminologique.

5.2 Point de départ : le réseau terminologique

La première étape de notre méthode GraphOnto est la construction d'un réseau terminologique qui décrit un vocabulaire de domaine tel qu'il est mentionné dans le texte. Cette étape repose sur une extraction automatique du matériau linguistique à partir du corpus d'acquisition. Le réseau terminologique $G_T(UT, RT)$ tel qu'il a été défini dans le chapitre 4 est constitué par un ensemble d'unités terminologiques (UT) qui représentent des termes et des entités nommées et un ensemble de relations terminologiques (RT) qui relient ces unités. Nous expliquons, tout d'abord, l'exploitation des termes pour la création des unités terminologiques. Ensuite, nous justifions l'intérêt de considérer les entités nommées au niveau du réseau terminologique. Enfin, nous décrivons l'ajout de relations terminologiques pour relier des unités terminologiques.

5.2.1 Termes

Dans cette section, nous décrivons le résultat de l'extraction terminologique par des outils de TAL à partir d'un corpus d'acquisition. Nous justifions l'intérêt de considérer les termes pour la construction d'un réseau terminoconceptuel. Nous présentons d'une manière générale la méthode d'extraction des termes et plus spécifiquement l'outil que nous avons utilisé. Enfin, nous présentons des exemples de termes extraits à partir de nos corpus.

Comme nous l'avons exposé dans le chapitre 2, l'utilisation des termes dans la littérature est relative à la création de concepts d'une ontologie à construire ou à la création du vocabulaire normalisé d'un thésaurus. En effet, les termes décrivent des notions partagées dans une communauté et un domaine spécifique. Les termes forment un vocabulaire spécialisé ayant généralement une sémantique stable dans un contexte précis.

Nous nous inscrivons dans la même branche que les travaux cités dans le chapitre 2. Nous considérons que l'exploitation des termes durant le processus

de normalisation permet la création des termino-concepts ou des types de relations termino-conceptuelles.

Il existe divers outils permettant d'extraire des termes à partir d'un corpus d'acquisition. Nous ne faisons pas ici un inventaire des outils existants [2]. Indépendamment de l'évaluation des outils d'extraction de termes qui a été faite (Mondary *et al.*, 2012), nous avons opté pour l'extracteur de termes *YaTeA* qui a été développé au LIPN et qui était utilisé avant notre arrivée. Dans la suite, nous décrivons les principes de l'extraction terminologique.

Les méthodes d'extraction des termes s'appuient sur le texte (propriétés linguistiques et statistiques) pour extraire des termes candidats ainsi que des relations syntaxiques les reliant. L'analyse terminologique permet la détection de termes dits « termes candidats » susceptibles de désigner des *notions du domaine* du fait de leurs propriétés syntaxiques. Notre méthode s'appuie sur l'analyse terminologique du corpus d'acquisition qui permet d'extraire le vocabulaire terminologique associé à un domaine spécifique. Les termes extraits sont souvent reliés entre eux par des relations syntaxiques (tête/modifieur). Prenons l'exemple du terme *airline* qui joue le rôle de modifieur pour les termes *airline participant, airline participant route, airline ticket* et *airline travel award*. Le même terme joue le rôle de tête pour les termes *AAdvantage participant airline, American airline, member airline* et *oneworld member airline*.

L'extracteur de termes candidats *YaTeA* (Yet Another Term ExtrActor) (Aubin & Hamon, 2006) [3] prend en entrée un corpus lemmatisé et donne en sortie une liste de termes candidats ainsi que leurs dépendances syntaxiques. *YaTeA* utilise comme entrée un fichier comportant pour chacun des mots extraits du texte sa catégorie grammaticale (par ex. nom, verbe, adjectif) et sa forme lemmatisée. Dans cette thèse, nous utilisons comme étiqueteur morpho-syntaxique l'outil *TreeTagger* (en anglais *Part of speech tagging* ou *POS tagging*). *TreeTagger* (Schmid, 1995) prend en entrée un texte brut qu'il découpe en phrases et en mots « tokens ». Il annote chaque mot appartenant à une phrase avec sa classe morpho-syntaxique (catégorie grammaticale, genre,

2. Divers outils sont cités dans l'ouvrage de (Cimiano, 2007).

3. YaTeA est un extracteur de termes développé au LIPN. http ://search.cpan.org/ thhamon/Lingua-YaTeA/ qui fait lui-même appel à TreeTagger (http ://www.ims.uni-stuttgart.fr/projekte/corplex/TreeTagger/).

etc.) et lui associe un lemme (la forme canonique du mot). Parfois, il y a des erreurs d'étiquetage des mots. Par exemple, l'outil ne fait pas de différence entre le point de ponctuation et le point d'abréviation (par ex. AA.airlines). La figure 5.1 décrit le processus d'extraction des termes.

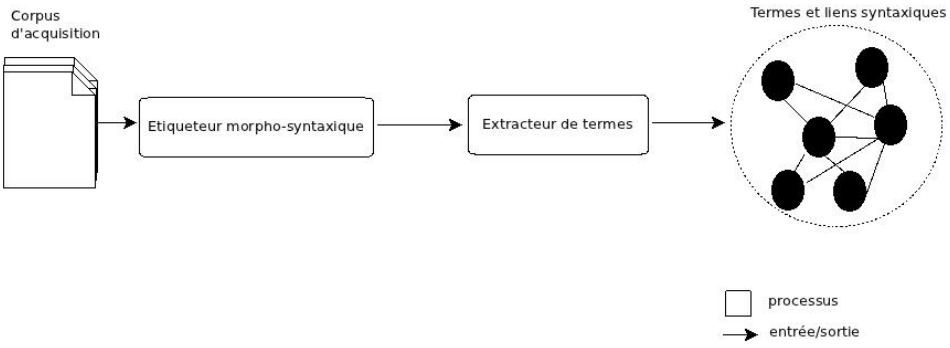

FIGURE 5.1 – Le processus d'extraction des termes.

Nous illustrons l'extraction terminologique par des exemples tirés des cas d'usage utilisés pour mener nos expérimentations. Le premier corpus est celui du cas d'usage de *American Airlines* qui contient environ 5 744 mots, 973 termes candidats ont été extraits par l'outil *YaTeA*. Un autre exemple, celui du corpus de *Audi* qui contient 3 704 mots, 1 003 termes candidats ont été extraits. Le tableau 5.1 présente les résultats d'extraction des termes pour les corpus AAdvantage et Audi.

Corpus	**# termes candidats trouvés**
AAdvantage	973 (5 744 mots)
Audi	1 003 (3 704 mots)

TABLE 5.1 – Nombres de termes extraits pour les cas d'usage AAdvantage et Audi.

Parfois un corpus comporte des fautes d'orthographe ou n'est pas nettoyé ce qui influence le résultat. Dans ce cas, les outils de TAL extraient des termes qui sont mal formés. Le tableau 5.2 présente quelques exemples de termes candidats mal formés du cas d'usage de AAdvantage.

L'extraction terminologique extrait des termes candidats qui peuvent être synonymes (Hamon, 2000). Ce sont généralement des variantes du même terme ou des termes reliés entre eux par une relation de type *Synonymie* (si ce type

Caractères incorrects (50 terms)	**Adjectifs** (64 terms)	**Mal formés** (161 terms)
administrative Ã©	additional	action cannot
admiral ClubÂ® member	actual	act and
S mileage accruement PlatinumÂ®	comparable	of mileage credit
check machineˆ	complete	accumulation
MalÃ©v Hungarian	complimentary	Per

TABLE 5.2 – Exemple de résultats bruités suite à l'extraction terminologique à partir du corpus AA.

de relation a été identifié dans le texte par un outil d'extraction de relations terminologiques). Le tableau 5.3 décrit quelques exemples de termes qui sont synonymes.

Dans le réseau terminologique $G_T(UT, RT)$, les termes candidats sont représentés par des unités terminologiques appartenant à l'ensemble UT. Les variantes d'un terme candidat sont associées comme *Marqueurs* de ce terme. Les occurrences d'un terme candidat dans le texte sont associées à la propriété *Occurrence* de l'unité terminologique correspondante. Les relations syntaxiques qui relient les termes candidats figurent parmi les relations terminologiques (RT) qui relient les unités terminologiques correspondantes. Dans le cas où il existe une ressource sémantique (par ex. un thésaurus de domaine), la propriété *Type sémantique* correspond au type sémantique de l'unité décrite dans la ressource sémantique. Le tableau 5.4 décrit les propriétés de l'unité terminologie *Elite status Addvantage member*.

Généralement les extracteurs de termes produisent des résultats bruités ou trop volumineux. Ils doivent être filtrés pour donner un résultat interprétable humainement. Par exemple, à partir du corpus AAdvantage, 973 termes candidats ont été extraits par l'outil *YaTeA* et plusieurs de ces termes candidats ne sont pas pertinents pour le domaine. Pour filtrer le résultat, nous avons supprimé par exemple des adjectifs, des termes mal-formés et nous avons regroupés des variantes de termes. Environ 30% des termes candidats ont été ainsi supprimés. Nous avons obtenu une nouvelle liste de 680 termes candidats. Parfois, les extracteurs de termes passent sous silence certains termes qui s'avèrent pertinents pour le domaine (par ex. l'emploi des patrons spécifiques

Terme	Synonymes/Variantes
Airline participant	*AAdvantage airline participant, Airline Representative, AA Airlines, member airline, oneword alliance member, oneworld member airline, participant*
Elite status	*Elite status member, Elite status aadvantage member*
AAdvantage mileage credit	*AAdvantage mileage accruement, AAdvantage mileage credit*
AAdvantage number	*AA number*
Account	*AAdvantage account*
Passenger	*Member*
Fare ticket	*Full fare ticket*

TABLE 5.3 – Exemples de termes et de leurs synonymes relatifs au cas d'usage de AAdvantage.

Propriété	Valeur
Label	Elite status AAdvantage member
Type sémantique	
Marqueur	Elite status AAdvantage member, Elite status members
Occurrence	All elite status AAdvantage members enjoy complimentary upgrades to the next class of service. Elite status members may earn a minimum of 500 AAdvantage base miles for every eligible flight.

TABLE 5.4 – Exemple des propriétés de l'unité terminologique *Elite status AAdvantage member*.

ne permet pas d'extraire toutes les mentions linguistiques d'un terme donné). Dans le cas d'usage de AAdvantage, par ex., le terme *itinerary* qui décrit la

distance minimale entre deux points de vol n'a pas été extrait. Cette distance correspond à un bonus égal à 500 miles. Ce terme n'apparaît qu'une seule fois dans le texte mais il est pertinent pour le cas d'usage de AAdvantage. L'occurrence du terme *itinerary* est mentionnée dans l'extrait tiré du cas d'usage de AAdvantage suivant :

AAdvantage members checking-in for an itinerary that requires 500-mile upgrades will be able to purchase the upgrades they need to complete their transaction at the Self-Service Check-In machine.

5.2.2 Entités nommées

Dans cette section, nous décrivons le résultat de l'identification des entités nommées par des outils de reconnaissance d'entités nommées (REN) à partir d'un corpus d'acquisition. Nous justifions l'intérêt de considérer les entités nommées pour la construction d'un réseau termino-conceptuel. Nous présentons d'une manière générale le principe de l'identification des entités nommées et plus spécifiquement l'outil que nous avons utilisé. Enfin, nous citons quelques exemples des résultats obtenus sur nos corpus.

Nous avons défini dans le chapitre 2 la notion d'entité nommée et les domaines d'application qui exploitent ce genre d'unité linguistique. Les entités nommées sont des mentions qui désignent de manière univoque des entités référentielles dans un contexte donné. Ce sont des unités textuelles qui renvoient à des « entités » du domaine et qui peuvent relever de différentes catégories linguistiques (des noms propres, des pronoms, etc.). Cet aspect référentiel fait que ce type d'unité linguistique a suscité de l'intérêt dans le domaine de l'ingénierie des connaissance. Les entités nommées sont généralement utilisées pour le peuplement des ontologies de domaine une fois que celles-ci sont construites.

Les termes et les entités nommées sont des types d'unités textuelles qui jouent chacun un rôle particulier par rapport au domaine. Ils ont un fonctionnement sémantique différent : les termes reflètent des notions alors que les entités nommées renvoient à des objets ou référents. De plus, généralement elles sont moins nombreuses que les termes dans les textes et expriment une valeur référentielle (elles font référence à des entités particulières).

Le fait que les entités nommées soient largement utilisées pour le peuplement d'ontologies ne signifie pas pour autant qu'elles soient à négliger pour

la création de concepts. Le fait que la conceptualisation repose traditionnellement et prioritairement sur l'exploitation des termes n'interdit pas d'exploiter d'autres indices textuels. Contrairement aux approches qui tendent à peupler les ontologies en dérivant automatiquement un type ontologique (instance) à partir de sa catégorie linguistique (entité nommée), notre méthode propose qu'à partir des entités nommées l'ingénieur de la connaissance crée des termino-concepts et des types de relations termino-conceptuelles. La prise en compte des entités nommées se situe au niveau terminologique et leur normalisation éventuelle au niveau termino-conceptuel.

De plus, dans la littérature, les entités nommées, avec leur sémantique référentielle, sont considérées comme des éléments clés du domaine et les relations qu'elles entretiennent sont aussi considérées comme des relations de domaine (Velardi *et al.*, 2006; Tanev & Magnini, 2008; Cimiano, 2006). Notre approche s'appuie sur cette même idée que les entités nommées sont des marqueurs de domaine mais pour la détection des concepts plutôt que des relations. Nous faisons en effet l'hypothèse que l'ancrage référentiel des entités nommées confère une valeur sémantique particulière à leurs contextes. Sous l'hypothèse que « les termes du domaine n'apparaissent pas seuls ou d'une manière arbitraire » (Maynard & Ananiadou, 1999), nous considérons que la valeur référentielle des entités nommées « déteint » sur leur contexte et qu'elles jouent le rôle de marqueurs de domaine pour les termes qui figurent dans leur voisinage. Elles peuvent ainsi aider à repérer les termes les plus pertinents pour un domaine donné. Nous considérons la détection des entités nommées pour le calcul d'un indice de pertinence attribué aux termes qui figurent dans les mêmes phrases que des entités nommées. Notre proposition sera évaluée dans le chapitre 6.

De nombreux outils de reconnaissance d'entités nommées ont été mis au point à la faveur des travaux sur l'extraction d'information et des campagnes d'évaluation. Ils reposent pour certains sur des règles d'extraction qui expriment un faisceau de contraintes (présence de certains marqueurs ou de certaines catégories syntaxiques, ordre des mots, etc.) couplées avec des dictionnaires ou à des catégories sémantiques. Les règles servent à reconnaître des formes variantes des entités des dictionnaires (*Roland Garros vs. R. Garros*) ou à découvrir de nouvelles entités d'un type donné [4]. Ces outils extraient des

4. Le fragment « Monsieur Roland Garros » permet d'identifier *Roland Garros* comme

entités nommées et leurs associent à chacune un ou plusieurs types sémantiques. La figure 5.2 décrit le processus d'extraction des entités nommées.

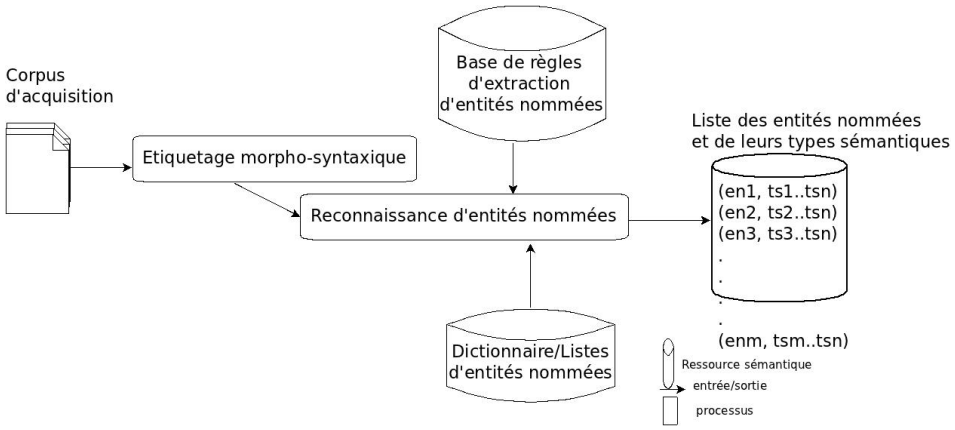

FIGURE 5.2 – Le processus d'extraction des entités nommées.

Les outils de reconnaissance d'entités nommées ont cependant leurs limites : ils ne reconnaissent que les mentions de type « noms propres » et avec une part d'erreur (Nadeau & Sekine, 2007). Par ailleurs, le repérage des entités nommées est guidé par les types sémantiques considérés comme pertinents et sélectionnés en fonction du domaine considéré (Ehrmann, 2008).

Nous avons néanmoins cherché à apprécier leur utilité pour le processus de normalisation. Nous avons pour cela choisi d'utiliser la plate-forme *Gate*. *Gate* est une plate-forme d'ingénierie linguistique qui permet d'appliquer toute sorte de traitements sur des textes, notamment un module de reconnaissance d'entités nommées appelé *Annie*[5]. De nombreux types d'entités peuvent être pris en compte. Sur nos corpus, *Annie* reconnaît les types d'entités nommées DATE, ADDRESSE, PERSONNE, EMPLOI (JOB TITLE), LOCALISATION et ORGANISATION. *Annie* utilise un composant appelé « Gazetter » qui manipule un ensemble de listes comportant chacune des noms d'entités nommées. Pour chacune des listes, un type sémantique principal est renseigné ainsi qu'un type optionnel. En utilisant ces listes couplées à des règles de grammaire, *Gazetter* détermine quel type associer à une entité identifiée dans le texte. Lorsque *Annie* ne reconnaît pas le type de certaines entités nommées, il leur associe le type sémantique AUTRES TYPES (UNKOWN). Le plus souvent ce type

une personne si celle-ci n'est pas déjà connue.

5. http ://gate.ac.uk/sale/tao/splitch6.html#chap :annie.

correspond à des entités relatives au domaine.

Nous illustrons la reconnaissance des entités nommées par des exemples tirés des cas d'usage utilisés pour mener nos expérimentations. Il y a des corpus factuels qui sont riches en entités nommées et d'autres qui sont « pauvres » en entités nommées. Nous n'exploitons pas les entités nommées de la même manière dans l'un et l'autre des cas (voir chapitre 6). Le tableau 5.5 présente les résultats d'extraction des entités nommées pour les corpus AAdvantage et Audi.

Corpus	**# EN trouvées**
AAdvantage	105 (5 744 mots)
Audi	90 (3 704 mots)

TABLE 5.5 – Nombres d'entités nommées extraits pour les cas d'usage AAdvantage et Audi.

Le corpus de AAdvantage n'est pas riche en entités nommées, néanmoins il en comporte 105 parmi lesquelles *American airlines*, *AAdvantage platinum member* (qui décrit un membre bénéficiant des avantages de niveau « platine ») ou encore *February 20, 2011*. Le tableau 5.6 indique le nombre d'entités nommées différentes et le nombre d'occurrences reconnues par *Annie* en détaillant les principaux types pour le cas d'usage de AAdvantage.

Type sémantique	**Annie**
DATE	13 (32)
PERSONNE	3 (15)
LOCALISATION	28 (58)
ORGANISATION	30 (314)
Autres types	31 (332)

TABLE 5.6 – Résultats de *Annie* en nombre d'entités nommées (le nombre d'occurrences est indiqué entre parenthèses) pour le cas d'usage de AAdvantage.

Nous avons en parallèle procédé à une analyse manuelle du corpus de AAdvantage pour déterminer quelles entités nommées, quelles occurrences et quels types d'entités sont pertinents à prendre en compte pour le cas d'usage

de AAdvantage. Cette analyse nous a permis d'évaluer les résultats fournis par *Annie* dans la perspective qui est la nôtre, celle de la construction d'un réseau termino-conceptuel décrivant un vocabulaire normalisé utilisé pour l'écriture des règles métier de AA. Nous avons évalué les entités nommées extraites par *Annie* avec celles qui sont mentionnées dans des passages réglementaires. Le tableau 5.7 présente ces résultats en termes de précision et de rappel pour chaque type d'entités nommées.

Type sémantique	**#EN pert. trouvées**	**#EN pert. non trouvées**	**#EN pert.**	**Précison**	**Rappel**	**EN mal classées**
DATE	13	1	14	1	0.928	0
PERSONNE	3	4	7	0.75	0.428	0
LOCALISATION	28	0	28	0.933	1	2
ORGANISATION	30	0	30	1	1	0

TABLE 5.7 – Évaluation des résultats de reconnaissance des entités nommées : précision (P), rappel (R) et nombre d'entités nommées (EN) mal classées pour le cas d'usage de AAdvantage.

Cette expérience reste modeste par sa taille mais on voit que les valeurs de précision et rappel sont globalement assez satisfaisantes, en tout cas pour les types DATE, LOCALISATION et ORGANISATION. Les entités nommées extraites sont globalement pertinentes pour le domaine et l'application visée (les systèmes de gestion des règles métier) même s'il reste un travail de sélection à faire. Dans ce contexte, on tend à privilégier le rappel sur la précision, et c'est donc le faible taux de rappel du type PERSONNE qui pose problème : *Annie* n'identifie pas comme entités nommées certaines mentions de personnes comme *AAdvantage Gold member*. Nous remarquons enfin des erreurs de classement pour certains types d'entités nommées : on trouve ainsi classés comme LOCALISATION, *Miles*, qui est en fait une unité de mesure, et *Ruby*, qui désigne une catégorie particulière de passagers.

Nous proposons de prendre en considération les entités nommées dans le réseau terminologique $G_T(UT, RT)$ en les modélisant comme des unités terminologiques (UT). Les types sémantiques sont conservés comme propriété

Type sémantique d'une unité terminologique. Les différentes mentions d'une entité nommée sont groupées dans la propriété *Marqueur* de l'unité terminologique correspondante. Les occurrences d'une entité nommée dans le texte sont associées à la propriété *Occurrence* de l'unité terminologique correspondante. L'exemple suivant tiré du cas d'usage AAdvantage décrit les propriétés de l'unité terminologique *American airlines* dans le tableau 5.8.

Propriété	**Valeur**
Label	American airlines
Type sémantique	ORGANISATION
Marqueur	American airlines, American Airlines, A.Airlines
Occurrence	American Airlines is not responsible for products and services offered by other participating companies. American airlines may amend its rules of the Program at any time without notice. Any claim for uncredited mileage must be received by A.Airlines within 12 months after the mileage credit was earned.

TABLE 5.8 – Exemple des propriétés de l'unité terminologique *American airlines*.

Les outils de REN extraient des entités nommées et leurs associent des types sémantiques même s'il peut y avoir des erreurs de types sémantiques. Parfois ces outils n'identifient pas certaines mentions d'entités nommées. C'est le cas généralement des entités très spécifiques à un domaine précis. Néanmoins, nous considérons que le résultat est « appréciable » et justifie l'exploitation de ce type d'unités linguistiques dans la création d'un réseau terminoconceptuel. Nous citons quelques exemples de normalisation des entités nommées dans la section 5.6.

5.2.3 Relations terminologiques

Dans cette section, nous décrivons le résultat de l'extraction des relations terminologiques par des outils de TAL ou d'extraction de relations sémantiques à partir d'un corpus d'acquisition. Nous montrons le rôle de l'extraction des relations terminologiques pour la construction d'un réseau termino-conceptuel. Nous présentons d'une manière générale les différentes méthodes de l'identification des relations terminologiques et plus spécifiquement ce que nous avons utilisé. Enfin, nous citons quelques exemples illustratifs de l'identification des relations terminologiques faite sur nos corpus.

Traditionnellement, les relations terminologiques sont utilisées pour la création de relations conceptuelles. Comme les termes représentent les futurs concepts d'une ontologie, les relations terminologiques qu'entretiennent ces termes peuvent se révéler être des mentions linguistiques des relations conceptuelles. Comme nous l'avons mentionné dans le chapitre 2, le repérage des relations terminologiques aide à la structuration des ressources sémantiques (par ex.thésaurus) et conceptuelles (par ex. ontologie). Nous nous inscrivons dans la même optique que celle des travaux cités dans le chapitre 2. Nous considérons que l'extraction des relations terminologiques contribue à la création des relations termino-conceptuelles.

La détection des relations terminologiques est une tâche difficile à cause de la diversité des types de relations terminologiques à identifier et la variabilité de performance des méthodes adoptées (Manzano-Macho *et al.*, 2008) en fonction des corpus d'acquisition utilisés. Plusieurs techniques sont utilisées pour l'identification des relations dans le texte. Cette tâche peut être assurée par les outils de TAL qui extraient des relations syntaxiques (*APourTête/APourModifieur*) entre les termes en s'appuyant sur la composition syntaxique de ces derniers ainsi que des informations syntaxiques (par ex. catégorie syntaxique des arguments). Prenons par exemple les deux unités terminologiques *AAdvantage mileage credit* et *AAdvantage mileage* qui sont reliées par une relation terminologique de type *APourModifieur*. Les analyseurs syntaxiques s'appuient sur les dépendances syntaxiques (sujet, objet) dans les phrases pour détecter des relation syntaxiques. D'autres outils d'extraction de relations terminologiques s'appuient sur des patrons linguistiques associés à des types prédéfinis de relations terminologiques pour l'extraction des rela-

tions à partir du corpus.

Dans cette thèse, nous n'avons pas utilisé d'outils pour l'extraction des relations spécialisées. Nous avons repéré des relations spécialisées à travers l'application manuelle de patrons à base de verbes décrits sous la forme « SN Verb SN » sur nos corpus. Par exemple, nous avons utilisé le patron « SN is equivalent to SN » pour extraire la relation de type *Synonymie* (Hamon, 2000) qui relie les deux unités terminologiques *AAdvantage Gold member* et *oneworld Ruby* dans la phrase « AAdvantage Gold is equivalent to oneworld Ruby ». Nous avons aussi exploité le résultat de l'outil *YaTeA* pour la détection des relations syntaxiques.

Nous avons extrait des relations terminologiques en appliquant manuellement des patrons à base de verbes sur le texte. Par application du patron « SN Verb SN » sur le corpus de AAdvantage, nous avons extrait 168 verbes dont 74 sont mentionnés dans des passages réglementaires comme par exemple les verbes *adhere*, *earn* et *reserve* qui sont identifiés dans les phrases suivantes :

1. *American Airlines reserves the right to change the eligible fare classes at any time without notice.*
2. *The member must adhere to the rules of the shareAAmiles program.*
3. *Mileage credit cannot be earned for the same flight.*
4. *American Airlines reserves the right to end the AAdvantage program with six months notice.*

Nous nous appuyons sur les résultats des outils de TAL et de l'extraction des relations spécialisées à base de patrons afin de créer dans le réseau terminologique des relations terminologiques. Nous obtenons des triplets $rt(ut_i, ut_k, ut_j)$ tels qu'une relation rt appartient à l'ensemble de relations terminologiques (RT, $rt \in RT$) avec les unités ut_i, ut_j jouant le rôle de *Source* et *Destination* de la relations extraite et ut_k représentant le type de la relation. Dans le graphe sous-jacent au réseau terminologique, les types de relations terminologiques sont considérés comme des étiquettes portées par les arcs qui relient des nœuds correspondant aux unités terminologiques. Prenons l'exemple de la relation terminologique de type *earns* qui relit les deux unités terminologiques *AAdvantage member* et *500-mile electronic upgrades* : $rt(AAdvantagemember, earns, 500-mileelectronicupgrades)$. Cette relation décrit qu'un membre gagne « 500 miles »(voir figure 5.3). Nous avons décrit

dans le chapitre 4 les différentes catégories de relations terminologiques. La relation de type *earns* est une relation spécialisée dénotant une relation de domaine entre deux unités terminologiques. Le tableau 5.9 décrit les propriétés de la relation terminologique de type *earns*.

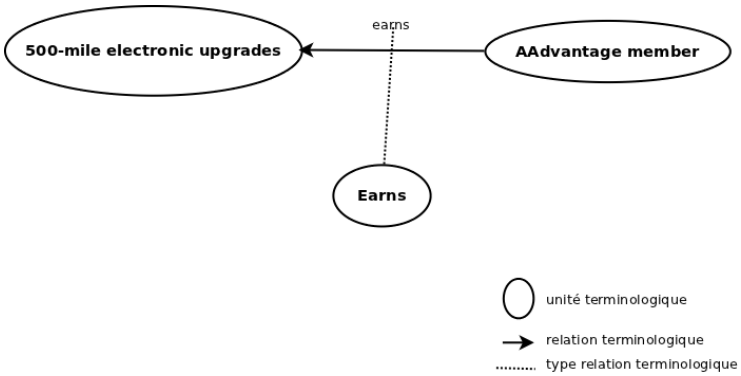

FIGURE 5.3 – Un exemple d'une relation terminologique tiré du cas d'usage de AAdvantage.

Propriété	Valeur
Type	earns
Catégorie	Relation spécialisée
Source	AAdvantage member
Destination	500-mile electronic upgrades

TABLE 5.9 – Exemple d'une relation terminologique de type *earns*.

Dans la suite de ce chapitre, les figures décrivant un réseau terminologique n'explicitent pas l'identité entre un type de relation terminologique (porté par un arc dans le graphe) et l'unité terminologique correspondant à ce type. Le tableau 5.10 décrit les propriétés de l'unité terminologique *earns*.

L'application des outils d'extraction de relations terminologiques à base de patrons nécessite l'adaptation de ces patrons au domaine étudié et au corpus d'acquisition puisque ces outils ne sont pas génériques (Manzano-Macho *et al.*, 2008). Ces outils extraient des relations spécialisées qui sont peu nombreuses. Les analyseurs syntaxiques et les extracteurs de termes génèrent beaucoup de relations terminologiques. Le résultat obtenu doit être nettoyé. Le réseau

Propriété	Valeur
Label	earns
Type sémantique	
Marqueur	earning, earn, earns
Occurrence	AAdvantage member earns four 500-mile electronic upgrades. The more you fly, the more upgrades you earn. Earning elite-qualifying miles when you purchase eligible tickets.

TABLE 5.10 – Exemple d'une unité terminologique décrivant un type de relation terminologique.

terminologique $G_T(UT, RT)$ obtenu est potentiellement volumineux et bruité. Il est formé par des unités terminologiques (UT) décrivant des termes et des entités nommées et des relations terminologiques (RT) représentées par des triplets d'unités terminologiques.

Dans cette thèse, nous avons extrait manuellement des relations terminologiques en appliquant des patrons à base de verbes. Mais la structure de connaissances manipulée définie dans le chapitre 4 supporte le résultat des outils d'extraction de relations terminologiques dans le cas où ces outils sont disponibles. Comme nous n'avons pas utilisé d'outil d'extraction de relations, les unités terminologiques sont reliées essentiellement à travers des relations syntaxiques de type *APourTête/APourModifieur*. Dans la section suivante, nous exposons les enjeux face auxquels notre méthodologie doit apporter des solutions.

5.3 Difficultés de la normalisation

Comme nous l'avons exposé dans la section précédente, le réseau terminologique obtenu suite à l'extraction du matériau linguistique à partir d'un corpus d'acquisition est un réseau potentiellement volumineux, hétérogène et bruité. Les unités terminologiques décrivent à la fois des notions et des rela-

tions du domaine. Certaines unités et relations terminologiques ne sont pas pertinentes à normaliser. D'autres unités sont ambiguës et leurs sens ne sont pas tous pertinents pour le domaine. A l'inverse, certaines unités ont des sens voisins et doivent être regroupées. La manipulation du réseau terminologique n'est pas donc une tâche évidente. Dans cette section, nous exposons les principales difficultés de ce travail.

5.3.1 Se repérer dans le réseau terminologique

Face aux unités et aux relations terminologiques formant le réseau terminologique et aux informations les caractérisant, l'ingénieur de la connaissance a besoin de savoir dans quel ordre normaliser ces unités et ces relations. L'exploration d'un vaste graphe de données est *a priori* une tâche difficile parce qu'il n'y a pas de parcours prédéfini à suivre. Il faut cependant s'assurer que l'ingénieur de la connaissance parcourt tout le réseau. Intuitivement, l'ingénieur de la connaissance parcourt le réseau terminologique de trois manières :

- en réduisant le réseau à une liste d'unités : dans ce cas, l'ingénieur de la connaissance s'intéresse qu'aux unités terminologiques et ignore les relations qui les lient ainsi que la sémantique qu'elles portent ;
- de proche en proche : l'ingénieur de la connaissance sélectionne à chaque fois le voisin d'une unité terminologique mais il est difficile de savoir quels voisins sélectionner ;
- d'autres parcours sont envisageables si on a une connaissance *a priori* de la structure du réseau. Par exemple, l'ingénieur de la connaissance parcourt le réseau en sélectionnant des unités terminologiques en fonction de leurs propriétés comme le nombre de relations qu'elles partagent avec d'autres unités.

En conclusion, l'ingénieur de la connaissance parcourt un réseau soit par le repérant des unités suivant un indice de pertinence (par ex. fréquence) soit en sélectionnant des unités en fonction de leur voisinage dans le réseau. Ces deux stratégies, poussées à l'extrême, conduisent à deux résultats différents si l'on considère que le graphe ne peut être exploité dans son intégralité (voir figure 5.4) :

1. la normalisation d'unités disparates dans le réseau sans considération de leurs voisinages. Dans le réseau terminologique (partie gauche de la

figure 5.4), les nœuds coloriés correspondent aux unités terminologiques normalisées tels que leurs voisins directs ne sont pas des unités normalisées ;

2. la normalisation d'un sous réseau du réseau terminologique sans considération d'autres parties du graphe. Cet exemple est décrit dans la partie droite de la figure 5.4. Les unités terminologiques sont normalisées elles et leurs voisins directs.

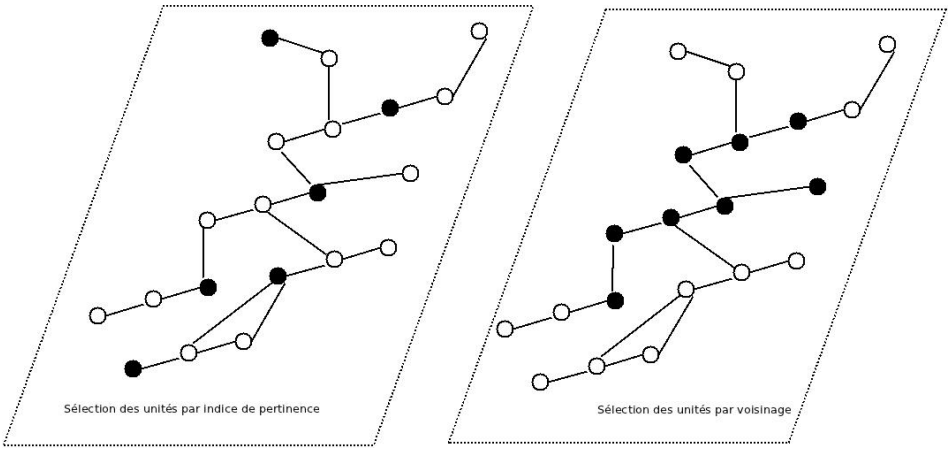

FIGURE 5.4 – Les deux stratégies de parcours dans le réseau terminologique.

Adopter l'une ou l'autre stratégie ne favorise pas un « équilibre » dans le traitement des unités et relations de domaine. L'ingénieur de la connaissance a besoin d'avoir une « appréciation » générale sur la composition du réseau en termes d'unités terminologiques extraites, de catégories de relations terminologiques et de leurs types. Il faut donc le guider dans le parcours dans le réseau terminologique. L'ingénieur de la connaissance a besoin de moyens permettant d'assurer une certaine « homogénéité » dans le parcours du réseau terminologique.

Notre méthode apporte une aide à l'ingénieur de la connaissance pour sélectionner des unités et des relations terminologiques. Nous proposons aussi des indicateurs afin de l'avertir dans le cas où le parcours qu'il adopte n'assure pas une couverture suffisante du graphe terminologique. Nous proposons trois possibilités de parcours dans le réseau terminologique :

– la sélection des unités terminologiques suivant un indice de pertinence (par ex. voir section 5.5.4.1) ;

– la sélection des unités par voisinage dans le réseau ;
– la sélection des unités suivant un choix propre à l'ingénieur de la connaissance : par exemple, il choisit de normaliser des unités terminologiques ayant un type sémantique pertinent par rapport à l'application visée (par ex. toutes les unités terminologiques ayant comme type sémantique *Organisation*).

5.3.2 Détecter des unités pertinentes

Le réseau terminologique extrait par des outils de TAL – de reconnaissance de termes, d'entités nommées et de relation terminologiques – , comporte des unités pertinentes par rapport au domaine et à l'application visée mais aussi du bruit. Les connaissances décrites au niveau terminologique peuvent être ambiguës et redondantes. Il s'ensuit que les relations terminologiques qui existent au sein du réseau terminologique doivent être validées par l'ingénieur de la connaissance. Le réseau terminologique est potentiellement volumineux et bruité. Il faut définir des mécanismes pour le tri des données. Notre méthode définit des critères de pertinence pour la sélection et la normalisation des unités et des relations terminologiques.

5.3.3 Arrêter le processus de normalisation

Durant le travail de normalisation, l'ingénieur de la connaissance a besoin d'indicateurs qui donnent une idée de la progression de la normalisation du réseau terminologique. L'analyse exhaustive du réseau terminologique n'est pas réaliste. En pratique, l'ingénieur de la connaissance filtre et sélectionne les données d'une manière un peu grossière pour aboutir à une première version d'un réseau normalisé qu'il enrichit durant d'autres parcours dans le réseau terminologique. Il doit être informé des nouvelles unités terminologiques qui sont au voisinage des unités déjà normalisées afin de vérifier si ces dernières sont importantes à prendre en considération pour la création du réseau termino-conceptuel.

Comme le travail de normalisation n'est pas totalement automatisable, l'arrêt de ce dernier reste une décision à prendre par l'ingénieur de la connaissance. Mais, nous proposons de lui fournir des informations sur la progression de la normalisation du réseau terminologique pour le guider dans cette prise de

décision. Par exemple, des informations qui permettent de donner une appréciation sur le taux des unités pertinentes non encore normalisées ou sur celles qui sont normalisées. Notre méthode définit des indicateurs de contrôle qui permettent de donner une idée de l'aboutissement du travail de normalisation.

5.4 Point d'arrivée : le réseau termino-conceptuel

Précisons tout d'abord que, durant la phase de normalisation, nous ne nous intéressons pas à l'enrichissement du réseau terminologique ni à la « terminologisation » du réseau termino-conceptuel. L'enrichissement du réseau terminologique consiste en l'ajout d'autres unités ou relations terminologiques dans le réseau terminologique qui n'apparaissent pas dans le corpus d'acquisition qui a servi à créer ce réseau. La terminologisation du réseau termino-conceptuel est le fait de relier par des liens de correspondance des termino-concepts, des types de relations termino-conceptuelles et des relations termino-conceptuelles à des unités ou relations terminologiques qui ne sont pas mentionnées. C'est utile quand on est parti d'une ontologie préexistante qui n'est pas ancrée dans le corpus d'acquisition.

A la fin de la phase de normalisation, nous obtenons un réseau formé par les unités et les relations terminologiques qui sont normalisées en unités termino-conceptuelles (termino-concepts et types de relations termino-conceptuelles) et des relations termino-conceptuelles ainsi que l'ensemble des liens de correspondance créés entre les structures de connaissances terminologiques et termino-conceptuelles utilisées. Le réseau conjoint formé, noté $G_T \& G_{TC}$, décrit un vocabulaire normalisé et structuré associé aux mentions linguistiques utilisées dans le corpus d'acquisition.

Au début de la phase de normalisation, le réseau terminologique $G_T(UT, RT)$ issu de l'extraction terminologique est potentiellement volumineux et bruité. Il est formé par des unités terminologiques (UT) décrivant des termes et des entités nommées et des relations terminologiques (RT) représentées par des triplets d'unités terminologiques $rt(ut_i, ut_k, ut_j)$ tels que les unités ut_i, ut_j jouent le rôle de *Source* et *Destination* d'une relation et le type d'une relation est représenté par l'unité terminologique ut_k. Le réseau termino-

conceptuel $G_{TC}(TC, TypeRTC, RTC)$ contient initialement les catégories des types de relations termino-conceptuelles décrites dans une liste notée $L_{TypeRTC}$ s'agissant des types de *Généricité/Spécificité* et *relation associative*. La figure 5.5 décrit les réseaux terminologique et termino-conceptuel [6]. Chaque nœud du graphe $G_T\&G_{TC}$ est représenté par un cercle dont la couleur caractérise son état. Les unités terminologiques validées par l'ingénieur de la connaissance sont représentées par des nœuds bleus [7]. Les termino-concepts créés sont représentés par des nœuds orangés [8]. Chaque relation est représentée par un arc étiqueté et orienté reliant deux nœuds et la couleur de l'étiquette portée par l'arc spécifie qu'il s'agit de la création d'un nouveau type de relation. Chaque nouveau type de relation termino-conceptuelle est représenté par une étiquette orangée [9] qui est ajoutée à la liste des types de relations termino-conceptuelles $L_{TypeRTC}$.

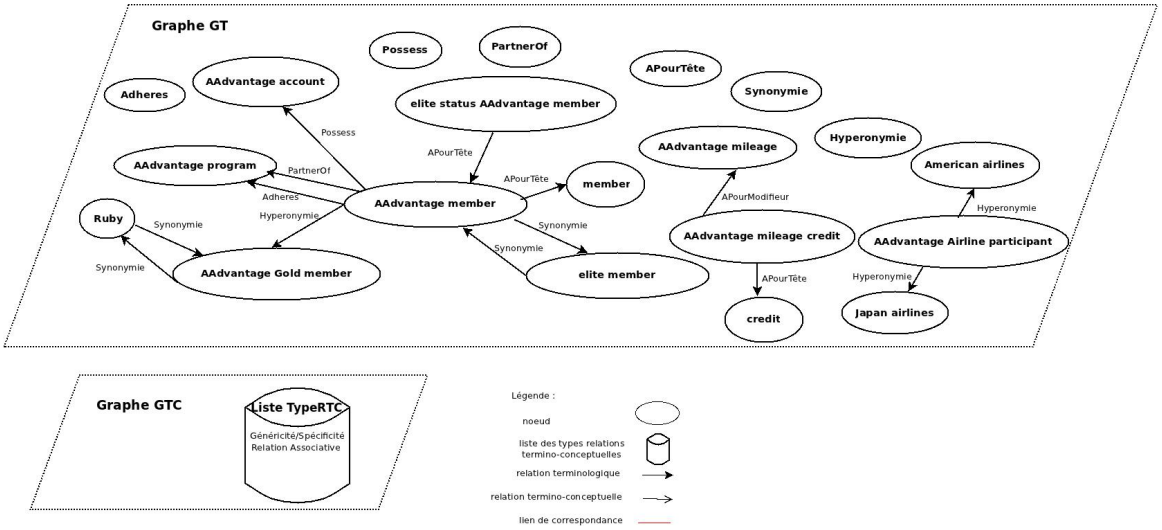

FIGURE 5.5 – Extrait du graphe $G_T\&G_{TC}$ au début de la phase de normalisation relatif au cas d'usage de *American Airlines*.

Afin de simplifier la visualisation des nouveaux nœuds et des nouveaux arcs créés dans le graphe $G_T\&G_{TC}$, nous ne représentons pas systématiquement

6. Les réseaux représentent des unités et des relations terminologiques qui sont extraites à partir du corpus du cas d'usage de *American Airlines*.
7. En mode impression sans couleur, le nœud apparaît en gris foncé.
8. En mode impression sans couleur, le nœud apparaît en gris clair.
9. En mode impression sans couleur, le nœud apparaît en gris clair.

les nœuds qui décrivent les unités terminologiques associées aux types des relations terminologiques, dans la suite de ce chapitre. Les types des relations terminologiques sont notés sur les arcs reliant des nœuds au niveau du graphe terminologique G_T.

A la fin de la phase de normalisation, le réseau terminologique est composé d'un ensemble d'unités terminologiques (UT) dont un sous ensemble UT' ($UT' \subset UT$) contient des unités normalisées et d'un ensemble de relations terminologiques dont un sous ensemble RT' ($RT' \subset RT$) contient des relations normalisées. Les unités normalisées sont des unités terminologiques qui sont reliées à des termino-concepts ou à des types de relations termino-conceptuelles représentés dans le réseau termino-conceptuel à travers des liens de correspondance. De la même manière, les relations normalisées sont des relations terminologiques qui sont reliées à des relations termino-conceptuelles à travers des liens de correspondance. Le réseau termino-conceptuel $G_{TC}(TC, TypeRTC, RTC)$ est composé d'un ensemble de termino-concepts TC interconnectés par des relations termino-conceptuelles RTC. Les deux réseaux terminologique et termino-conceptuel forment un réseau appelé $G_T\&G_{TC}$ qui représente les deux réseaux ainsi que les liens de correspondance définis entre les éléments des réseaux mis en correspondance. En effet, à la fin de la phase de normalisation, le réseau $G_T\&G_{TC}$ est enrichi avec des liens de correspondance créés entre des unités ou des relations terminologiques et des termino-concepts, des types de relations termino-conceptuelles ou des relations termino-conceptuelles. La figure 5.6 décrit un extrait du réseau ($G_T\&G_{TC}$), relatif au cas d'usage de *American Airlines*, obtenu à la fin du travail de normalisation.

Dans la section suivante, nous définissons chacune des opérations élémentaires qui permettent de sélectionner, valider, créer et mettre à jour le réseau $G_T\&G_{TC}$. Nous définissons un processus assurant cette normalisation sémantique qui prend en entrée les réseaux terminologique et termino-conceptuel et les liens de correspondance tels qu'ils sont décrits dans la section 5.4 et donne en sortie ces réseaux enrichis suite à des modifications faites au niveau de l'un ou des deux réseaux terminologique et termino-conceptuel.

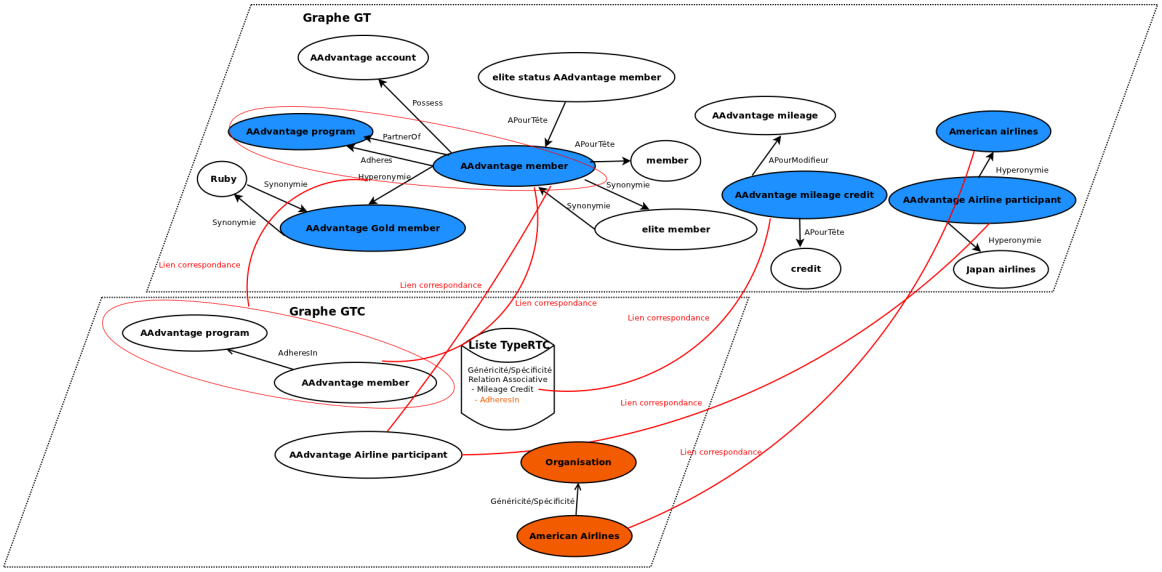

FIGURE 5.6 – Extrait du graphe G_T&G_{TC} relatif au cas d'usage de *American Airlines* à la fin de la phase de normalisation.

5.5 Normalisation et contrôle des transformations élémentaires

La méthode que nous proposons vise à créer un réseau termino-conceptuel $G_{TC}(TC, TypeRTC, RTC)$ à partir d'un réseau terminologique $G_T(UT, RT)$. Après la phase d'extraction terminologique (cf section 5.2), vient la phase de normalisation. La phase de normalisation se fonde sur des opérations élémentaires dites de transformation qui permettent de normaliser des unités et des relations terminologiques et de créer et mettre à jour des termino-concepts, des types de relations termino-conceptuelles et des relations termino-conceptuelles. Ces opérations influent sur la structure du réseau G_T&G_{TC}. Ces opérations de normalisation et de mise à jour sont définies dans des opérations composées (macro-opérations) qui permettent de modifier et de contrôler l'évolution du travail de la normalisation des réseaux terminologique et termino-conceptuel. Dans cette section, nous décrivons les opérations élémentaires et composées de normalisation.

5.5.1 Opérations élémentaires

Durant la phase de normalisation, l'ingénieur de la connaissance valide, invalide des unités terminologiques, crée et met à jour des termino-concepts et des relations termino-conceptuelles. Nous définissons des opérations élémentaires qui permettent de sélectionner, valider, créer et mettre à jour des unités et des relations au niveau du graphe $G_T \& G_{TC}$.

Nous avons choisi de modéliser les opérations de transformation sur les graphes par des diagrammes d'activités définis dans la modélisation UML. La modélisation des opérations avec des diagrammes d'activités explicite le comportement interne des opérations. En effet, les opérations sont décrites par un ensemble d'actions élémentaires qui s'enchaînent et qui sont contrôlées par des conditions d'exécution et d'arrêt. Les diagrammes d'activités décrivent aussi les objets manipulés par ces actions. Ce type de diagramme UML décrit deux types de flots : le flot de contrôle et le flot des objets. Le flot de contrôle représente le déroulement des actions et le flot d'objets décrit les objets véhiculés entre les actions. Le formalisme des diagrammes d'activités repose sur un ensemble de concepts décrits dans la figure 5.7 :

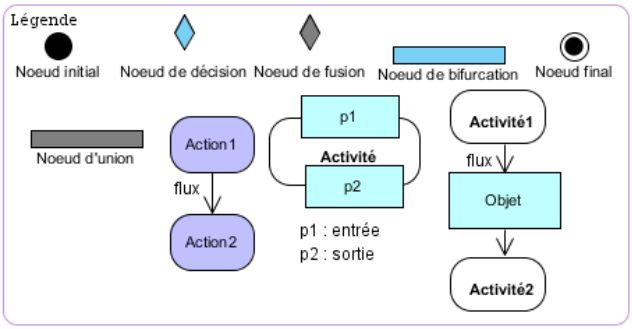

FIGURE 5.7 – Légende du diagramme d'activités.

- les nœuds d'exécution (initial et final) permettent de déclencher ou d'arrêter un flux ;
- les nœuds de bifurcation permettent de créer plusieurs flux sortants à partir d'un seul flux entrant. Ils sont représentés par une barre de synchronisation ;
- les nœuds de fusion permettent de produire un seul flux sortant à partir de plusieurs flux entrant de la barre de synchronisation. Le flux sortant peut être exécuté lorsqu'au moins un flux entrant est activé. Ce nœud

est le symétrique du nœud de bifurcation ;
- les nœuds de décision permettent à partir d'un seul flux entrant de produire plusieurs flux sortants qui correspondent aux différentes valeurs que peut prendre une condition ;
- les pins d'entrée ou de sortie (annotés $p1$ et $p2$ dans la figure 5.7) sont des paramètres spécifiés respectivement comme entrée et sortie d'une action ou d'une activité. Un pin peut être un objet passé comme paramètre à une action ou une activité. Une action ou une activité peut recevoir un ou plusieurs pins d'entrée ;
- les flots de contrôle [10] décrivent l'enchaînement des actions ;
- les flots d'objets permettent de représenter un flux d'objets entre des actions ou des activités [11] ;
- les concepts objet représentent des objets. Un objet peut être décrit par un pin d'entrée ou de sortie d'une action ou une activité ;
- les actions correspondent à des traitements (élémentaires au sens d'instructions en programmation ou composés de plusieurs actions) qui affectent l'état d'un système ;
- les activités représentent un ensemble d'actions et leur enchaînement (flot de contrôle). Une activité peut recevoir des paramètres (pouvant être des objets, des variables, etc) d'entrée et en produire en sortie.

Dans la suite de ce chapitre, nous décrivons les opérations de normalisation à travers des diagrammes d'activités qui décrivent le déroulement de ces opérations. Les opérations élémentaires de normalisation manipulent comme objets des unités et des relations terminologiques. Nous associons une nouvelle propriété notée *Statut* à une unité terminologique afin de contrôler le travail de normalisation. Cette propriété représente l'état de cette dernière durant le travail de normalisation. Au début, une unité a le statut « candidat » : c'est une unité qui n'est pas encore sélectionnée par l'ingénieur de la connaissance. Une fois qu'elle est sélectionnée, l'ingénieur de la connaissance peut la valider (son statut devient « validé ») ou la supprimer du réseau terminologique (son statut devient « invalidé »). Parmi les unités terminologiques ayant comme statut *validé*, il y a des unités qui ne sont pas normalisées en termino-concepts ou types de relations termino-conceptuelles. Seule la créa-

10. Ce flot est représenté par une flèche reliant deux actions.
11. Ce flot est représenté par une flèche reliant une action à un objet.

tion des liens de correspondance reliant des unités terminologiques à des unités termino-conceptuelles permet de détecter si des unités sont normalisées ou pas. Sauvegarder les statuts des unités terminologiques permet de vérifier, durant l'étape de contrôle de normalisation, s'il existe encore des unités non normalisées et détecter celles qui doivent l'être.

Toutes les opérations élémentaires de transformation influent sur la structure du graphe $G_T\&G_{TC}$. Dans les diagrammes d'activités correspondant à ces opérations, le graphe $G_T\&G_{TC}$ est représenté comme un objet global pour toutes ces opérations. Dans la suite, chacune des opérations présentées agit sur le graphe $G_T\&G_{TC}$. L'évolution de la structure du graphe est décrite au fur et à mesure dans la définition des opérations élémentaires de transformation.

5.5.1.1 Sélection d'une unité terminologique

La normalisation d'une unité terminologique suppose, d'abord, de sélectionner dans le graphe G_T une unité candidate. Nous définissons l'opération « *Sélection unité terminologique* » qui permet la sélection d'une unité terminologique *ut* dans le graphe G_T. L'ingénieur de la connaissance a le choix de trois méthodes de sélection d'une unité terminologique :

1. sélection suivant un indice de pertinence : l'unité terminologique est sélectionnée suivant son rang dans une liste ordonnée d'unités terminologiques, notée *LW* (appel à l'opération *Sélection dans liste LW*) ;
2. sélection par voisinage : une unité terminologique est sélectionnée parce qu'elle est voisine d'une autre unité précédemment sélectionnée (appel à l'opération *Sélection par voisinage*) ;
3. sélection aléatoire dans le graphe G_T : une unité terminologique est sélectionnée de manière libre par l'ingénieur de la connaissance (appel à l'opération *Sélection dans un graphe*).

Ces trois méthodes de sélection correspondent aux différents types de parcours (*cf* section 5.3) que l'ingénieur de la connaissance peut entreprendre dans le réseau terminologique. A la fin de cette opération, une unité terminologique *ut* est sélectionnée. Elle sert de point de départ pour les opérations de création de termino-concepts et de types de relations termino-conceptuelles. L'opération *Sélection unité terminologique* est modélisée par un diagramme d'activités décrit dans la figure 5.8.

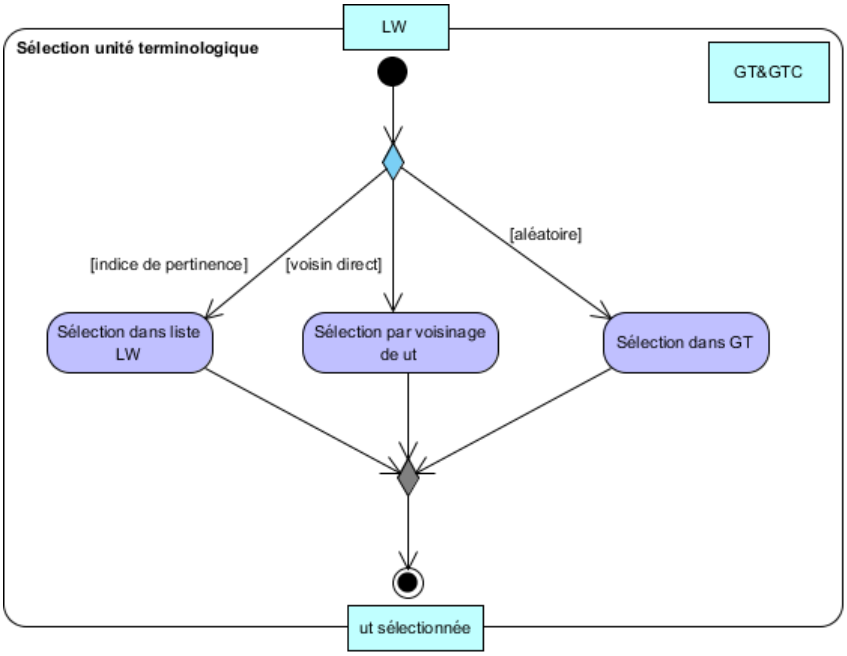

FIGURE 5.8 – Le diagramme d'activités de l'opération *Sélection unité terminologique*.

5.5.1.2 Validation d'une unité terminologique

Cette opération permet de mettre à jour les propriétés d'une unité terminologique ut ($ut \in UT$) en principe dans le but de la normaliser. La validation d'une unité terminologique permet de spécifier si elle est pertinente ou pas, si une elle est ambiguë ou synonyme d'une autre unité (à travers l'étude de ses occurrences).

Plus concrètement, l'ingénieur de la connaissance valide ou modifie les propriétés associées à une unité terminologique sélectionnée afin de ne garder que les informations qui sont utiles pour décrire l'unité courante. La validation d'une propriété donnée consiste à garder la même valeur de cette propriété initiale. La modification d'une propriété donnée est une mise à jour de sa valeur ou sa suppression. L'ingénieur de la connaissance ajoute ou supprime un ou plusieurs types sémantiques, un ou plusieurs marqueurs et une ou plusieurs occurrences de l'unité courante. Il n'a pas le droit de supprimer le label de l'unité terminologique mais il peut le modifier. Le tableau 5.11 décrit les propriétés de l'unité terminologique *AAdvantage Gold member*.

L'opération *Validation unité terminologique* est représentée par le dia-

Propriété	Valeur
Label	AAdvantage Gold member
Type sémantique	
Marqueur	AAdvantage Gold member, AAdvantage Gold members, AAdvantage Gold
Occ	AAdvantage Gold is equivalent to oneworld Ruby. AAdvantage Gold member earns four 500-mile electronic upgrades for every 10,000 qualifying base miles flown. AAdvantage Gold members may upgrade from any individual published coach ticket to the next class of service.
Statut	candidat

TABLE 5.11 – Les propriétés de l'unité terminologique *AAdvantage Gold member*.

gramme d'activités de la figure 5.10.

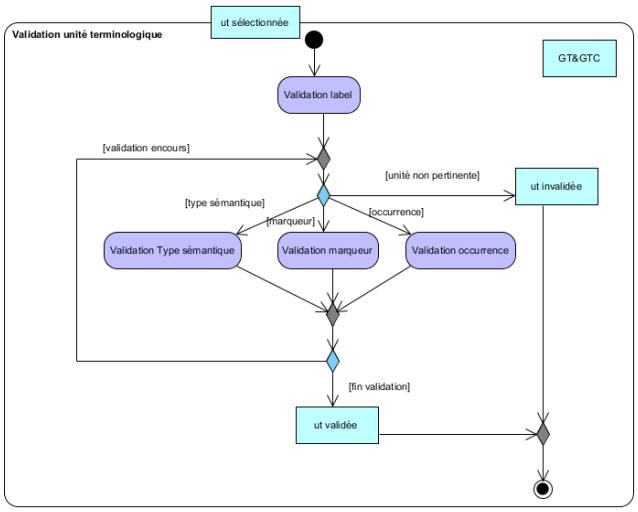

FIGURE 5.9 – Le diagramme d'activités de l'opération *Validation unité terminologique*.

La validation ou la modification de chacune des propriétés de l'unité courante est optionnelle. L'ingénieur de la connaissance choisit les propriétés qu'il

décide de modifier. La validation des propriétés de l'unité terminologique courante n'implique pas la validation de l'unité. Seule la modification du statut de « candidat » à « validé » fait que l'unité est validée. L'ingénieur de la connaissance peut considérer une unité comme invalide (son statut est « invalidé ») s'il juge que cette unité n'est pas pertinente pour le domaine. Dans le graphe $G_T\&G_{TC}$, une unité validée figure en bleu. Si une unité est invalidée alors elle est supprimée du $G_T\&G_{TC}$ ainsi que l'ensemble des relations terminologiques qu'elle partage avec d'autres unités. Si l'ingénieur de la connaissance n'invalide pas une unité sélectionnée, le graphe $G_T\&G_{TC}$ reste inchangé. La couleur bleu indique que l'unité terminologique *AAdvantage Gold member* est validée. La figure 5.10 décrit ce changement au niveau du graphe $G_T\&G_{TC}$.

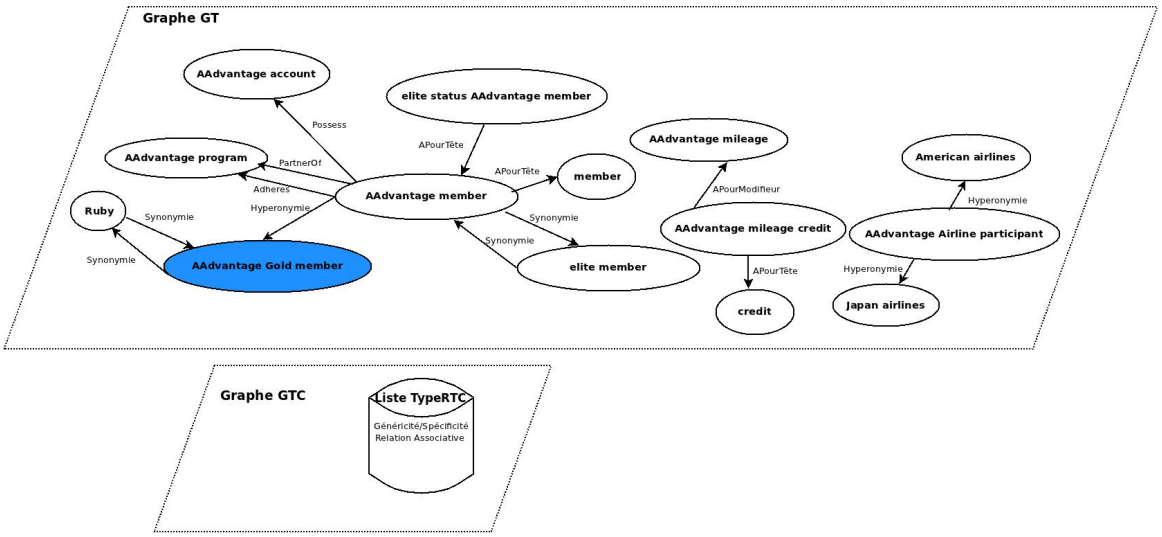

FIGURE 5.10 – Validation de l'unité terminologique *AAdvantage Gold member*.

A la fin de cette opération, les propriétés de l'unité terminologique courante sont mises à jour. Si une unité a un statut « invalidé » alors son score de pertinence est mis à zéro dans la liste LW.

5.5.1.3 Création d'un termino-concept

L'opération *Création termino-concept* permet la création d'un nouveau termino-concept *tc* au niveau du réseau termino-conceptuel $G_{TC}(TC, TypeRTC, RTC)$. L'ingénieur de la connaissance peut créer un nouveau termino-concept de deux manières :

– à partir de zéro ;
– à partir d'une unité terminologique.

Dans les deux cas, le nouveau termino-concept doit posséder un label préféré. Si le nouveau termino-concept est créé à partir de zéro alors l'ingénieur de la connaissance ajoute manuellement ce label. Dans le cas où il s'agit d'une création à partir d'une unité terminologique, le label préféré par défaut correspond au label de l'unité terminologique qui est ajouté automatiquement. L'affectation du label de l'unité au nouveau termino-concept valide automatiquement les propriétés de l'unité terminologique sélectionnée si elle n'est pas déjà validée. De plus, un lien de correspondance est créé entre l'unité terminologique validée et le nouveau termino-concept. L'ingénieur de la connaissance ajoute optionnellement un ou plusieurs labels alternatifs, une définition et une ou plusieurs relations termino-conceptuelles de type *Généricité/Spécificité* qui relient le nouveau termino-concept à d'autres termino-concepts existants dans le réseau G_{TC}. L'enchaînement de l'opération *Création termino-concept* est représentée par le diagramme d'activités de la figure 5.11. A la fin de cette opération, un nouveau termino-concept *tc* est créé et ajouté à l'ensemble des termino-concepts (TC) du réseau G_{TC}. L'unité terminologique qui a servi pour la création du nouveau termino-concept est validée automatiquement au moment de la création du lien de correspondance. La figure 5.12 décrit le résultat de la création du termino-concept *AAdvantage member* à partir de zéro. Le nouveau termino-concept créé n'est pas relié à un termino-concept existant par une relation de type *Généricité/Spécificité*, dans le réseau termino-conceptuel.

Le tableau 5.12 représente les propriétés du nouveau termino-concept *AAdvantage member* créé à la fin de cette opération.

Propriété	**Valeur**
Label préféré	AAdvantage member
Labels alternatifs	AA member, member
Définition	est une catégorie de passagers qui adhérent dans le programme de fidélité de AA.

TABLE 5.12 – Les propriétés du nouveau termino-concept *AAdvantage member*.

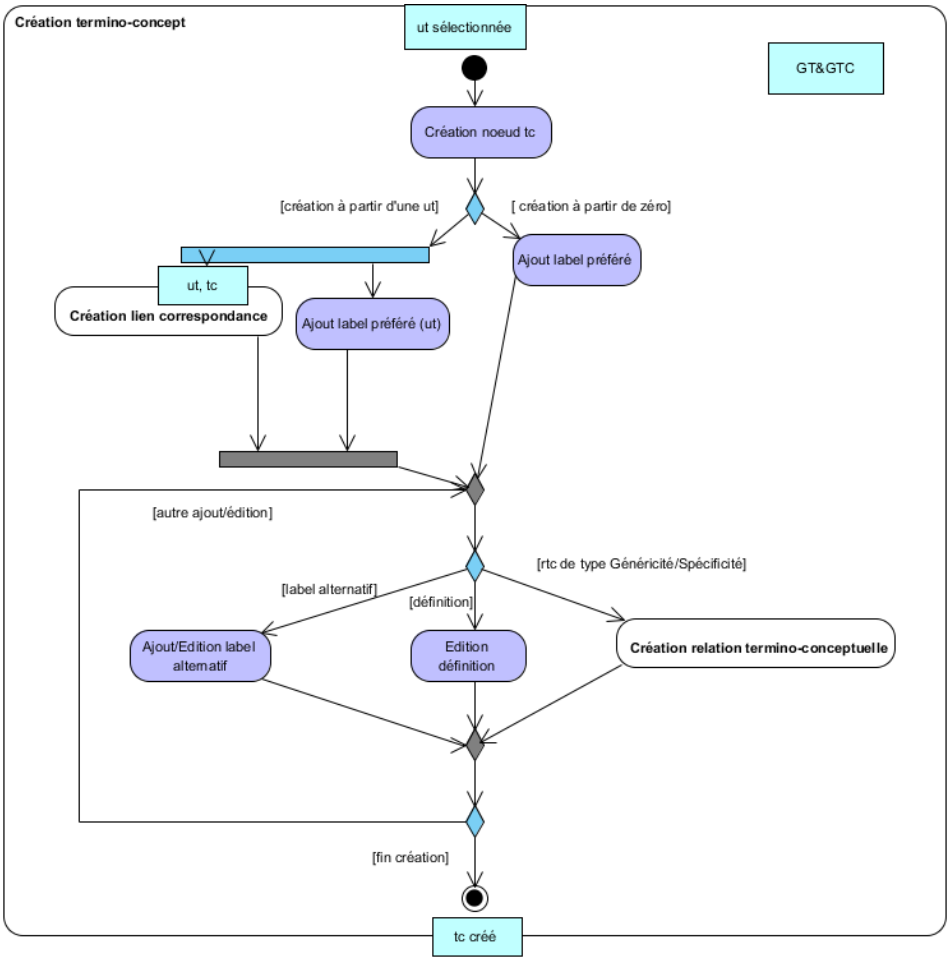

FIGURE 5.11 – Le diagramme d'activités de l'opération *Création termino-concept*.

5.5.1.4 Création d'un type de relation termino-conceptuelle

L'opération *Création type relation termino-conceptuelle* permet la création d'un nouveau *typeRTC* dans l'ensemble des types de relations termino-conceptuelles ($typeRTC \in TypeRTC$) du graphe G_{TC}. L'ingénieur de la connaissance peut créer un nouveau type de relation termino-conceptuelle de deux manières :

- à partir de zéro ;
- à partir d'une unité terminologique.

Dans les deux cas, le nouveau type de relation termino-conceptuelle doit posséder un label. Si le nouveau type de relation termino-conceptuelle est créé à partir de zéro alors l'ingénieur de la connaissance ajoute manuellement ce

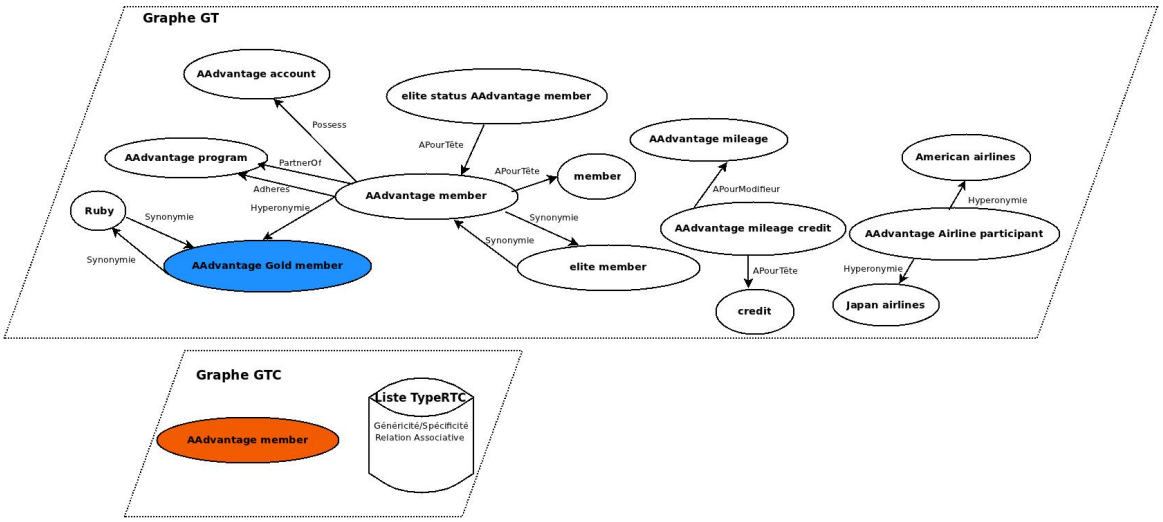

FIGURE 5.12 – Création d'un termino-concept au niveau du graphe G_T&G_{TC}.

label. Dans le cas où il s'agit d'une création à partir d'une unité terminologique, un label par défaut correspond au label de l'unité terminologique qui est ajouté automatiquement. Le fait d'affecter le label de l'unité au nouveau type de relation termino-conceptuelle valide automatiquement les propriétés de l'unité terminologique sélectionnée si elle n'est pas déjà validée. De plus, un lien de correspondance est créé entre l'unité terminologique validée et le nouveau type de relation termino-conceptuelle. L'ingénieur de la connaissance ajoute optionnellement une ou plusieurs propriété(s). L'enchaînement de l'opération *Création termino-concept* est représenté par le diagramme d'activités de la figure 5.13.

A la fin de cette opération, un nouveau type de relation termino-conceptuelle *typeRTC* est créé et ajouté à l'ensemble des types de relation termino-conceptuelle ($TypeRTC$). Prenons l'exemple d'un nouveau type de relation termino-conceptuelle *Mileage credit* qui est créé à partir d'une unité terminologique *AAdvantage mileage credit*. La figure 5.14 décrit le résultat de l'opération au niveau du graphe G_T&G_{TC}. Le nouveau type de relation termino-conceptuelle possède comme valeur de la propriété *Catégorie* « relation associative » [12]. L'unité terminologique *AAdvantage mileage credit* qui a servi à créer ce nouveau type de relation termino-conceptuelle est validée.

12. L'ajout de cette catégorie de relation est fait d'une manière automatique.

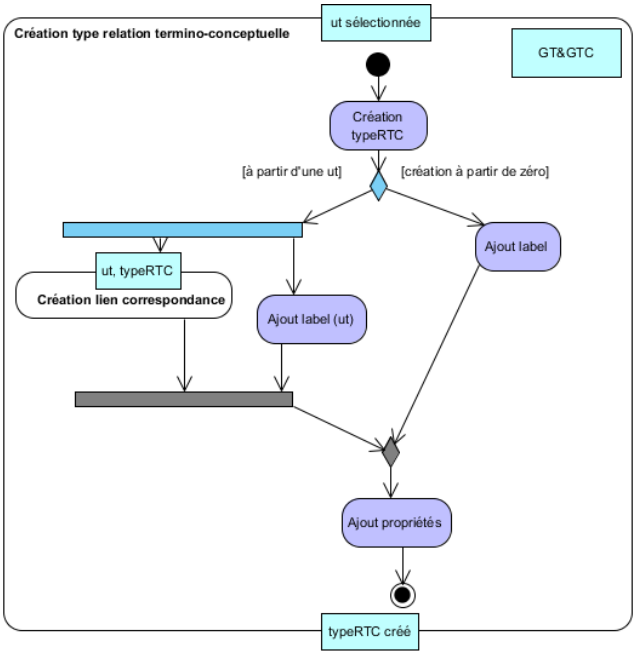

FIGURE 5.13 – Le diagramme d'activités de l'opération *Création type relation termino-conceptuelle*.

Le tableau 5.13 décrit les propriétés du nouveau type de relation termino-conceptuelle *Mileage credit*.

Propriété	**Définition**
Label	Mileage credit
Catégorie	relation associative

TABLE 5.13 – Les propriétés du nouveau type de relation termino-conceptuelle *Mileage credit*.

5.5.1.5 Mise à jour d'un termino-concept

La construction du réseau termino-conceptuel peut parfois entraîner des modifications au niveau du réseau termino-conceptuel. L'opération *Mise à jour termino-concept* permet de mettre à jour d'un termino-concept existant tc appartenant à l'ensemble des termino-concepts ($tc \in TC$). Plus spécifiquement, la mise à jour d'un termino-concept existant consiste en la modification

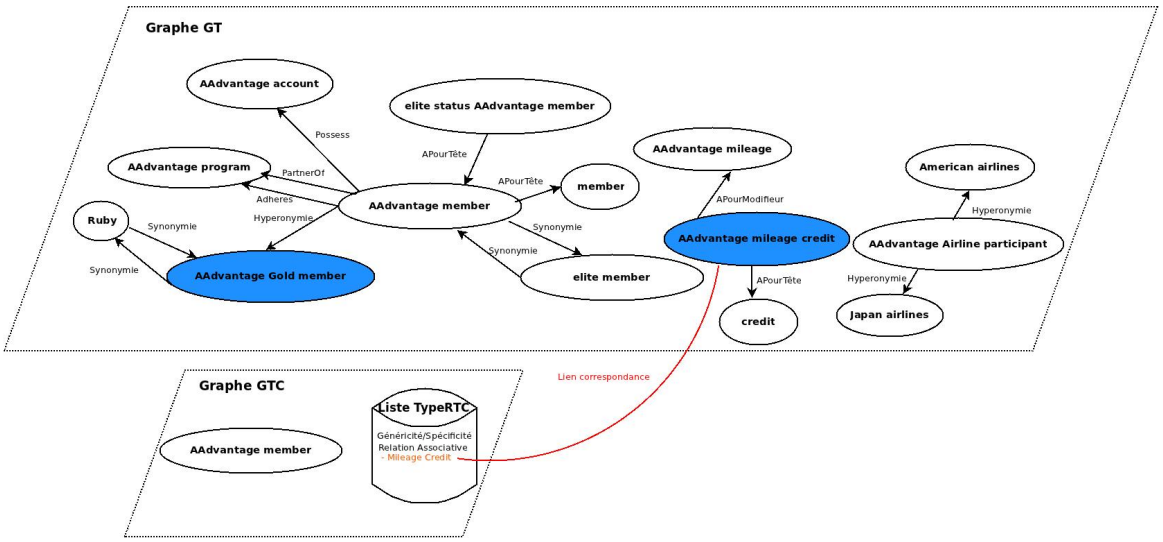

FIGURE 5.14 – Création d'un type de relations termino-conceptuelles.

de ses propriétés ou sa suppression s'il s'agit d'une suppression, un termino-concept est supprimé du réseau termino-conceptuel ainsi que toutes les relations termino-conceptuelles qu'il entretient avec d'autres termino-concepts.

Un des cas particuliers de modification des propriétés d'un termino-concept existant consiste à le relier à une unité terminologique par la création d'un lien de correspondance. Ce type de modification entre dans le cadre de la « terminologisation » d'unités termino-conceptuelles. Dans ce cas, l'ingénieur de la connaissance peut modifier le label préféré du termino-concept existant tel que ce nouveau label corresponde au label de l'unité terminologique. L'unité terminologique qui a servi pour la mise à jour du termino-concept est validée automatiquement au moment de la création du lien de correspondance.

D'autres modifications peuvent être faites par l'ingénieur de la connaissance comme par exemple la modification d'un ou plusieurs labels alternatifs, d'une définition ou encore d'une ou plusieurs relations termino-conceptuelles ayant comme type *Généricité/Spécificité* reliant le termino-concept courant à d'autres termino-concepts.

A la fin de cette opération, le termino-concept courant tc est mis à jour dans le graphe $G_T \& G_{TC}$. La figure 5.15 représente le diagramme d'activités de l'opération *Mise à jour termino-concept*.

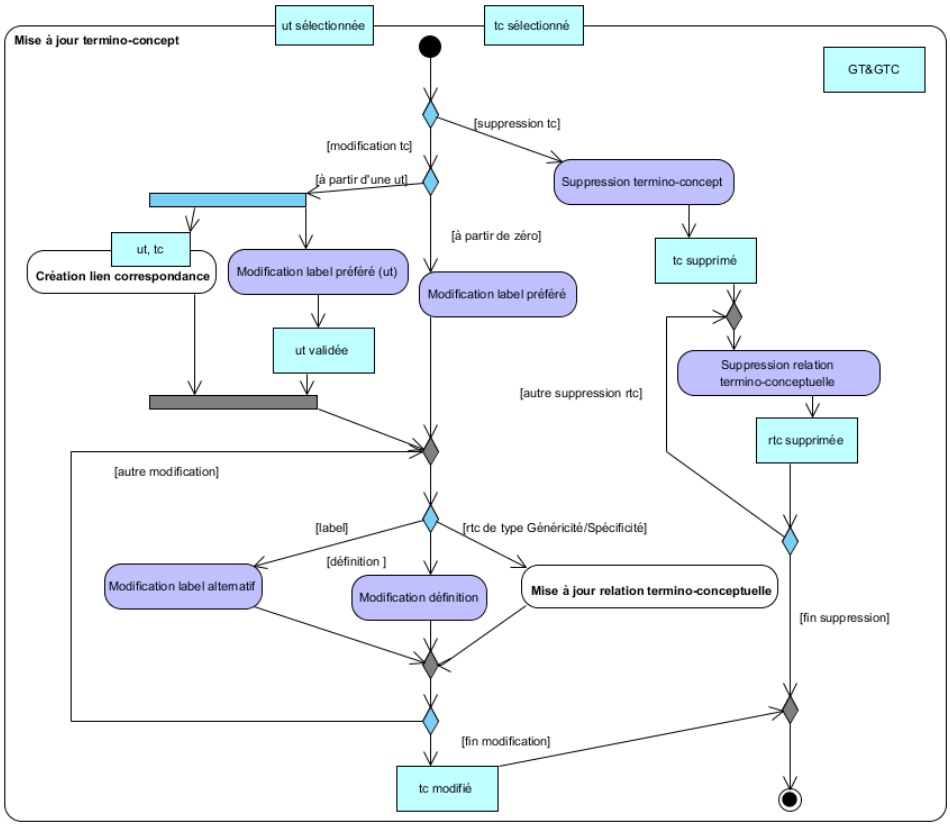

FIGURE 5.15 – Le diagramme d'activités de l'opération *Mise à jour termino-concept*.

5.5.1.6 Mise à jour d'un type de relation termino-conceptuelle

L'opération *Mise à jour type relation termino-conceptuelle* permet la mise à jour d'un type de relation termino-conceptuelle existant, $typeRTC$, appartenant à l'ensemble des types de relation termino-conceptuelles ($typeRTC \in TypeRTC$). Cette mise à jour consiste en la modification des propriétés du type ou éventuellement sa suppression. La suppression d'un type de relation termino-conceptuelle entraîne sa suppression de la liste des types de relations termino-conceptuelles $TypeRTC$. En plus, toutes les relations termino-conceptuelles ayant comme types de relations termino-conceptuelles $typeRTC$ doivent obligatoirement avoir un autre type de relation termino-conceptuelle que l'ingénieur de la connaissance doit sélectionner parmi la liste de $TypeRTC$ ou ajouter à cette liste.

Un des cas particuliers de modification des propriétés d'un type de relation termino-conceptuelle existant consiste à le relier à une unité terminologique

par la création d'un lien de correspondance. Dans ce cas, l'ingénieur de la connaissance peut modifier le label du type de relation termino-conceptuelle courant par le label de l'unité terminologique. L'unité terminologique qui a servi pour la mise à jour du type de relation termino-conceptuelle est validée automatiquement au moment de la création du lien de correspondance. D'autres modifications peuvent être faites par l'ingénieur de la connaissance.

A la fin de cette opération, le type de relation termino-conceptuelle courant *typeRTC* est mis à jour dans la liste des types de relation termino-conceptuelles *typeRTC*. La figure 5.16 représente le diagramme d'activités de l'opération *Mise à jour type relation termino-conceptuelle*.

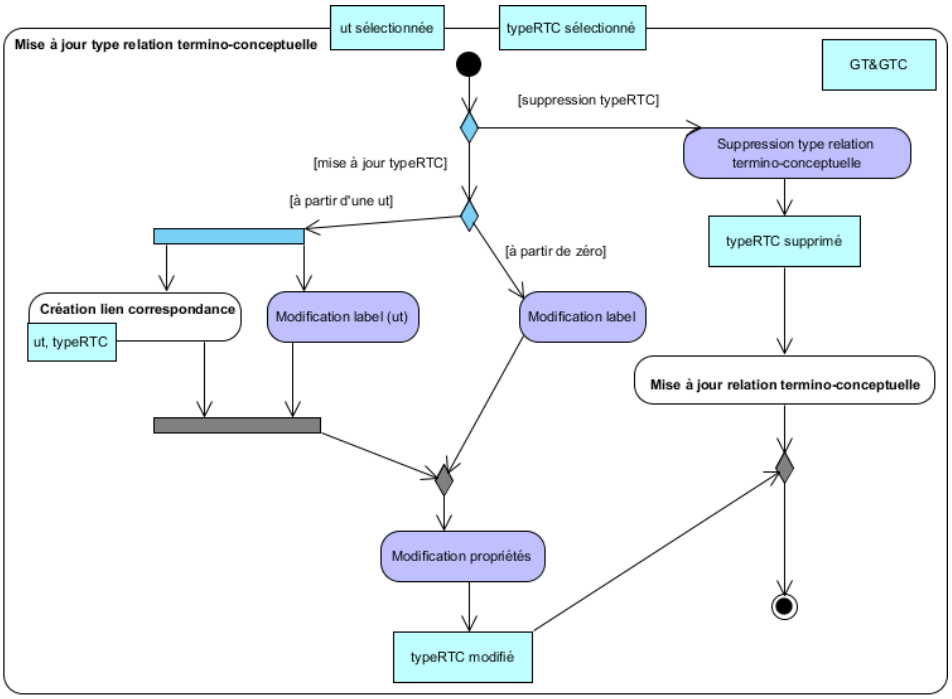

FIGURE 5.16 – Le diagramme d'activités de l'opération *Mise à jour type relation termino-conceptuelle*.

5.5.1.7 Création d'une relation termino-conceptuelle

Cette opération permet la création d'une nouvelle relation termino-conceptuelle $rtc(tc_i, typeRTC, tc_j)$ dans le réseau termino-conceptuel. L'ingénieur de la connaissance a deux choix de création d'une nouvelle relation termino-conceptuelle :

- à partir des termino-concepts et d'un type de relation termino-conceptuelle existants dans le réseau termino-conceptuel ;
- à partir d'une relation terminologique $rt(ut_i, ut_k, ut_j)$.

Dans le premier cas de figure, l'ingénieur de la connaissance choisit un termino-concept tc_i jouant le rôle de source de la relation et un autre termino-concept tc_j jouant le rôle de destination de la même relation. Il choisit aussi un type de relation termino-conceptuelle existant $typeRTC$ et crée le triplet $rtc(tc_i, typeRTC, tc_j)$.

Dans le deuxième cas de figure, la création automatique de la relation termino-conceptuelle est faite à partir d'une relation terminologique sélectionnée. Dans ce cas un termino-concept tc_i jouant le rôle de source de la relation est créé automatiquement à partir de l'unité terminologique ut_i et un autre termino-concept tc_j jouant le rôle de destination de la même relation est créé automatiquement à partir de l'unité terminologique ut_j. Un type de relation termino-conceptuelle $typeRTC$ est créé automatiquement à partir de l'unité terminologique ut_k. Enfin, un lien de correspondance est établi entre la relation terminologique sélectionnée et la relation termino-conceptuelle créée. La figure 5.17 représente le diagramme d'activités de l'opération *Création relation termino-conceptuelle*.

A la fin de cette opération, il y a une nouvelle relation termino-conceptuelle $rtc(tc_i, typeRTC, tc_j)$ dans le réseau termino-conceptuel. Si cette relation est créée à partir d'une relation terminologique alors un lien de correspondance est créé entre ces deux relations dans le graphe $G_T \& G_{TC}$.

Prenons l'exemple où l'ingénieur de la connaissance crée une relation termino-conceptuelle à partir de la relation terminologique $rt(AAdvantagemember, Adheres, AAdvantageprogram)$. La relation termino-conceptuelle créée est composée du termino-concept créé automatiquement à partir de l'unité terminologique *AAdvantage member* jouant le rôle de source de la relation, du termino-concept créé automatiquement à partir de l'unité terminologique *AAdvantage program* jouant le rôle de destination de la même relation et du type de relation termino-conceptuelle créé automatiquement à partir de l'unité terminologique *Adheres* et dont le label a été modifié par l'ingénieur de la connaissance en *AdheresIn*. La nouvelle relation termino-conceptuelle $rtc(AAdvantagemember, Adheres, AAdvantageprogram)$ est créée dans le réseau termino-conceptuel et un lien de correspondance est créé

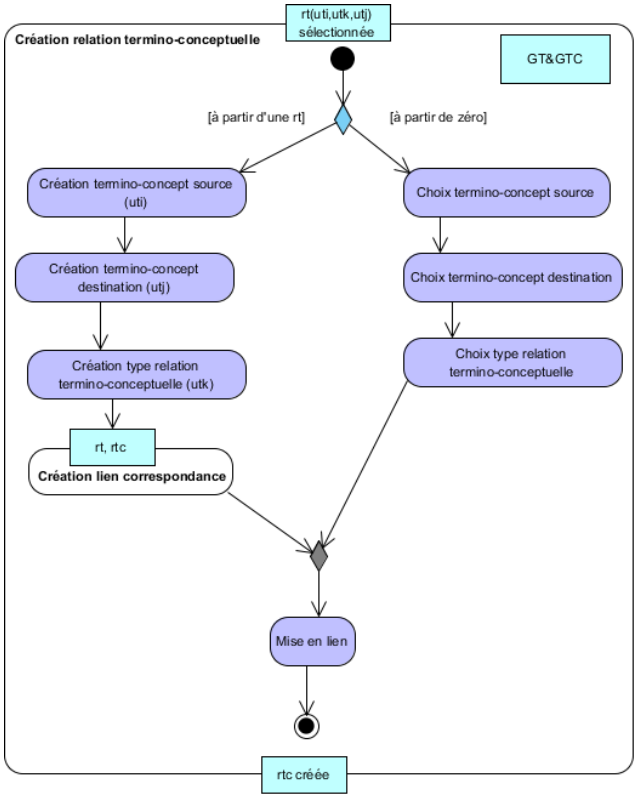

FIGURE 5.17 – Le diagramme d'activités de l'opération *Création relation termino-conceptuelle*.

entre ces deux relations dans le graphe $G_T \& G_{TC}$. La figure 5.18 décrit les deux réseaux G_T et G_{TC} transformés.

5.5.1.8 Mise à jour d'une relation termino-conceptuelle

L'opération *Mise à jour relation termino-conceptuelle* permet la modification ou la suppression d'une relation termino-conceptuelle $rtc(tc_i, typeRTC, tc_j)$ existante dans le réseau termino-conceptuel.

Dans le cas où une relation termino-conceptuelle courante *rtc* est supprimée alors le lien qui relit les termino-concepts source et destination de la relation est supprimé dans le réseau termino-conceptuel. Si une relation termino-conceptuelle possède un ou plusieurs liens de correspondance avec une ou plusieurs relations terminologiques, alors tous ces liens sont supprimés dans le graphe $G_T \& G_{TC}$.

Le deuxième cas de mise jour d'une relation termino-conceptuelle est la

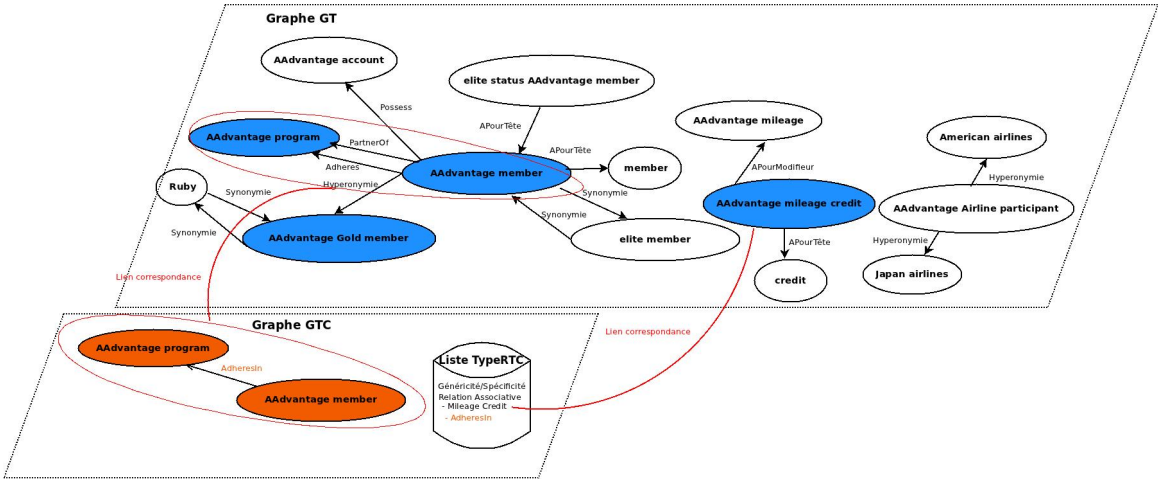

FIGURE 5.18 – Création d'une relation termino-conceptuelle.

modification d'un ou de plusieurs composants du triplet $rtc(tc_i, typeRTC, tc_j)$ à savoir les termino-concepts source ou destination (tc_i, tc_j) ou le type de relation termino-conceptuelle $(typeRTC)$. L'ingénieur de la connaissance choisit le ou les composants à modifier. Par exemple il choisit de mettre à jour le termino-concept source de la relation tc_i en reliant ce dernier avec une unité terminologique. Un cas particulier de mise jour d'une relation termino-conceptuelle consiste à relier cette relation avec une relation terminologique $rt(ut_i, ut_k, ut_j)$ par la création d'un lien de correspondance entre ces deux relations. Les unités terminologiques qui constituent la relation terminologique (ut_i, ut_k, ut_j) sont automatiquement validées. La figure 5.19 décrit l'enchaînement de l'opération *Mise à jour relation termino-conceptuelle* qui est représentée par un diagramme d'activités.

5.5.1.9 Création d'un lien de correspondance

L'opération *Création lien de correspondance* permet la création d'un lien de correspondance entre un élément terminologique et un élément termino-conceptuel au niveau du graphe $G_T \& G_{TC}$. Un élément terminologique correspond à unité terminologique ut ou à une relation terminologique $rt(ut_i, ut_k, ut_j)$. Un élément termino-conceptuel correspond à un termino-concept tc, à un type de relation termino-conceptuelle $typeRTC$ ou à une relation termino-conceptuelle $rtc(tc_i, typeRTC, tc_j)$. La figure 5.20 décrit le

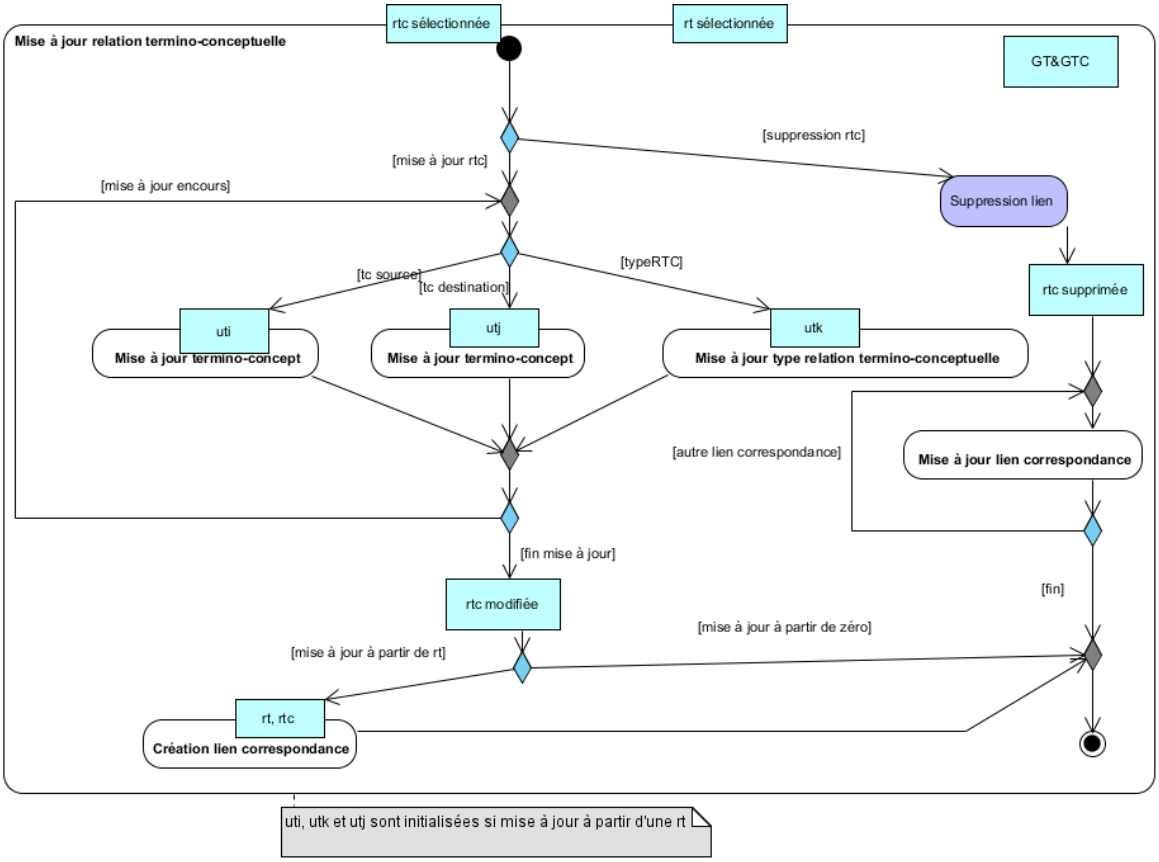

FIGURE 5.19 – Le diagramme d'activités de l'opération *Mise à jour relation termino-conceptuelle*.

diagramme d'activités de l'opération *Création lien de correspondance*. Cette correspondance entre des éléments terminologiques et des éléments termino-conceptuels est décrite dans le chapitre 4 (voir figure 4.13).

La création d'un lien de correspondance nécessite, d'abord, que l'unité terminologique soit validée. L'ingénieur de la connaissance choisit l'unité source et l'unité destination du lien de correspondance. A la fin de cette opération, un lien de correspondance est ajouté entre les unités sélectionnées dans le graphe $G_T \& G_{TC}$. La figure 5.20 décrit l'enchaînement de l'opération *Création lien de correspondance* qui est représenté par un diagramme d'activités.

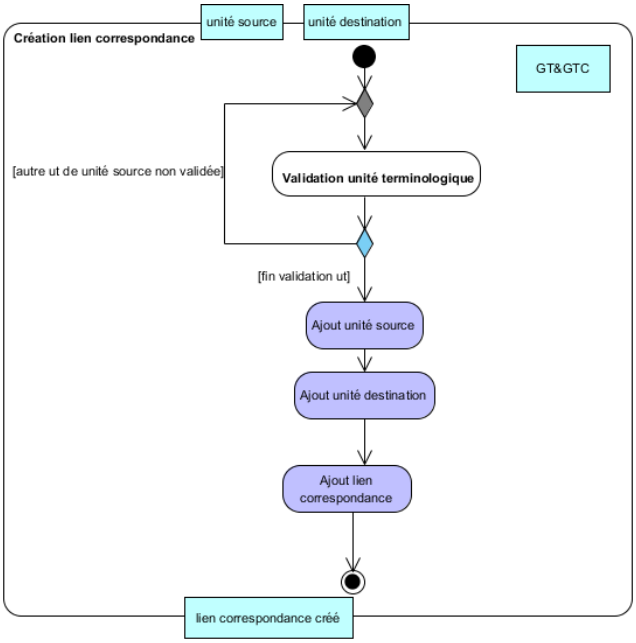

FIGURE 5.20 – Le diagramme d'activités de l'opération *Création lien de correspondance*.

5.5.1.10 Mise à jour d'un lien de correspondance

L'opération *Mise à jour lien de correspondance* permet la modification ou la suppression d'un lien de correspondance dans le graphe $G_T\&G_{TC}$.

Dans le cas où il s'agit de supprimer un lien de correspondance courant, mais les unités liées ne sont pas supprimées du graphe $G_T\&G_{TC}$. Dans le cas où l'ingénieur de la connaissance veut modifier l'une ou les deux unités qui sont reliées par ce lien de correspondance, il choisit dans le graphe $G_T\&G_{TC}$ les nouvelles unités à mettre en correspondance. Parmi les unités terminologiques ou les relations terminologiques, l'ingénieur de la connaissance choisit une unité terminologique ut ou relation terminologique rt comme premier argument du lien de correspondance courant. De la même manière, l'ingénieur de la connaissance choisit un termino-concept tc, un type de relation termino-conceptuelle $typeRTC$ ou une relation termino-conceptuelle $rtc(tc_i, typeRTC, tc_j)$ comme deuxième argument du lien de correspondance courant. La modification de l'un ou des deux arguments du lien de correspondance est optionnel. L'unité terminologique qui a servi comme argument du lien de correspondance est validée automatiquement, de même que toutes les unités extraites d'une relation

terminologique argument du lien de correspondance courant. La figure 5.21 décrit l'enchaînement de l'opération *Mise à jour lien de correspondance* qui est représenté par un diagramme d'activités.

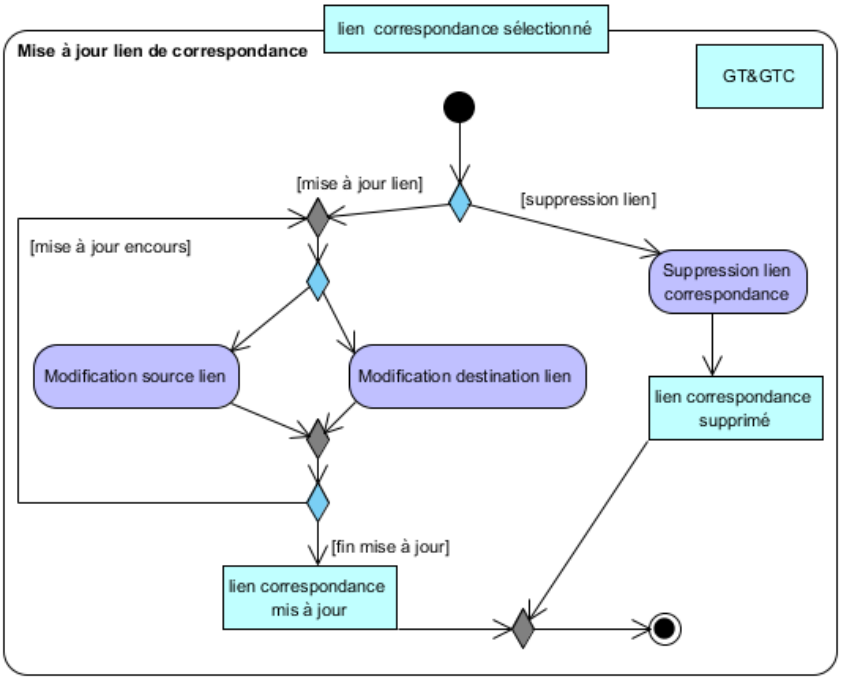

FIGURE 5.21 – Le diagramme d'activités de l'opération *Mise à jour lien de correspondance*.

5.5.2 Opérations composées de normalisation du réseau terminologique

Les opérations élémentaires de transformation du graphe $G_T\&G_{TC}$ créent, modifient ou suppriment des composants du graphe (nœud, arc, triplet). Il faut assurer un enchaînement entre les opérations élémentaires de transformation du graphe $G_T\&G_{TC}$ afin de propager les conséquences de chaque type de transformation sur le graphe et de guider l'ingénieur de la connaissance dans la sélection et la normalisation des unités terminologiques, la mise à jour des unités termino-conceptuelles et leur mise en correspondance. Cette section présente le processus de normalisation.

La normalisation et la mise à jour supposent, d'abord, de fixer une zone d'intérêt dans le réseau. En effet l'ingénieur de la connaissance ne peut pas

analyser tout le réseau à la fois. La sélection d'une « zone d'intérêt » permet d'avoir une vision centrée sur certains éléments du réseau global et de commencer une série de transformations sur ces éléments. Cette zone d'intérêt est représentée par un sous-réseau du réseau global, éventuellement réduit à une seule unité. D'autres configurations d'éléments constituant le sous-réseau sont décrites dans la section 5.5.2.1. Nous définissons les opérations *Sélection graphe de travail* G_T et *Sélection graphe de travail* G_{TC} pour guider l'ingénieur de la connaissance dans la sélection des éléments respectivement à normaliser ou à mettre à jour. Une fois qu'un sous-graphe est sélectionné, l'objectif est donc de le normaliser ou de le mettre à jour. Nous définissons les opérations *Normalisation graphe de travail* et *Mise à jour graphe de travail* respectivement pour la normalisation ou la mise à jour d'un sous-graphe termino-conceptuel.

Afin de guider l'ingénieur de la connaissance, nous définissons des opérations qui permettent de prendre en compte plusieurs indices durant le travail de normalisation afin de juger de la progression de la transformation du réseau $G_T\&G_{TC}$. Ces opérations permettent aussi de vérifier si certaines zones du réseau terminologique ne sont pas encore normalisées dans le but de mettre l'accent sur des unités qui sont potentiellement pertinentes et qui ne sont pas encore normalisées.

Le travail de normalisation est défini par une séquence d'itérations de normalisation du réseau terminologique ou de mise à jour du réseau termino-conceptuel jusqu'à la construction du réseau termino-conceptuel. A chaque itération, des opérations de sélection de sous-graphes terminologiques et termino-conceptuels permettent de sélectionner des zones des graphes globaux. Puis, suivant la nature (terminologique ou termino-conceptuel) du sous-graphe à transformer, des opérations de normalisation et de mise à jour permettent d'appliquer des modifications sur le sous-graphe. A la fin de chaque itération, une opération d'analyse de l'évolution du travail de la normalisation permet de vérifier la progression de la normalisation et guide l'ingénieur de la connaissance dans ses choix de sélection et de transformation pour la prochaine itération. La figure 5.22 décrit le travail de normalisation du graphe $G_T\&G_{TC}$ représenté par un diagramme d'activités.

Dans la suite, nous exposons les opérations composées de normalisation du réseau terminologique et de mise à jour du réseau termino-conceptuel.

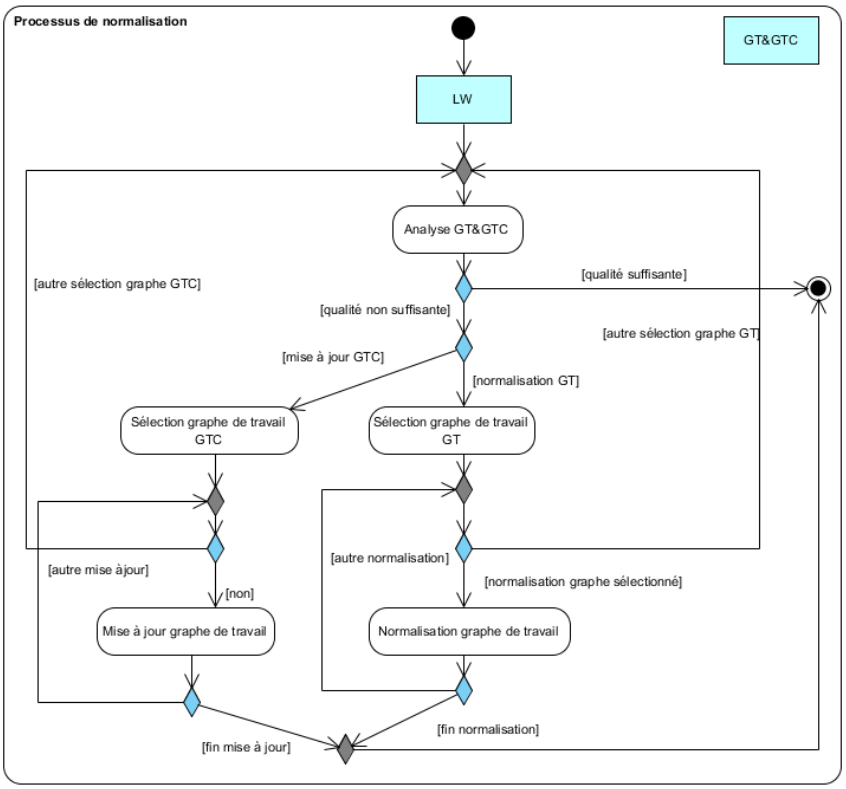

FIGURE 5.22 – Le diagramme d'activités de la normalisation du graphe G_T&G_{TC}.

5.5.2.1 Sélection d'un graphe de travail terminologique

Nous définissons un *graphe de travail* G'_T comme étant un espace de travail pour la normalisation. Il s'agit d'un sous-graphe du graphe G_T ($G'_T(UT', RT') \subset G_T(UT, RT)$) tel que UT' est un sous ensemble des unités terminologiques de UT et RT' est un sous ensemble des relations terminologiques de RT les unités sources et destination faisant partie de UT'.

L'opération *Sélection graphe de travail terminologique* permet à l'ingénieur de la connaissance de sélectionner un sous graphe comme étant un espace de travail pour la normalisation. A chaque itération du travail de normalisation, l'ingénieur de la connaissance décide de la partie du réseau terminologique qu'il veut normaliser à savoir :

- une unité terminologique ;
- un ensemble d'unités terminologiques ;
- une relation terminologique ;
- un sous-réseau terminologique d'une unité terminologique correspondant

à l'ensemble de ses successeurs et prédécesseurs (les voisins directs) ;
- un sous-graphe qui représente le chemin le plus court entre deux unités terminologiques.

Ces configurations de graphes de travail nous ont paru intéressantes à explorer lors de nos expériences mais l'ingénieur de la connaissance est libre de choisir ce qu'il voulait. Par exemple, il peut choisir de normaliser des unités terminologiques ayant un type sémantique pertinent par rapport à l'application visée (par ex. toutes les unités terminologiques ayant comme type sémantique *Organisation*).

A la fin de cette opération, un graphe de travail $G'_T(UT', RT')$ est créé. Les unités terminologiques décrites dans cet espace de travail ont un statut « candidat » sauf celles qui étaient déjà validées (elles ont un statut « validé »). Par défaut, si l'ingénieur de la connaissance ne sélectionne pas de graphe de travail, le sous graphe correspond à l'unité terminologique suivante dans la liste ordonnée des unités terminologiques LW.

La sélection d'un graphe de travail terminologique est représentée par le diagramme d'activités de la figure 5.23.

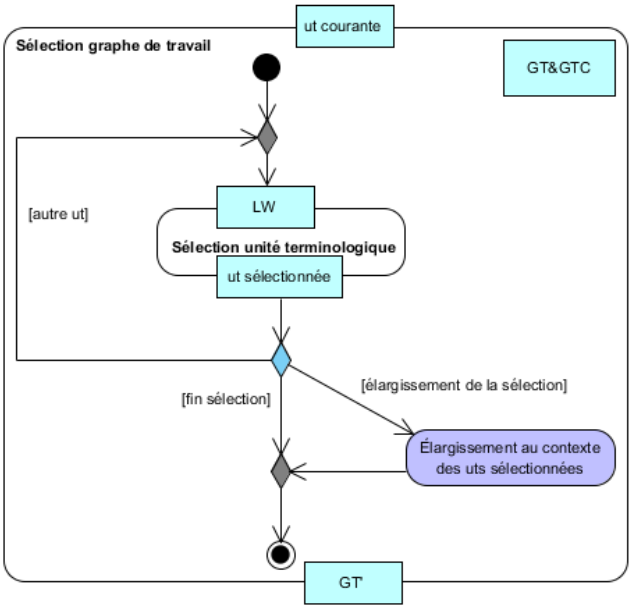

FIGURE 5.23 – Le diagramme d'activités de l'opération *Sélection graphe de travail*.

5.5.2.2 Normalisation d'un graphe de travail

Une fois un graphe de travail terminologique G'_T ($G'_T(UT', RT') \subset G_T$) sélectionné, l'ingénieur de la connaissance procède à le normalisation des unités et des relations qui composent le graphe de travail et qui sont pertinents pour le domaine à modéliser. Il choisit les unités ou les relations terminologiques qu'il veut normaliser. Un graphe normalisé est un graphe dont le travail de normalisation considéré comme achevé par l'ingénieur de la connaissance. En effet, il peut choisir de normaliser une seule unité terminologique ou une seule relation ou tous les composants du graphe de travail. L'opération *normalisation graphe de travail* G'_T est représentée par le diagramme d'activités de la figure 5.24.

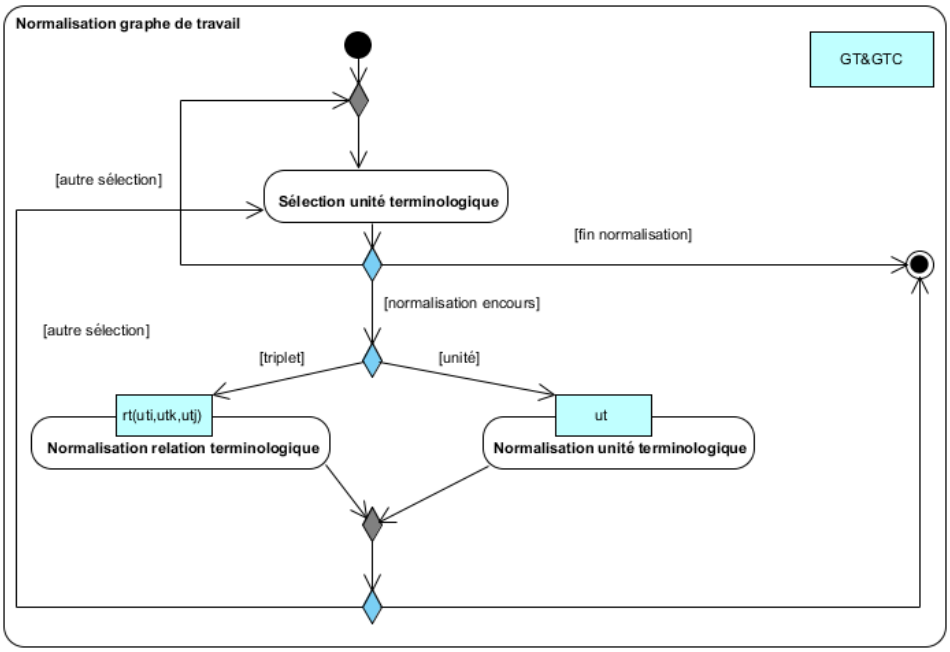

FIGURE 5.24 – Le diagramme d'activités de l'opération *normalisation graphe de travail*.

5.5.2.3 Normalisation d'une unité terminologique

L'opération *Normalisation unité terminologique* permet la normalisation d'une unité terminologique sélectionnée *ut* en un ou plusieurs termino-concepts ou en un ou plusieurs types de relation termino-conceptuelle. L'ingénieur de la connaissance doit, d'abord, valider l'unité terminologique sélec-

tionnée *ut*, puis la normaliser en fonction du domaine et de l'application visée. Il peut par exemple, normaliser une unité terminologique en :

- un ou plusieurs termino-concepts ;
- un ou plusieurs types de relations termino-conceptuelles ;
- un termino-concept et un type de relation termino-conceptuelle.

Quand une unité terminologique est normalisée, elle est reliée par des liens de correspondance à des unités termino-conceptuelles correspondantes. A la fin de cette opération, l'unité terminologique est validée et la ou les unités termino-conceptuelles correspondantes et leurs liens de correspondance sont créés ou mis à jour dans le graphe $G_T\&G_{TC}$. L'enchaînement de l'opération *Normalisation unité terminologique* est décrit par la figure 5.25 qui représente le diagramme d'activités.

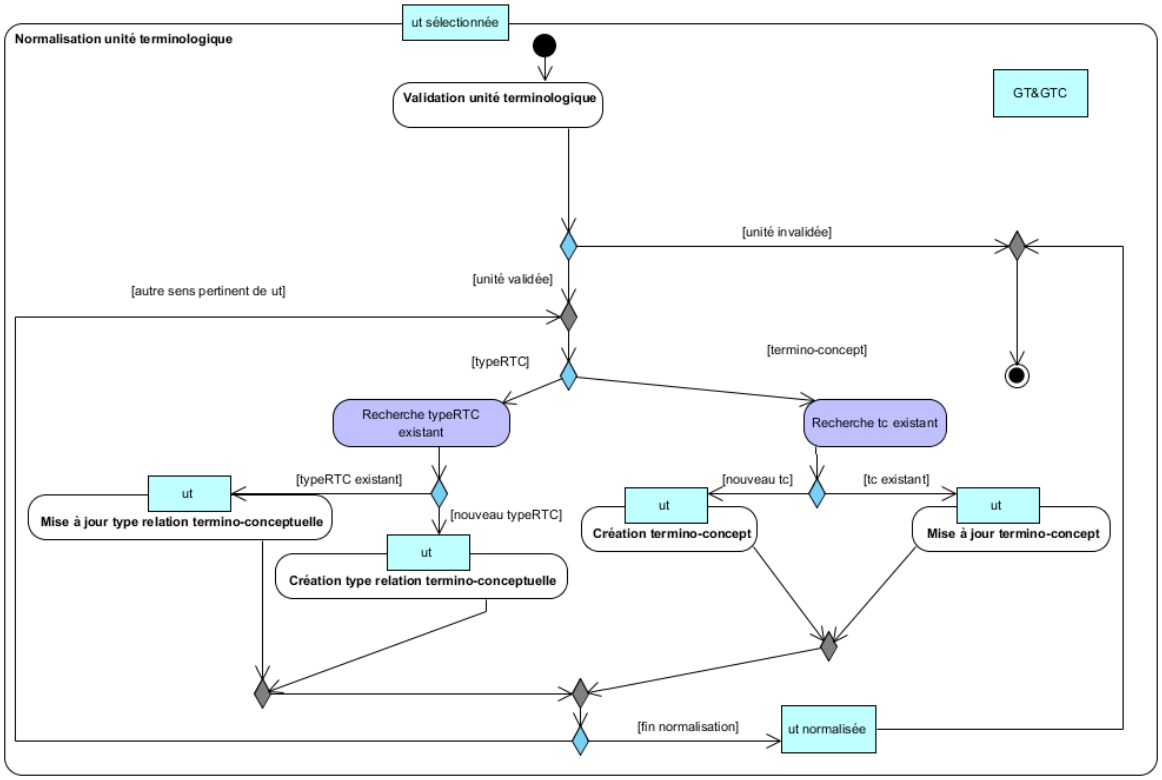

FIGURE 5.25 – Le diagramme d'activités de l'opération *Normalisation unité terminologique*.

Le cas où une unité terminologique est normalisée en plusieurs unités termino-conceptuelles (termino-concept ou type de relation termino-

conceptuelle) est défini lorsque cette unité dénote plusieurs sens pertinents à normaliser dans le texte (unité ambiguë). Ce scénario de normalisation d'une unité terminologique est décrit dans la section 5.6.1. Dans le cas où une unité terminologique dénote un seul pertinent pour le domaine, l'ingénieur de la connaissance décide de la normaliser en un termino-concept ou un type de relation termino-conceptuelle

Prenons l'exemple d'une unité terminologique *AAdvantage mileage credit* normalisée en un type de relation termino-conceptuelle *Mileage credit*. Elle est créée à partir d'une unité terminologique. L'unité terminologique *AAdvantage mileage credit* est d'abord validée par l'ingénieur de la connaissance. Ensuite, l'ingénieur de la connaissance choisit de normaliser cette unité en un type de relation termino-conceptuelle. Il vérifie si ce type n'existe pas dans la liste $TypeRTC$. Dans ce cas, il normalise cette unité terminologique en un type de relation termino-conceptuelle ayant comme catégorie « relation associative ». Il crée également un lien de correspondance entre ces deux unités dans le graphe $G_T\&G_{TC}$. La figure 5.14 décrit le résultat de l'opération au niveau du graphe $G_T\&G_{TC}$.

5.5.2.4 Normalisation d'une relation terminologique

Cette opération vise à normaliser une relation terminologique $rt(ut_i, ut_k, ut_j)$ en une relation termino-conceptuelle $rtc(tc_i, typeRTC, tc_j)$. Il s'agit de normaliser chacune des unités terminologiques qui constituent la relation rt :

- l'unité terminologique ut_i est normalisée en un termino-concept tc_i ;
- l'unité terminologique ut_k est normalisée en un type de relation termino-conceptuelle $typeRTC$;
- l'unité terminologique ut_j est normalisée en un termino-concept tc_j.

L'ingénieur crée une relation termino-conceptuelle $rtc(tc_i, typeRTC, tc_j)$ en indiquant l'ordre des arguments du triplet qui sont créés suite à la normalisation des unités terminologiques. A la fin de cette opération, les unités terminologiques ut_i, ut_k, ut_j sont normalisées en des unités termino-conceptuelles nouvelles ou existantes $tc_i, typeRTC, tc_j$ et la relation terminologique rt est reliée par un lien de correspondance à la relation termino-conceptuelle rtc dans le graphe $G_T\&G_{TC}$. L'enchaînement de l'opération *Normalisation rela-*

tion terminologique est décrit par la figure 5.26 qui représente le diagramme d'activités.

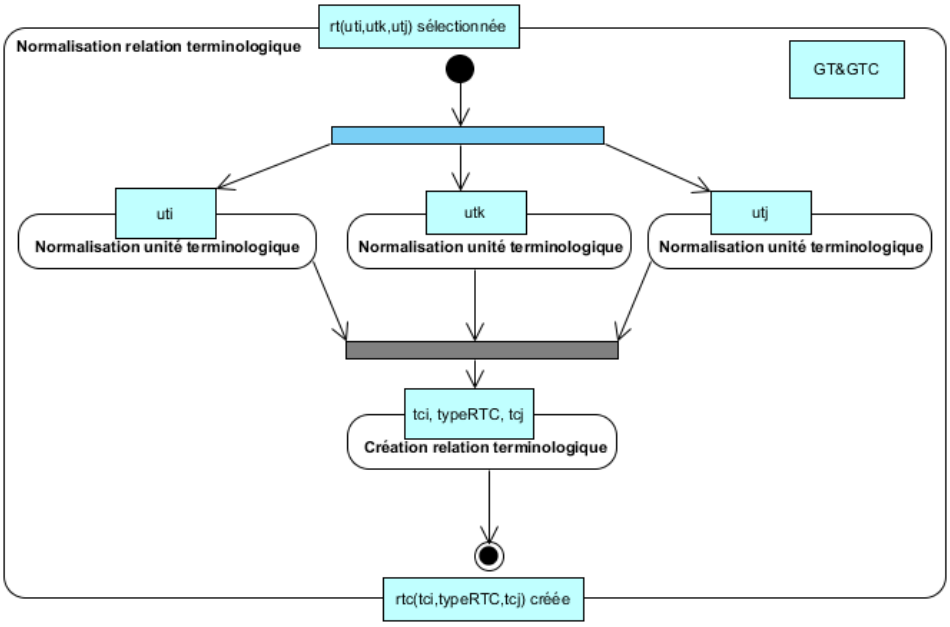

FIGURE 5.26 – Le diagramme d'activités de l'opération *Normalisation relation terminologique*.

5.5.3 Opérations composées de mise à jour du réseau termino-conceptuel

Dans cette section, nous présentons les opérations composées de mise à jour du réseau termino-conceptuel.

5.5.3.1 Sélection d'un graphe de travail termino-conceptuel

Comme dans la sélection d'un graphe de travail terminologique, nous définissons un *graphe de travail* G'_{TC} comme un espace de travail pour la mise à jour comme un sous graphe du graphe $G'_{TC}(TC', TypeRTC', RTC')$ tel que TC' est un sous ensemble des termino-concepts de TC, $TypeRTC'$ est un sous ensemble des types de relations termino-conceptuelles de $TypeRTC$ et RTC' est un sous ensemble des relations termino-conceptuelles de RTC dont les unités sources et destination font partie de l'ensemble TC'.

La sélection d'un graphe de travail termino-conceptuel est décrite par un diagramme d'activités dans la figure 5.27.

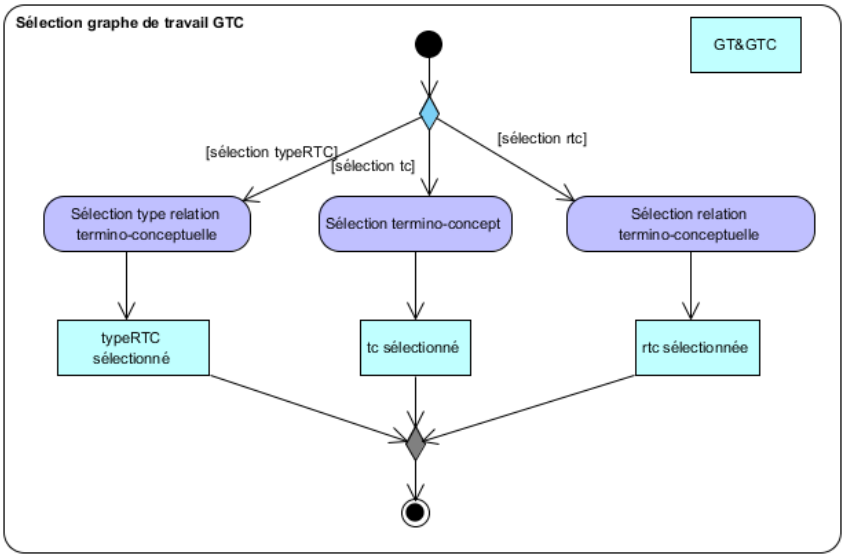

FIGURE 5.27 – Le diagramme d'activités de l'opération *Sélection graphe termino-conceptuel*.

5.5.3.2 Mise à jour d'un graphe termino-conceptuel

L'opération *Mise à jour graphe termino-conceptuel* permet la mise à jour d'une ou de plusieurs unités termino-conceptuelles à savoir de(s) termino-concept(s), de(s) type(s) de relation termino-conceptuelle(s) ou de(s) relation(s) termino-conceptuelle(s) dans le graphe $G_T\&G_{TC}$. Dans le premier cas, la mise à jour d'un termino-concept consiste en la modification des propriétés d'un termino-concept (y compris la création d'un lien de normalisation) ou sa suppression. Le deuxième cas est la mise à jour d'un type de relation termino-conceptuelle. L'ingénieur de la connaissance choisit de modifier ou de supprimer un type de relation termino-conceptuelle. L'opération qui joue ce rôle est *Mise à jour type relation termino-conceptuelle*. Le dernier cas de figure concerne la mise à jour d'une relation termino-conceptuelle. En effet, l'ingénieur peut la mettre à jour en modifiant à jour les termino-concepts source et destination, en changeant son type ou bien en la supprimant du graphe $G_T\&G_{TC}$.

A la fin de cette opération, l'unité termino-conceptuelle (termino-concept,

type de relation termino-conceptuelle, relation termino-conceptuelle) est mise à jour dans le graphe $G_T \& G_{TC}$. L'enchaînement des opérations élémentaires de transformation de l'opération *Mise à jour graphe termino-conceptuel* est représenté par un diagramme d'activités décrit dans la figure 5.28.

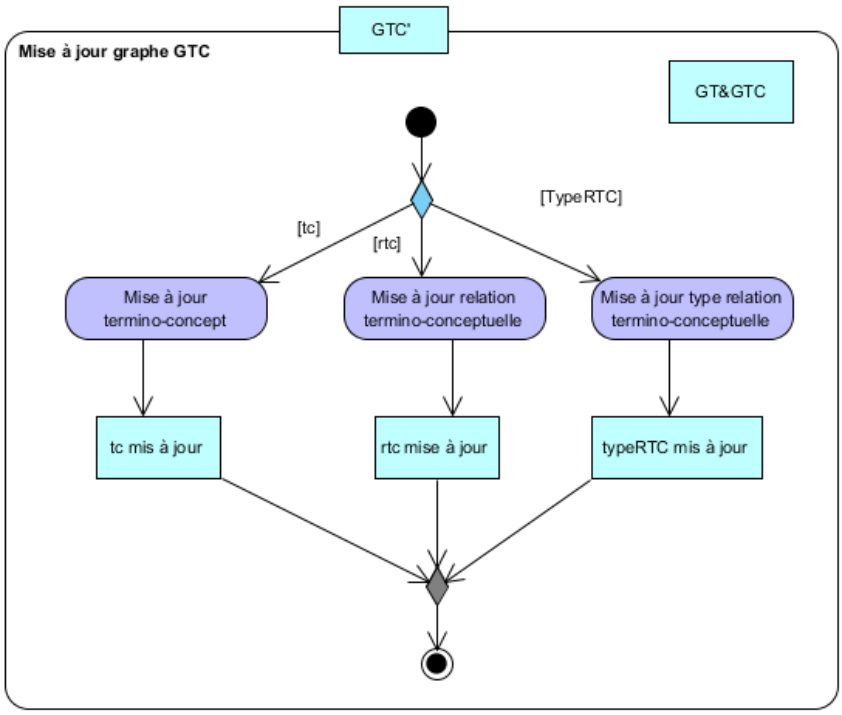

FIGURE 5.28 – Le diagramme d'activités de l'opération *Mise à jour graphe termino-conceptuel*.

5.5.4 Contrôle des réseaux terminologique et termino-conceptuel

Nous avons également proposé des indicateurs de suivi qui permettent à l'ingénieur de mesurer l'avancement du travail de normalisation. Toute la difficulté est effectivement de déterminer à quel moment interrompre le travail de normalisation, étant donné qu'il n'est généralement pas raisonnable de faire ce travail de manière exhaustive mais qu'il faut néanmoins que le résultat soit de qualité suffisante et couvre le corpus d'acquisition et/ou le domaine à modéliser de manière homogène.

Pour guider l'ingénieur de la connaissance dans ses choix de sélection de

graphe de travail terminologique, nous proposons de pondérer les unités terminologiques au niveau du réseau terminologique en fonction de plusieurs critères relatifs au domaine à modéliser et aux caractéristiques du corpus d'acquisition, le but étant d'ordonner des unités extraites du corpus suivant leur pertinence pour le domaine à modéliser. Cette pondération permet à l'ingénieur de la connaissance de sélectionner des unités qui sont potentiellement pertinentes à normaliser pour le domaine. La mesure sémantique que nous proposons comme indice de pertinence est présentée dans la section 5.5.4.1. Dans le processus de normalisation, la priorité est donnée *a priori* aux unités ayant des poids élevés (en appliquant l'indice de pertinence).

Il y a aussi un risque lié à la couverture partielle du réseau terminologique. En effet, durant la normalisation, l'ingénieur de la connaissance peut ne pas repérer certaines unités et relations pertinentes pour le domaine. Cela vient du fait qu'il n'y a pas de contrôle sur l'ordre dans lequel des opérations de normalisation sont définies dans le processus de normalisation. L'ingénieur de la connaissance fait toute une série de choix plus ou moins indépendants. Ces choix n'assurent pas nécessairement une bonne couverture d'exploration du réseau terminologique. Nous proposons des mesures permettant de contrôler l'homogénéité de la normalisation du réseau terminologique et la cohérence de l'analyse terminologique dans le but de s'assurer que, pour les parties explorées, il n'y a pas de transformation importante qui ait été oubliée.

L'enjeu est enfin de savoir à quel moment arrêter le processus de normalisation et comment contrôler ce dernier. Nous essayons de répondre aux questions suivantes dans la deuxième section : quelles sont les mesure d'avancement à définir ? Pour quels intérêts ? Comment les utiliser ?

5.5.4.1 Pondération des unités terminologiques

Nous définissons plusieurs critères de pertinence qui permettent de mettre l'accent sur les unités terminologiques susceptibles de dénoter des notions du domaine. Certains indices sont relatifs aux caractéristiques du corpus d'acquisition. D'autres critères sont reliés aux caractéristiques du domaine à modéliser. Nous présentons les indices suivants :

– **PDomaine**(ut) décrit le poids attribué à une unité terminologique ut par rapport au domaine à modéliser. Ce poids est calculé en fonction du

voisinage de *ut* avec des entités nommées qui relèvent du domaine ;

- **PDiscours(*ut*)** représente la saillance discursive de l'unité terminologique. Celle-ci est appréciée au regard de l'application visée. Comme nous traitons des textes réglementaires qui décrivent des règles métier, l'exploration de ces passages mènent à l'identification des unités terminologiques potentiellement pertinentes pour le domaine ;
- **PDegre(*ut*)** représente la centralité de l'unité terminologique dans le réseau terminologique. Il dépend du degré de l'unité terminologique dans le réseau, c'est-à-dire du nombre de relations terminologiques que l'unité terminologique *ut* entretient avec d'autres unités terminologiques au sein du réseau terminologique. Plus un nœud est relié à plusieurs autres nœuds plus il est considéré comme un nœud central dans le graphe. Par exemple dans le cas d'usage de AA, l'unité terminologique *flight* a comme degré $PDegre(flight) = 15$ car le réseau terminologique, cette unité est reliée aux unités terminologiques *Mileage credit, connecting flight, single plane flight, AAdvantage member, AAdvantage flight, flight award, departure time of the flight, shorter flight, cancelled flight, flight segment, flight departure, flight reservation, flight number, flight document et flight departs.*

L'ensemble de ces indices de pertinence contribue à la définition d'une mesure de pondération sémantique notée $W : UT \rightarrow \mathbb{R}$ qui permet d'attribuer des *poids* $W(ut)$ aux unités terminologiques lors de l'analyse terminologique et qui décrit leur pertinence par rapport au domaine. Nous supposons que plus le poids d'une unité terminologique est élevé plus l'unité terminologique est potentiellement pertinente pour la modélisation du domaine. Les unités terminologiques sont ordonnées suivant leurs poids dans une liste notée LW. Cette mesure est définie comme suit :

$$W(ut) = \frac{Freq_{voisEN}(ut) + Freq_{voisPP}(ut) + D(ut)}{FreqTotale(ut)}$$

où $Freq_{voisEN}(ut)$ est le nombre des occurrences de UT qui sont au voisinage des entités nommées, $Freq_{voisPP}(ut)$ est le nombre des occurrences de UT qui sont mentionnées dans des passages réglementaires, $D(ut)$ est le nombre de relations dont l'unité *ut* joue de rôle de nœud *source* ou *destination* et $FreqTotale(ut)$ est la fréquence totale de *ut* (nombre d'occurrences).

Notre méthode de normalisation propose à l'ingénieur de la connaissance deux manières de parcourir le réseau terminologique :

1. un parcours par ordre de pertinence suivant le poids $W(ut)$. L'ingénieur de la connaissance normalise les unités terminologiques dans le réseau terminologique suivant les poids attribués à ces dernières. En appliquant notre mesure sémantique de calcul de poids $W(UT)$, la sélection d'une unité ut est faite à chaque fois par l'ingénieur de la connaissance telle que :
$$W(ut) = Max(\cup_{i \to n} W(ut_i))$$ [13]
2. un parcours s'appuyant sur les relations de voisinage dans le réseau. L'ingénieur de la connaissance sélection de proche en proche des unités terminologiques. A chaque fois qu'il décide de normaliser une unité ut, l'unité suivante à normaliser est le prédécesseur ou le successeur de l'unité ut.

5.5.4.2 Contrôle du processus de normalisation

La normalisation des éléments terminologiques entraînent des changements dans les deux réseaux terminologique et termino-conceptuel. L'ingénieur de la connaissance doit être informé de la progression du travail de la normalisation. Il doit pouvoir détecter d'éventuelles « anomalies » dans la couverture du réseau terminologique en termes de normalisation, comme par exemple l'existence d'unités encore non normalisées mais potentiellement pertinentes pour le domaine. Contrôler le processus de normalisation permet d'accompagner l'ingénieur de la connaissance dans ses choix de normalisation et d'éviter qu'il « perde de vue » les zones du réseau terminologique intéressantes mais non encore normalisées.

Nous proposons à l'ingénieur de la connaissance des indicateurs qui lui permettent de suivre l'avancement du travail de normalisation. L'idée est qu'il puisse savoir « où il est» pour pouvoir éventuellement arrêter le travail de normalisation quand le travail lui paraît « suffisamment » avancé.

Contrôle des unités pertinentes non encore normalisées Pour assurer un juste équilibre entre les deux stratégies de parcours du réseau terminolo-

13. où n est le nombre des unités terminologiques restant à normaliser.

gique, nous proposons de mettre à jour le poids $W(ut)$ en mesurant le degré d'un nœud par rapport aux relations normalisées. Nous introduisons un nouvel indicateur dit de contrôle de la normalisation ICN_{UT}, qui prend en considération l'effet de la normalisation dans le réseau terminologique. Cet indicateur se calcule de la manière suivante pour une unité terminologique ut :

$$ICN_{ut} = \frac{Freq_{voisEN}(ut) + Freq_{voisPP}(ut) + Freq_{voisUN}(ut)}{FreqTotale(ut)}$$

où $Freq_{voisEN}(ut)$ est le nombre des occurrences de ut qui sont au voisinage des entités nommées, $Freq_{voisPP}(ut)$ est le nombre des occurrences de ut qui sont mentionnées dans des passages réglementaires, $Freq_{voisUN}(ut)$ est le nombre des voisins normalisés de l'unité ut et $FreqTotale(ut)$ la fréquence totale de ut (nombre d'occurrences).

Nous définissons une mesure Γ_{UN} qui décrit le pourcentage des unités normalisées par rapport à celles qui sont pondérées dans la liste ordonnée des unités terminologiques :

$$\Gamma_{UN} = \frac{N_{UN}}{N_{UT}}$$

où N_{UN} est le nombre des unités terminologiques normalisées par l'ingénieur de la connaissance dans la liste ordonnée et N_{UT} est le nombre des unités terminologiques dont le poids ICN_{ut} figure dans l'intervalle $[Max(ICN_{(}ut_i)), Min(ICN_{(}ut_j))]$ où ut_i et ut_j sont des unités normalisées.

Couverture du vocabulaire normalisé par rapport à l'application visée La couverture du réseau termino-conceptuel par rapport au corpus d'acquisition est un indice révélateur de la couverture du vocabulaire normalisé par rapport à celui mentionné dans le texte. Nous définissons l'indicateur σ_{UN2UC} qui décrit le pourcentage des unités normalisées par rapport à celles mentionnées dans le texte :

$$\sigma_{UN2UC} = \frac{UN}{UC}$$

où UN est le nombre des unités normalisées dans le texte et UC est le nombre total des unités candidates extraites du corpus d'acquisition.

Cette couverture peut être calculée par rapport au vocabulaire mentionné dans des passages saillants (par ex. passages réglementaires) si nous avons dé-

tecté au départ ces passages. Elle permet d'apprécier plus finement le réseau termino-conceptuel par rapport à l'application visée. Nous définissons l'indicateur $\sigma_{UN*2UC*}$ qui décrit le pourcentage des unités normalisées par rapport à celles mentionnées dans des passages saillants :

$$\sigma_{UN*2UC*} = \frac{UN*}{UC*}$$

où $UN*$ est le nombre des unités normalisées dans des passages saillants et $UC*$ est le nombre total des unités candidates mentionnées dans des passages saillants.

Couverture du vocabulaire validé au niveau terminologique Durant le travail de normalisation, l'ingénieur de la connaissance valide des unités terminologiques dans le but de les normaliser. Des unités validées peuvent ne pas être normalisées si l'ingénieur de la connaissance décide de poursuivre d'abord son travail de validation. La couverture des unités terminologiques validées par rapport au vocabulaire mentionné dans le texte est donnée par le taux du vocabulaire validé par l'ingénieur de la connaissance par rapport à celui extrait du corpus d'acquisition selon la formule suivante :

$$\sigma_{UV2UC} = \frac{UV}{UC}$$

où UV est le nombre des unités validées dans le texte et UC est le nombre total des unités candidates extraites du corpus d'acquisition.

L'indicateur peut également être calculé par rapport au vocabulaire mentionné dans des passages saillants (par ex. passages réglementaires) si ces derniers ont été déjà identifiées. L'indicateur de couverture $\sigma_{UV*2UC*}$ est défini comme suit :

$$\sigma_{UV*2UC*} = \frac{UV*}{UC*}$$

où $UV*$ est le nombre des unités validées dans des passages saillants et $UC*$ est le nombre total des unités candidates extraites des passages saillants.

5.6 Cas particuliers de normalisation

Dans cette section, nous mettons l'accent sur des scénarios de normalisation qui correspondent à des séquences d'opérations fréquentes qui sont au cœur du travail de normalisation. Nous décrivons les cas de la désambiguisation d'une unité terminologique, du regroupement de plusieurs unités terminologiques jugées comme synonymes par l'ingénieur de la connaissance, et enfin le cas particulier de la normalisation d'une unité terminologique ayant un type sémantique pertinent pour le domaine à modéliser. Nous prenons des exemples du cas d'usage de *American Airlines*.

5.6.1 Désambiguisation d'une unité terminologique

La désambiguisation d'une unité terminologique sélectionnée consiste à identifier les différents sens pertinents à normaliser et à associer chacun des sens pertinents à un termino-concept distinct. L'ingénieur de la connaissance détecte si une unité terminologique est ambiguë s'il identifie plusieurs sens parmi ses occurrences dans le texte ou si l'unité entretient des relations terminologiques qui lui semblent contradictoires ou peu compatibles avec d'autres unités dans le sous-réseau de l'unité ambiguë.

Prenons l'exemple de l'unité terminologique *AAdvantage member* qui décrit à la fois un membre adhérant au le programme de fidélité et une compagnie aérienne participant au le même programme de fidélité. Le tableau 5.14 décrit les propriétés relatives à cette unité terminologique. Si le premier sens de l'unité est celui du « membre adhérant au programme de fidélité », les phrases mentionnées en rouge (gris clair en mode impression sans couleur) révèlent un autre sens pertinent, celui de « compagnie aérienne participant au programme de fidélité ».

L'ingénieur de la connaissance sélectionne dans le réseau terminologique un graphe de travail définissant le voisinage de cette unité (le sous-réseau formé par l'ensemble des successeurs et prédécesseurs de *AAdvantage member*, voir figure 5.29). Celui-ci montre que l'unité terminologique *AAdvantage member* entretient une relation d'hyperonymie avec l'unité terminologique *AAdvantage Gold member* et une relation spécialisée de type *PartnerOf* avec l'unité terminologique *AAdvantage program*. Le rapprochement de ces deux types de relations montre que l'unité terminologique *AAdvantage member* décrit à la

Propriété	Valeur
Label	AAdvantage member
Type sémantique	
Marqueur	AAdvantage members, member, AAdvantage member, members
Occ	All elite status AAdvantage members enjoy complimentary upgrades to the next class of service. Credit is not transferable and may not be combined among AAdvantage members. Non-elite status AAdvantage members must purchase a full-fare Economy Class ticket The minimum mileage amount earned may be less than 500 miles for travel on oneworld member airlines and AAdvantage participating airlines. Mileage will be credited only to the account of the AAdvantage member who flies , rents a car , stays at a hotel or earns mileage. ... members may be prohibited from redeeming mileage credits for an AAdvantage award or ticket. You can also enjoy benefits of tier status on all oneworld alliance members.
Statut	candidat

TABLE 5.14 – Les propriétés de l'unité terminologique *AAdvantage member*.

fois un membre adhérant et un membre partenaire dans le même programme de fidélité.

L'ingénieur de la connaissance décide de ne garder pour le termino-concept

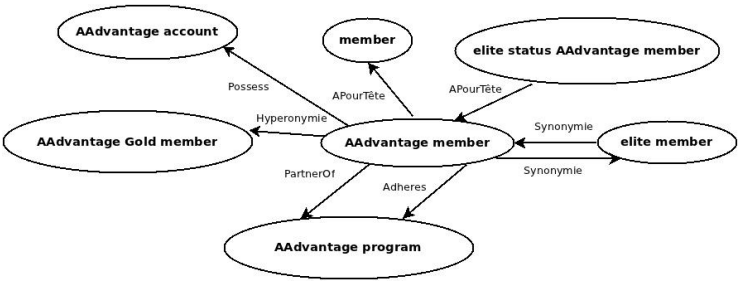

FIGURE 5.29 – Le graphe de travail de l'unité terminologique *AAdvantage member*.

correspondant à l'unité terminologique *AAdvantage member* que les occurrences décrivant le sens de « passager adhérant au programme de fidélité ». Il crée un lien de correspondance entre ces deux unités dans le graphe $G_T\&G_{TC}$ (voir figure 5.30).

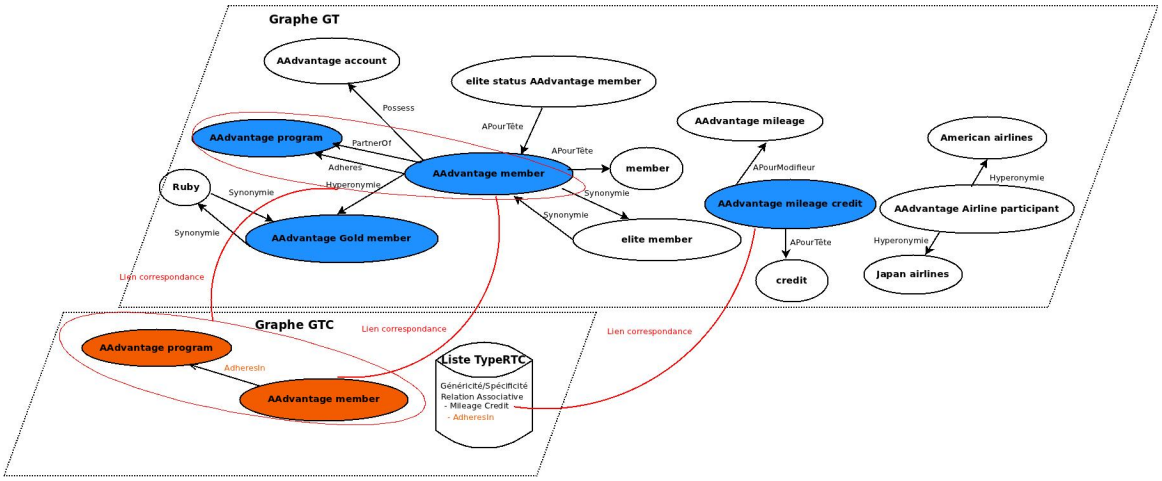

FIGURE 5.30 – Création d'un lien de correspondance au niveau du graphe $G_T\&G_{TC}$.

5.6.2 Regroupement des unités terminologiques

Le regroupement des unités terminologiques est un cas particulier de la normalisation où il s'agit de relier plusieurs unités terminologiques à un même termino-concept si l'ingénieur de la connaissance décide que ces unités terminologiques décrivent un même sens par rapport à un usage précis. Généralement, l'opération de regroupement peut être considérée par l'ingénieur de la connaissance dans les cas suivants :

- deux unités terminologiques sont reliées par une relation de type *Synonymie* ;
- des unités sont des variantes d'une même unité terminologique même si elles n'ont pas été identifiées comme marqueurs de l'unité en question ;
- deux unités terminologiques sont reliées par une relation de type *APourTête/APourModifieur*.

Dans ces cas de figure, le graphe de travail G'_T est un sous graphe formé par un ensemble d'unités terminologiques que l'ingénieur de la connaissance considère comme synonymes. Prenons un exemple où le regroupement de deux unités terminologiques se fonde sur la constatation par l'ingénieur de la connaissance que ces dernières sont synonymes dans le corpus de *American Airlines* : *AAdvantage member* et *AAdvantage Airline participant*. La première opération est la sélection d'un graphe de travail terminologique composé par ces deux unités terminologiques. L'ingénieur de la connaissance sélectionne chacune de ces unités dans le but de valider leurs propriétés (voir tableaux 5.15 et 5.14). Le parcours de leurs occurrences (pour *AAdvantage member*, seules les occurrences coloriées en rouge décrivent une compagnie aérienne) montre qu'il s'agit de deux unités terminologiques synonymes qui désignent les compagnies aériennes participant au programme de fidélité de *American Airlines*.

L'ingénieur de la connaissance décide de valider ces deux unités terminologiques et de les normaliser de telle sorte que l'une des unités terminologiques forme le label préféré du nouveau termino-concept créé et que l'autre soit un label alternatif. L'ingénieur de la connaissance choisit l'unité terminologique *AAdvantage Airline participant* comme label préféré du nouveau termino-concept *AAAdvantage Airline participant* et l'unité terminologique *AAdvantage member* comme label alternatif. La normalisation de ces deux unités terminologiques entraîne la création de deux liens de correspondance entre ces dernières et le nouveau termino-concept. La figure 5.31 représente le graphe transformé $G_T \& G_{TC}$ où les unités terminologiques *AAdvantage Airline participant* et *AAdvantage member* sont validées (les nœuds correspondants sont coloriés en bleu), le nouveau termino-concept *AAdvantage Airline participant* est ajouté au le réseau termino-conceptuel et deux liens de correspondance relient les deux unités validées au termino-concept.

Propriété	Valeur
Label	AAdvantage Airline participant
Type sémantique	Organisation
Marqueur	AAdvantage airline participants, AAdvantage airline participant, AAdvantage Participant Airline
Occurrence	Accruing mileage credit on any AAdvantage airline participant. On American Airlines and other AAdvantage airline participants , you 'll receive AAdvantage mileage credit only for the class of service on which your fare is based when you are ticketed. Each AAdvantage Participant Airline is responsible for its awards only and not for the awards of other participating companies.
Statut	validé

TABLE 5.15 – Exemple d'une unité terminologique *AAdvantage Airline participant*.

5.6.3 Normalisation d'une unité terminologique ayant un type sémantique

Ici ce qui est important, ce n'est pas la valeur sémantique des unités terminologiques entant que telle mais le fait qu'elles soient associées à un type. Les unités terminologiques qui sont extraites par un outil de REN sont souvent associées à un type sémantique. La normalisation d'une unité terminologique ayant un ou plusieurs types sémantiques est un cas particulier du travail de normalisation du réseau terminologique. En effet, l'ingénieur de la connaissance peut normaliser l'unité terminologique et son type sémantique comme deux termino-concepts (nouveaux ou existants) reliés par une relation termino-conceptuelle de type *Généricité/Spécificité*. Prenons un exemple. L'ingénieur de la connaissance sélectionne l'unité terminologique *American Airlines* qui a comme type sémantique « organisation » (voir les propriétés de cette unité

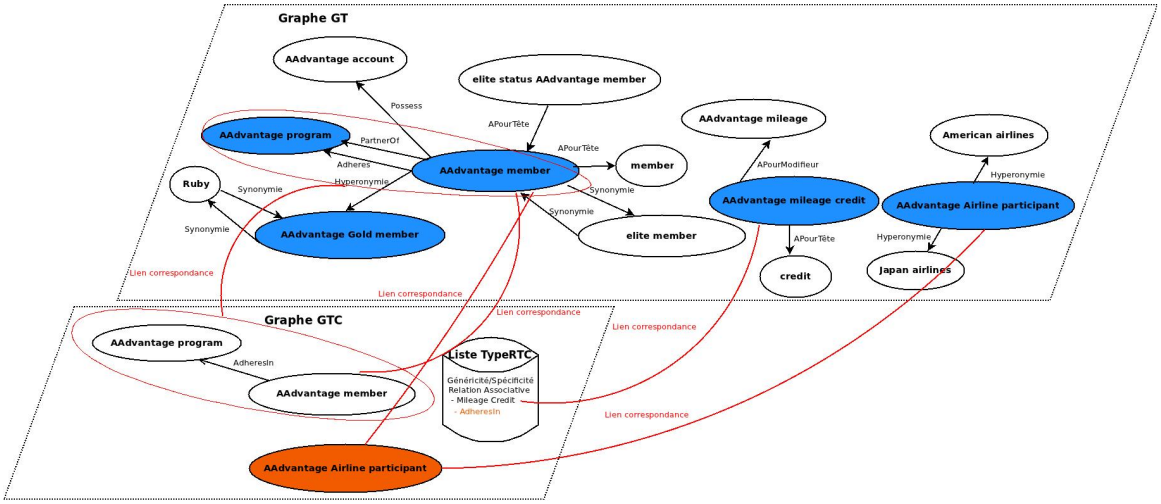

FIGURE 5.31 – Création d'un nouveau termino-concept au niveau du graphe G_T&G_{TC}.

dans le tableau 5.16). Il valide les propriétés relatives à cette unité et décide de la normaliser en un nouveau termino-concept *American Airlines*. En

Propriété	**Valeur**
Label	American Airlines
Type sémantique	ORGANISATION
Marqueur	American Airlines, A.A
Occurrence	On American Airlines and other AAdvantage airline participants, you'll receive AAdvantage mileage credit. A.A reserves the right to end the AAdvantage program with six months notice.
Statut	validé

TABLE 5.16 – Les propriétés relatives à l'unité terminologique *American Airlines*.

plus, il crée un lien de correspondance qui relit ce nouveau termino-concept à l'unité terminologique *American Airlines*. L'ingénieur de la connaissance décide aussi de normaliser son type sémantique (organisation) en un nouveau

termino-concept. Il crée un nouveau termino-concept *Organisation*.

A la fin de la normalisation de l'unité terminologique *American Airlines*, cette dernière est validée et deux nouveaux termino-concepts sont ajoutés dans le réseau termino-conceptuel ainsi que leurs liens de correspondance dans le graphe G_T&G_{TC}. La figure 5.32 décrit le G_T&G_{TC} transformé.

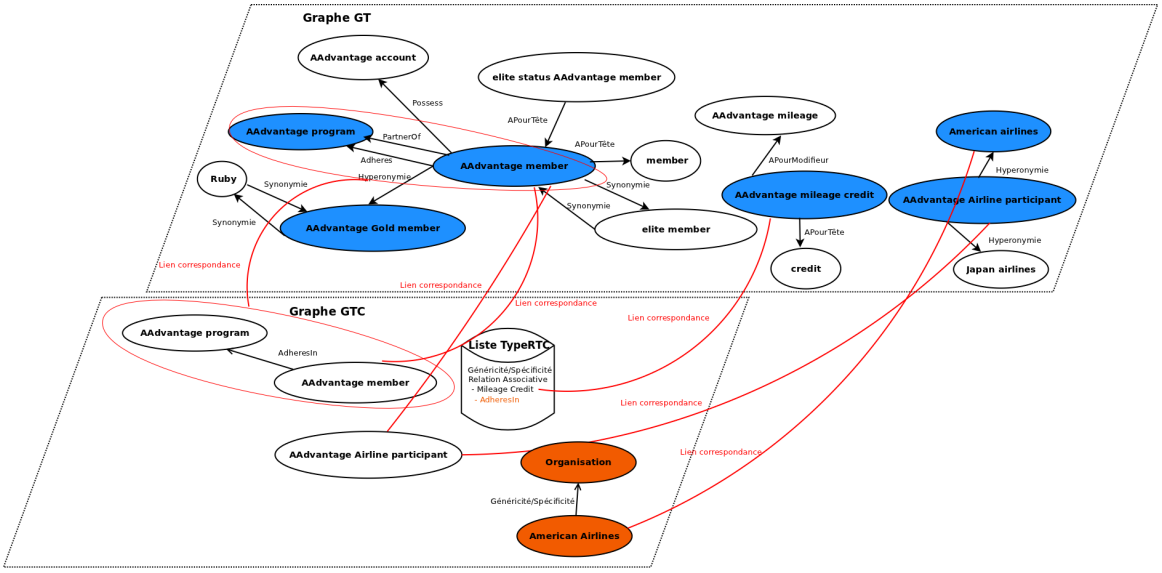

FIGURE 5.32 – Création de deux termino-concepts au niveau du graphe G_T&G_{TC}.

5.7 Conclusion

Dans ce chapitre, nous avons proposé une méthode pour normaliser le réseau terminologique extrait à partir d'un corpus d'acquisition. La normalisation aboutit à la construction d'un réseau termino-conceptuel qui peut servir de point de départ pour la construction d'ontologies de domaine. Contrairement aux méthodes proposées dans l'état d'art, notre méthode combine des termes, des entités nommées et des relations terminologiques au sein d'un réseau pour la construction d'un thesaurus de domaine.

Nous avons défini les opérations élémentaires de normalisation du réseau terminologique et de mise à jour du réseau termino-conceptuel. Ces opérations sont encapsulées dans des macro-opérations permettant la sélection de sous-graphes de travail, la normalisation des unités et des relations terminologiques

et la mise à jour du réseau termino-conceptuel. Nous avons défini un processus itératif de normalisation qui prend en compte ces macro-opérations et qui peut être contrôlé par des indicateurs de l'avancement du travail de normalisation.

Une des propositions originales de cette thèse est la présentation d'une mesure sémantique qui prend en compte différents critères de pertinence liés aux caractéristiques du domaine ainsi que du corpus d'acquisition. Cette mesure permet de donner des poids aux unités terminologiques et de guider l'ingénieur de la connaissance dans ses choix de normalisation. Notre méthode, à la différence de la plupart des méthodes définies dans l'état d'art, fait intervenir l'ingénieur de la connaissance dans toutes opérations de sélection, validation et création d'unités de domaine. Il doit faire des choix de normalisation et vérifier le bon déroulement du processus de normalisation mais il peut aussi s'appuyer sur différents éléments – le poids des unités terminologiques et différentes mesures de couverture – pour se guider dans ce travail. Afin de l'aider, nous lui proposons, tout au long du travail de normalisation, des éléments informationnels qui le permettent de détecter des unités pertinentes.

Chapitre 6

Expérimentations et évaluations

Sommaire

6.1 Introduction

Dans ce chapitre, nous présentons les expérimentations effectuées dans le cadre du projet ONTORULE [1].

Les systèmes de gestion des règles métier (SGRMs) sont des systèmes qui aident les utilisateurs à définir et maintenir des règles métier dans le but d'automatiser des processus de prise de décision. Les règles métier décrivent les aspects métier et les contraintes relatives aux connaissances manipulées dans une organisation. Elles proviennent de différentes sources (réglementations internationales, documents d'entreprises ou directement introduites par des experts métier). Le projet ONTORULE visait à construire des systèmes d'aide à la décision pour des domaines réglementaires et leurs bases de règles métier. L'objectif majeur du projet est de permettre à différents utilisateurs (experts métier, développeurs, opérationnels) de manipuler, maintenir et faire évoluer les bases de règles métier suivant ce que leur rôle leur permet de faire comme opérations. Un expert métier a besoin de moyens qui le guident et lui facilitent les tâches de création et maintien d'une base de règles métier. Un opérationnel qui utilise un SGRMs pour effectuer des tâches liées à la prise de décision, joue un rôle différent : appliquer les règles produites par ce dernier pour prendre des décisions. Le projet ONTORULE devait répondre aux besoins de ces différents utilisateurs en les accompagnant dans l'exécution de leurs tâches. En plus, le projet ONTORULE devait assurer une « séparation » entre la couche métier qui définit les connaissances manipulées et la couche opérationnelle qui manipule ces connaissances au sein d'un système de gestion de règles métier. En effet, dans des systèmes d'information comme les SGRMs il faut faire une distinction entre la couche du système qui gère des connaissances métier et celle qui définit l'ensemble des opérations manipulant ces connaissances dans le but d'assurer une flexibilité pour la maintenance et l'évolution des différentes couches d'un SGRMs.

L'un des objectifs du projet ONTORULE était d'acquérir le vocabulaire conceptuel employé pour exprimer des règles métier à partir des textes réglementaires et d'intégrer ces documents aux SGRMs. En effet, l'acquisition des connaissances pertinentes pour un domaine est un défi majeur des systèmes

1. Ces expérimentations ont été faites dans le cadre du projet FP7 231875 ONTORULE. http://ontorule-project.eu/.

de partage de connaissances. Dans ce contexte, le projet ONTORULE avait pour but d'acquérir le vocabulaire métier à partir des textes réglementaires et d'unifier et de structurer ce vocabulaire sous forme d'une ontologie lexicalisée pour permettre aux gens du métier d'analyser et de modifier des règles métier suivant leur propre vocabulaire. L'utilisation d'une ontologie lexicalisée facilite le partage d'un vocabulaire commun pour écrire des règles métier et permet de garder une trace du vocabulaire dans le texte. Cela permet d'assurer une traçabilité entre les règles métier (décrites d'une manière formelle) et les fragments de textes à partir desquels elles sont dérivées. Ainsi, ce lien entre les passages réglementaires et les règles formelles permet de vérifier la cohérence de la base de règles quand les documents évoluent. Autre avantage est que les utilisateurs retrouvent les mentions textuelles des règles métier formelles qui leurs facilitent leur analyse et leur compréhension.

Les expériences ont été faites sur deux cas d'usage du projet ONTORULE. Un autre cas d'usage a servi de cas d'école au début du projet ONTORULE, celui de American Airlines. Il a été utilisé pour spécifier notre méthode et des exemples étaient extraits de ce cas d'usage et ont été présentés dans le chapitre 5 pour illustrer quelques scénarios du travail de la normalisation. Nous avons aussi utilisé un autre cas d'usage du jeu de Golf pour élargir nos expérimentations avec un corpus d'acquisition différent, plus factuel et en français, même si c'est toujours un texte réglementaire. Les documents utilisés dans nos expériences sont extraits des cas d'usage étudiés dans le cadre du projet ONTORULE. Dans ce contexte applicatif et pour chacun de ces cas d'usage, nous cherchons à construire une ontologie lexicalisée de domaine à partir de textes qui décrivent les règles métier en vigueur. Une ontologie est composée par les concepts qui décrivent des notions du domaine. Chaque concept est lié à un ou plusieurs termes utilisés dans les règles décrites dans le corpus. Ce lien se fait par des termino-concepts.

Ce chapitre est structuré en cinq sections. La première section présente les cas d'usage et les corpus que nous avons utilisés pour mener nos expérimentations. La deuxième section décrit les ontologies de domaine construites pour les deux cas d'usage American Airlines et Audi. Dans la troisième section, nous appliquons notre méthode de construction de réseau termino-conceptuel sur l'un des cas d'usages présentés dans la section précédente. Avant de conclure, la quatrième section décrit les protocoles et les résultats d'évaluation des prin-

cipales contributions de cette thèse : la normalisation des entités nommées et la pondération des unités terminologiques en fonction de leur voisinage avec des marqueurs du domaine.

6.2 Présentation des cas d'usage et des corpus

Dans cette section, nous décrivons les cas d'usage et les corpus que nous avons utilisés comme sources d'acquisition pour la construction de ressources sémantiques (réseau termino-conceptuel, ontologie lexicalisée). Ces corpus sont des textes réglementaires qui sont spécifiques à des domaines précis. Ils sont écrits en anglais ou en français. La taille des corpus d'acquisition varie d'un cas d'usage à un autre.

6.2.1 Cas d'usage American Airlines

Le cas d'usage de American Airlines (AA) vise a développer une application qui permet de gérer les avantages que gagnent un voyageur membre du programme de fidélité AA pendant la période de validité de son abonnement. Le corpus relatif à ce cas d'usage est constitué d'un seul document écrit en anglais, qui fait 11 pages et composé de 243 phrases. C'est un texte de petite taille mis en ligne pour informer les adhérents du programme de fidélité dans le but de les informer des différents statuts et des bonus relatifs à ces statuts. Prenons un extrait du corpus de American Airlines décrit ci-après :

To ensure you receive credit for your AAdvantage transactions, please provide your AAdvantage number when you make your travel reservations or use the services of our participants. Retain all flight documents until the mileage credit appears on your mileage summary. Members may track their summary online at www.aa.com or by subscribing to receive the AAdvantage eSummarySM.

Le tableau 6.1 décrit le corpus AA en termes de nombre de mots, de termes et d'entités nommées extraits. Nous faisons la distinction entre les listes de termes candidats et celle des entités nommées comme nous allons utiliser ces deux types d'unités terminologiques différemment pour la construction d'une

ontologie lexicalisée [2] qui modélise le vocabulaire conceptuel employé pour décrire les règles régissant le programme de fidélité AA.

Corpus	#Mots	#Termes	#Entités nommées
AAdvantage	5 744	973	105

TABLE 6.1 – Le nombre de mots, de termes et d'entités nommées dans le corpus AAdvantage.

Plus spécifiquement, le corpus décrit les règles et conditions d'attribution de « miles » et de leurs avantages pour des voyageurs qui utilisent les services de American Airlines ou d'autres compagnies aériennes participant au même programme de fidélité. Ces règles décrivent les conditions d'attribution de miles (points de fidélité), de leur accumulation, de leur durée de validité et des avantages résultants. Le statut qu'un voyageur possède dépend des miles accumulés. Selon son statut, un membre gagne des bonus ou avantages durant toute la validité de son abonnement au programme de fidélité AA. Par exemple, il peut bénéficier d'offres de réduction de billets d'avion, de séjours en hôtels et de location de voitures. Ces bonus sont offerts à un membre après vérification de plusieurs critères (le nombre de miles cumulés, des pays visités, etc). Nous avons analysé la couverture des règles métier par rapport au corpus d'acquisition en calculant le nombre de phrases qui décrivent des règles métier par rapport aux nombre total des phrases du corpus. Nous avons par ailleurs calculé la couverture des règles métier par rapport au corpus de AA qui est égale à 41% du corpus. Prenons un extrait [3] décrivant des règles métier dans le corpus AA :

If your account has no qualifying activity in any 18-month period, all miles in the account will expire except for those miles earned prior to July 1, 1989 in accounts established prior to January 1, 1989 whose mileage credit will not expire.

2. La distinction entre la liste des termes extraits et celle des entités nommées est valable pour tous les cas d'usage utilisés dans ce chapitre.

3. Les unités terminologiques coloriées en bleu (gris foncé si impression sans couleur) sont celles qui sont reconnues comme des termes, ou des entités nommées par les outils de TAL coloriées en orangé (gris clair si impression sans couleur) quand nous les appliquons à ce corpus.

Earn elite-qualifying points, miles, and segments when you purchase eligible fare tickets and fly on American Airlines, American Eagle, AmericanConnection, Alaska Airlines (including Horizon Air), British Airways, Cathay Pacific Airways,Japan Airlines (including Japan Asia Airways, JALways, Japan Transocean Air, JAL Express, and J-Air), LAN Airlines (including LAN Argentina, LAN Ecuador, LAN Express and LAN Peru), Malév Hungarian Airlines, Qantas Airways, Royal Jordanian, and American Airlines codeshare flights where the ticket reflects an American Airlines coded flight number.

6.2.2 Cas d'usage de Audi

Le cas d'usage de Audi visait à enrichir le système de gestion de règles métier de Audi par le vérifiant dans la conformité des procédures mises en œuvre par l'entreprise par la compagnie Audi avec les normes de sécurité des véhicules définies à l'échelle internationale. Les règles métier contrôlent des techniques de tests utilisées pour la vérification des produits (les véhicules à quatre roues) avant le lancement de leur production. Par exemple, les processus de tests vérifient les paramètres de confort de la conduite et la sécurité du conducteur à partir des modèles définis pour la fabrication des voitures.

Plusieurs difficultés émergent de l'exécution de ces processus de vérification de produits. D'abord, ils utilisent souvent les mêmes paramètres. Ces paramètres sont décrits plusieurs fois par des experts métier ce qui engendre une redondance d'information au niveau de la base de règles métier. De plus, les règles sont écrites dans un langage formel [4]. Les experts métier trouvent une difficulté à écrire ces règles dans un langage en logique puisqu'ils doivent avoir une compréhension préalable du vocabulaire utilisé. Parfois, ils n'arrivent pas à comprendre la sémantique des règles déjà sauvegardées dans une base de règles métier. De plus, des expert métier peuvent utiliser différents labels pour décrire les mêmes connaissances dans l'écriture des règles métier. Par exemple, le terme *ceinture de sécurité* peut avoir des synonymes en contexte comme par exemple *belt, ceinture*. Il y a donc un problème de partage de connaissances et d'hétérogénéité du vocabulaire employé pour décrire les mêmes tests.

Le corpus Audi est un extrait d'une directive internationale qui décrit les

4. Les règles métier sont écrites en F-Logic.

règles et procédures que les véhicules à quatre roues ainsi que leurs équipements doivent satisfaire pour tout ce qui touche aux ceintures de sécurité. Ce corpus est comme le précédent composé d'un seul document mais il fait 130 pages. Prenons un extrait du corpus de Audi :

Two belts or restraint systems are required for the buckle inspection, the low-temperature buckle test, the low-temperature test described in paragraph 7.5.4. below where necessary, the buckle durability test, the belt corrosion test, the retractor operating tests, the dynamic test and the buckle-opening test after the dynamic test. One of these two samples shall be used for the inspection of the belt or restraint system.

Nous nous sommes intéressés plus particulièrement au chapitre 7 de la directive internationale qui décrit l'ensemble des tests appliqués pour vérifier la conformité des ceintures de sécurité à l'égard de la réglementation en vigueur. Ce chapitre est constitué de 138 phrases. Le tableau 6.2 décrit le corpus Audi en termes de nombre de mots, de termes et d'entités nommées extraits.

Corpus	**#mots**	**#Termes**	**#Entités nommées**
Audi	3 704	1003	90

TABLE 6.2 – Le nombre de mots, de termes et d'entités nommées dans le corpus Audi.

Plus spécifiquement, le texte décrit différents tests qui permettent de vérifier l'adéquation des composants matériels des véhicules (chaise, volant, etc) avec des normes définies dans la directive internationale pour assurer le bon fonctionnement des ceintures de sécurité. Nous avons analysé la couverture des règles métier par rapport au corpus d'acquisition en calculant le nombre de phrases qui décrivent des règles métier par rapport aux nombre total des phrases du corpus. La couverture des règles métier par rapport au corpus de Audi est 33.8% du corpus. Prenons l'exemple d'un extrait décrivant des règles métier du corpus de Audi :

During calibration of the stopping device , the speed of the trolley shall be 50 km/h +- 1 km/h and the stopping distance shall be of 40 cm +- 2 cm.

The samples to be submitted to the Micro Slip test shall be kept for a minimum of 24 hours in an atmosphere having a temperature of 20 + 5 CC and a relative humidity of 65 + 5 per cent.
the Exposure test shall proceed continuously for a period of 50 hours, except for short interruptions, to check and replenish the salt solution.

6.2.3 Cas d'usage de Arcelor Mittal

Le cas d'usage de Arcelor Mittal visait à automatiser des techniques de tests pour la vérification de la qualité des produits fabriqués avec les paramètres décrits dans des commandes de clients. Durant le processus de galvanisation (*Galvanisation Line*) de bobines (*coils*), un ou plusieurs défauts peuvent affecter les produits fabriqués. Selon la criticité du défaut par rapport à la commande du client une décision est prise concernant la destination du produit fabriqué : il peut être délivré au client, mis à la poubelle ou adressé à un laboratoire pour réparation. Le cas d'usage de Arcelor Mittal visait à définir une ontologie qui décrit le vocabulaire conceptuel utilisé pour écrire des règles de vérification de conformité des bobines avec des commandes. Un extrait du texte est représenté dans le paragraphe suivant :

Example of Isolated defect for galvanizing line : Dross, Slivers, Blister.. A defect with a severity s+1 has a weight higher than the same defect with a severity s. The letters A, B, C, D describe the aspect level for the isolated defects. "A" aspect is the most severe. If the process parameter Skin-Pass elongation exceeds by 50% or more the maximum elongation of the order during more than 100 meters, then there is a mechanical defect and the coil is set to be repaired. The parameters of the process : Skin Elongation and flattener elongation, are stable within 15% with regard to the average value of the whole product and the temperature measure at the heating exit in a 7%, then the coils will be classified as OK and the assignment of the order will be kept.

Le tableau 6.3 décrit le corpus Arcelor Mittal en termes de nombre de mots, de termes et d'entités nommées.

Nous avons analysé la couverture des règles métier par rapport au corpus d'acquisition en calculant le nombre de phrases qui décrivent des règles métier

Corpus	#Mots	#Termes	#Entités nommées
Arcelor Mittal	569	79	18

TABLE 6.3 – Le nombre de mots, de termes et d'entités nommées dans le corpus Arcelor Mittal.

par rapport aux nombre total des phrases du corpus. La couverture des règles métier par rapport au corpus de Arcelor Mittal est 40% du corpus. Prenons l'exemple d'un extrait décrivant des règles métier du corpus de Audi :

The evaluation of the defects can result in one of three outcomes : The existing defects, if any, can be accepted within the parameters of the order, and the assignment is kept. The existing defects cannot be accepted within the parameters of the order, and the coil is set for scrapping. None of the above can be stated with certainty, and the decision is put off.

6.2.4 Cas d'usage du Golf

Le cas d'usage du Golf est relatif à l'identification et la structuration du vocabulaire termino-conceptuel qui est employé pour décrire les règles du jeu sous forme d'un réseau termino-conceptuel. Ce cas d'usage est un cas d'école que nous avons utilisé pour expérimenter notre approche sur un autre cas d'usage différent des cas d'usage d'ONTORULE décrits précédemment : le corpus du Golf décrit les règles du jeu de Golf destiné aux joueurs participant à une partie de Golf. Il est constitué de 1005 phrases. C'est un corpus écrit en français, factuel et volumineux. Prenons un exemple d'un extrait du corpus du Golf décrit ci-après :

Pénalité pour infractionà la Règle Locale : Match play : perte du trou - Stroke play : deux coups. Etat du terrain : boue, extrême humidité, mauvais état et protection du terrain a. Dégagement pour une balle enfonce. La règle 25-2 prévoit un dégagement sans pénalité pour une balle enfoncée dans son propre impact dans toute zone tondue du parcours. Sur le green, une balle peut être relevée et le dommage causé par l'impact de la balle peut être réparé (règles 16-1b et c). Quand la permission de se dégager pour une balle enfonce n'importe où sur le parcours se justifie, la règle Locale suivante est

recommande : « Sur le parcours, une balle qui est enfonce dans son propre impact dans le sol peut être relevée sans pénalité, nettoyer et dropper aussi près que possible de l'emplacement où elle reposait mais pas plus près du trou ».

Le tableau 6.4 décrit le corpus Golf en termes de nombre de mots, de termes et d'entités nommées extraits.

Corpus	**#Mots**	**#Termes**	**#Entités nommées**
Golf	112898	3711	411

TABLE 6.4 – Le nombre de mots, de termes et d'entités nommées dans le corpus Golf.

Le corpus du Golf décrit des règles métier qui spécifient les règles du jeu. Nous n'avons calculé pour ce cas d'usage la couverture des règles métier parce que le repérage systématique des règles n'est pas fait lors de l'analyse du corpus du Golf. Prenons un exemple d'une règle du jeu de Golf décrite ci-après :

Un golfeur amateur ne doit pas accepter un prix ou un bon d'achat dont la valeur de vente au détail dépasse 500 dollars ou l'équivalent 750 euros, ou toute valeur inférieure qui pourrait être décidée par l'Autorité gouvernante. Cette limite s'applique au total des prix ou des bons d'achat reçus par un golfeur amateur au cours de toute compétition ou série de compétitions.

6.3 Construction des ontologies de domaine : cas d'usage American Airlines et Audi

Nous avons défini notre méthode GRAPHONTO en repensant les étapes définies dans la méthode TERMINAE. Nous avons construit, pour chacun des cas d'usage American Airlines et Audi, une ontologie lexicalisée qui décrit le vocabulaire conceptuel correspondant à l'ensemble des termes qui sont utilisés pour décrire des règles métier dans le corpus d'acquisition. Durant l'analyse du processus de conceptualisation des ontologies, nous nous sommes intéressés à l'amélioration du travail de normalisation qui constitue un passage entre

les niveaux terminologique et conceptuel définis dans la méthode TERMINAE. Notre méthode GRAPHONTO se fonde sur la méthode TERMINAE et se fonde sur des principes qui ont été définis suite aux expérimentations faites sur ces cas d'usage. Dans cette section, nous décrivons les réseaux termino-conceptuels construits et nous évaluons les résultats obtenus par rapport à l'application visée. Nous concluons par dresser un bilan expérimental.

6.3.1 Construction des réseaux termino-conceptuels

Dans le cas d'usage de AA, nous avons créé des termino-concepts à partir des unités terminologiques (termes et entités nommées) qui ont été détectées dans l'analyse terminologique du corpus AA. Par exemple, nous avons repéré les termes *AAdvantage airlines participant*, *member airline* et *Airline representative* qui décrivent le même sens dans le corpus. Nous avons donc associé le termino-concept ***AAdvantage_Airline_Participant*** à ces termes. Nous avons détecté que le terme *AAdvantage member* est polysémique dans le corpus AA. En effet, il décrit à la fois un membre adhérant dans le programme de fidélité AA et une compagnie aérienne offrant des services dans le même programme. Ces deux sens du terme *AAdvantage member* sont pertinents à modéliser pour le domaine. Nous avons créé deux termino-concepts ***AAdvantage member*** et ***AAdvantage_Airline_Participant*** où chacun décrit un sens du terme.

Nous avons créé des relations termino-conceptuelles à partir de l'analyse des relations terminologiques de type *Tête/Modifieur* et de la détection manuelle des relations spécialisées. Par exemple, nous avons créé une relation termino-conceptuelle de type ***Généricité/Spécificité*** entre les termino-concepts ***Member*** et ***AAdvantage_Member*** à partir de la relation terminologique ***AAdvantage member, APourTête, member***. Un autre exemple d'une relation termino-conceptuelle de type *expireOn* a été créée entre les termino-concepts ***AAdvantage_Miles*** et ***18_Month_Period*** à partir de la relation terminologique ***AAdvantage miles, expire, 18 Month Period***.

Dans le corpus de Audi, nous avons détecté certains termes polysémiques. Prenons comme exemple, le terme *Airbag* (coussin gonflable) qui décrit généralement un dispositif qui permet d'assurer la sécurité du conducteur et

celle des occupants d'une voiture. Dans le corpus de Audi, ce terme décrit en plus une des fonctions de sécurité mise en place pour assurer la sécurité des occupants d'un véhicule. Nous avons donc créé deux termino-concepts ***AirbagDevice*** et ***AirbagFunction*** dans le réseau termino-conceptuel et qui correspondent au terme *Airbag*.

Nous avons repéré quelques termes qui sont synonymes. Prenons par exemple les termes *test* et *method* qui décrivent dans le texte la notion de test effectué sur des composants du véhicule. Nous avons associé à ces deux termes, un même termino-concept ***Method*** dans le réseau termino-conceptuel.

Pour la création des relations termino-conceptuelles de type ***Généricité/Spécificité***, nous avons exploité les relations terminologiques de types *Tête/Modifieur* et *Hyperonymie*. Par exemple, nous avons créé les termino-concepts ***VirtualMethod*** et ***PhysicalMethod*** comme des termino-concepts plus spécifiques du termino-concept ***Method*** et qui sont reliés avec ce dernier par une relation termino-conceptuelle de type ***Généricité/Spécificité***. Nous avons créé aussi les termino-concepts ***LowTemperatureTest, BuckleTest*** et ***BreakingStrengthTest*** comme des termino-concepts qui sont plus spécifiques que le termino-concept ***PhysicalMethod***.

Nous avons créé des relations termino-conceptuelles associatives entre des termino-concepts comme par exemple la relation ***assuredBy*** qui relie les termino-concepts ***Function*** et ***Method*** ou encore la relation ***isPartOf*** entre les termino-concepts ***Buckle*** et ***SatefyBeltAssembly***.

Les réseaux termino-conceptuels construits sont présentés dans le tableau 6.5.

Cas d'usage	Réseau termino-conceptuel	#Termino-concepts	#Relations Termino-conceptuelles
AAdvantage	$G_{TC}AA$	182	74
Audi	$G_{TC}Audi$	77	19

TABLE 6.5 – Les réseaux termino-conceptuels construits des deux cas d'usage.

6.3.2 Evaluation des réseaux termino-conceptuels par rapport à l'application visée

Une fois que nous avons créé les réseaux termino-conceptuels relatifs aux cas d'usage, nous avons évalué ces réseaux termino-conceptuels par rapport à l'application visée. Cette évaluation consiste en la couverture du vocabulaire termino-conceptuel par rapport à celui mentionné dans des passages réglementaires. Pour chacun des cas d'usage American Airlines et Audi, nous avons évalué le vocabulaire termino-conceptuel construit par rapport, d'une part, à sa couverture dans le corpus d'acquisition et, d'autre part, à sa couverture dans les passages réglementaires mentionnés dans le corpus d'acquisition. Cette évaluation a fait l'objet d'un article scientifique (Nazarenko *et al.*, 2011). Nous avons appliqué, pour la couverture du vocabulaire termino-conceptuel par rapport au corpus d'acquisition, la formule suivante

couverture $G_{TC}2T$ = AnnTextOcc/TextOcc

telle que $AnnTextOcc$ est le nombre de mots dans le texte qui sont annotés par le vocabulaire termino-conceptuel G_{TC} et $TextOcc$ est le nombre total des occurrences de mots pleins (MP). Seuls les mots ayant comme catégories syntaxiques nom, verbe ou adjectif sont considérés. Puis, nous avons calculé la couverture du vocabulaire termino-conceptuel par rapport au passages réglementaires en appliquant la formule suivante :

couverture $G_{TC}2R$ = AnnRegleOcc/RegleOcc

telle que $AnnRegleOcc$ est le nombre des occurrences de MP qui sont annotés par le vocabulaire termino-conceptuel G_{TC} et $RegleOcc$ est le nombre total des occurrences de MP. Nous avons obtenu les résultats décrits dans le tableau 6.6 en terme de pourcentage du vocabulaire termino-conceptuel mentionné dans le corpus d'acquisition et dans les passages réglementaires.

Dans le cas d'usage de Audi, nous remarquons que 1/3 des mots sont annotés par le vocabulaire termino-conceptuel. Les réseaux termino-conceptuels créés pour les deux cas d'usage couvrent mieux les passages réglementaires que les autres passages de corpus d'acquisition. L'évaluation des réseaux termino-conceptuels montre que les ressources sémantiques construites reflètent bien le vocabulaire utilisé pour décrire des règles métier. En effet, dans les deux cas d'usage, les vocabulaires termino-conceptuels couvrent mieux les passages réglementaires que les autres passages dans le texte (par ex. dans le cas d'usage

Cas d'usage	Couverture RéseauTC par rapport au texte ($G_{TC}2T$)	Couverture RéseauTC par rapport aux passages réglementaires ($G_{TC}2R$)
AA	46.4%	54.8%
Audi	33.8%	40%

TABLE 6.6 – Couvertures des réseaux termino-conceptuels construits par rapport aux textes et aux passages réglementaires des deux cas d'usage.

de AA, le vocabulaire termino-conceptuel couvre de 54.8% les passages réglementaires et de 46.4% le texte).

La construction et l'évaluation des réseaux termino-conceptuels a permis de mettre l'accent sur certains aspects qui peuvent améliorer le processus de conceptualisation d'ontologies lexicalisées (Omrane *et al.*, 2010). Plus particulièrement, nous nous sommes intéressés à la phase de normalisation qui constitue le pont entre la phase terminologique et la phase conceptuelle. Notre objectif étant de définir une méthodologie de normalisation de réseau terminologique qui assure ce passage dans la méthode TERMINAE.

Nous avons remarqué que les entités nommées jouent un rôle particulier dans les textes réglementaires. Outre le fait qu'elles sont utilisées comme marqueurs de domaine qui font référence à des objets du domaine, elles permettent de contraindre certains aspects des règles métier. En effet, les entités nommées mentionnées dans des passages réglementaires décrivent des connaissances importantes à modéliser comme par exemple certaines entités nommées de type DATE qui jouent le rôle de date butoir dans l'exécution des règles métier. Nous avons proposé de considérer les entités nommées comme étant un autre type d'unité linguistique pour la construction d'ontologies à partir de textes. Nous avons défini, dans notre méthode, que à partir des termes et des entités nommées, l'ingénieur de la connaissance crée des termino-concepts et des types de relations termino-conceptuelles. Ces deux types d'unités terminologiques forment un réseau terminologique qui va servir de point de départ pour le travail de normalisation.

La détection des passages saillants dans le texte (passages réglementaires) a permis d'identifier le vocabulaire employé pour décrire ces règles. Nous avons conclu que, dans le cas où on peut identifier ces passages automatiquement, il

est pertinent de détecter les unités terminologiques qui sont mentionnées dans ces passages dans le but de les normaliser. Ces unités normalisées constituent le vocabulaire utilisé pour décrire des règles métier dans le texte. Nous avons enrichi notre mesure sémantique définie dans le chapitre 5 en ajoutant le critère de la fréquence des unités terminologiques figurant dans des passages réglementaires.

Nous avons constaté que, durant la création des ontologies correspondant aux différents cas d'usage, l'ingénieur de la connaissance a besoin d'indices informationnels qui donnent une appréciation du travail de normalisation. Nous avons défini dans notre méthode des indicateurs d'avancement du travail de normalisation dans le but de guider l'ingénieur de la connaissance dans la construction d'un réseau termino-conceptuel.

6.4 Application de la méthode GraphOnto : cas d'usage de Arcelor Mittal

Notre méthode GraphOnto, définie dans le chapitre 5, guide l'ingénieur de la connaissance dans le travail de normalisation d'un réseau terminologique pour la construction d'un réseau termino-conceptuel. Dans cette section, nous exposons le travail de normalisation qui a été fait pour le cas d'usage de Arcelor Mittal. Nous décrivons le réseau terminologique obtenu suite à une extraction automatique du matériau linguistique par des outils de TAL. Puis, nous exposons quelques scénarios de normalisation et nous chiffrons les résultats obtenus. Enfin, nous décrivons le réseau termino-conceptuel créé à la fin du processus de normalisation.

6.4.1 Réseau terminologique

Le réseau terminologique $G_T(UT, RT)$ est construit à partir des résultats des outils de TAL et de reconnaissance d'entités nommées. Nous avons utilisé l'extracteur de termes *YaTeA* pour l'extraction des termes candidats. Prenons comme exemple les termes candidats suivants : *order*, *defect*, *yield strength of the steel* et *chemical composition*. La détection des entités ainsi que de leurs types sémantiques est assurée par la chaîne de traitement *Annie* de la plateforme *Gate*. Le texte de Arcelor Mittal est pauvre en entités nommées (18

entités nommées extraites en total), *Annie* n'a détecté que des entités nommées de type sémantique ORGANISATION, POURCENTAGE et AUTRES TYPE. Une seule entité nommée, de type ORGANISATION a été extraite : *yield*. Trois entités nommées (*50%, 15%* et *7%*), de type POURCENTAGE, ont été extraites à partir du corpus de Arcelor Mittal. Les autres entités nommées extraites par *Annie* correspondent au type sémantique AUTRES TYPE. L'outil de reconnaissance d'entités nommées identifie ces unités terminologiques comme des entités nommées (généralement à cause de la présence d'une majuscule à l'initial) mais n'arrive pas à déterminer leur type sémantique. Il leur associe donc au type sémantique AUTRES TYPES. C'est le cas par de *Skin-Pass*, *Galvanisation Line* et *Slivers*.

Nous avons appliqué manuellement des patrons à base de verbes (sous la forme « SN Verbe SN ») pour l'identification des relations spécialisées. Nous avons ainsi extrait les triplets suivants : (*Galvanisation line, processes, coils*), (*metallurgical structure of the steel, is assigned, to an order*) ou encore (*the defects, present in, the coil*). Le tableau 6.7 décrit le nombre de types de relations terminologiques extraites.

Corpus	#Relations Syntaxiques	#Relations Lexicales	#Relations Spécialisées
Arcelor Mittal	63	1	18

TABLE 6.7 – Nombre et types des relations terminologiques extraites du corpus Arcelor Mittal.

6.4.2 Normalisation du réseau terminologique : scénarios

Nous avons pondéré les unités terminologiques en appliquant notre mesure sémantique $W(UT)$ qui prend en considération trois critères de pondération. Le premier critère décrit par l'indice **PDomaine** est relatif aux unités terminologiques qui sont au voisinage des entités nommées dans le texte. Le deuxième critère décrit par l'indice **PDegré** est relatif aux unités terminologiques qui sont reliées à d'autres unités dans le réseau terminologique. Le troisième critère, décrit par l'indice **PDiscours**, correspond aux unités terminologiques qui sont mentionnées dans des passages réglementaires (*cf* 5).

Nous avons identifié manuellement ces passages en repérant des mots clefs comme par exemple *if, then* et *must*. Le tableau 6.8 représente une partie de la liste des unités terminologiques ordonnée LW_{UT} suivant les poids attribués avec notre mesure $W(UT)$.

Unité terminologique	**ordre par poids** $W(UT)$	**ordre par fréquence**
yield strength of the steel	1	31
lab test result	2	39
order assignment	3	40
long strip of the steel	4	38
Galvanisation line	5	21
coil	6	1
surface aspect	7	44
Thermal cycle	8	42
coating zinc	9	41
assignement of the coil	10	32
strip	11	47
surface	12	46
aspect	13	45
target yield strength	14	33
.....	....	....
defect	40	10

Table 6.8 – Un extrait de la liste ordonnée des unités terminologiques LW_{UT} suivant leur poids $W(UT)$ et leur fréquence.

La figure 6.1 décrit le graphe sous-jacent au réseau terminologique. L'ensemble des nœuds décrivent les termes et les entités nommées extraits du texte de Arcelor Mittal. L'ensemble des arcs sont étiquetés par des types de relations syntaxiques (*APourTête/APourModifieur*), de relations lexicales (*Hyperonymie, Synonymie*) et de relations spécialisées (i.e. *isAssignedTo*).

Nous avons démarré le processus de normalisation en choisissant de normaliser, d'abord, les unités terminologiques ayant les poids les plus élevés et de naviguer dans le réseau terminologique par voisin de plus fort poids. Puis

FIGURE 6.1 – Le graphe sous jacent au réseau terminologique du cas d'usage de Arcelor Mittal.

nous avons exploré le voisinage des unités terminologiques qui sont déjà normalisées pour la détection d'autres unités terminologiques qui peuvent être potentiellement pertinentes à normaliser pour le domaine.

Pour la création des relations termino-conceptuelles, nous avons exploré

les sous-graphes de travail des unités terminologiques normalisées dans le but d'étudier les relations terminologiques qu'elles entretiennent avec d'autres unités terminologiques et de normaliser celles qui sont relatives au domaine.

Dans la suite de cette section, nous décrivons quelques exemples particuliers de la normalisation du réseau terminologique en un réseau termino-conceptuel. Nous mettons l'accent, plus spécifiquement, sur les cas de regroupement des unités terminologiques synonymes, de normalisation des unités polysémiques, de rôle de la pondération des unités et l'étude de leurs voisinage pour le travail de la normalisation.

Exemple de regroupement d'unités terminologiques

Dans le cas de regroupement d'unités terminologiques, l'ingénieur de la connaissance fait correspondre un termino-concept à deux ou plusieurs unités terminologiques qu'il juge synonymes. Prenons par exemple, les deux unités terminologiques *yield strength of the steel* et *yield strength* qui sont reliés par une relation terminologique de type *APourModifieur* et qui décrivent dans le texte le même sens. L'ingénieur de la connaissance sélectionne ces unités terminologiques et visualise leurs propriétés dans le but de les normaliser (voir tableaux 6.9, 6.10).

L'ingénieur de la connaissance regroupe ces deux unités et les relie à un termino-concept ***Yield Strength Of The Steel*** tel que le termino-concept ***Yield Strength Of The Steel*** a comme label préféré l'unité *yield strength of the steel* et comme label alternatif l'unité *yield strength*. Un lien de correspondance est créé entre chacune des unités terminologiques et le termino-concept ***Yield Strength Of The Steel***.

L'exploration d'un sous graphe d'une unité sélectionnée permet à l'ingénieur de la connaissance de vérifier s'il existe des unités qui partagent un même type de relation terminologique et qui peuvent donc être synonymes dans le but de les regrouper. Par exemple, le sous graphe de l'unité *order* est composé de deux relations terminologiques *(order, sets, target)* et *(order, sets, target yield strength)*. Afin de vérifier que ces relations décrivent une même sémantique, l'ingénieur de la connaissance étudie les occurrences des unités jouant le rôle de source et destination de ces relations (voir tableaux 6.11 et 6.12).

Propriété	Valeur
Label	yield strength
Type sémantique	
Marqueur	yield strength
Occurrence	The order sets a target yield strength for the coil , as well as upper and lower tolerances ; if at any point the yield strength is outside of this range there exists a mechanical defect. During the process the mechanical properties , such as the yield strength of the steel are also changed due to the thermal cycle it goes through. Otherwise , if the source data for the yield strength is a model the assignment is set on hold until the lab tests results are available ; if the problem persists when the data from the tests is available the coil is sent to scrapping.
Statut	validé

TABLE 6.9 – Propriétés de l'unité *yield strength*.

L'ingénieur de la connaissance regroupe les deux unités *target* et *target yield strength* pour la création du termino-concept (***Yield Strength Of The Steel***). Il associe l'unité *target yield strength* comme le label préféré du termino-concept créé et l'unité *target* comme son label alternatif. Il créé un lien de correspondance entre chacune des unités terminologiques et le termino-concept créé.

Dans le cas d'usage de Arcelor Mittal, 41 unités terminologiques sont regroupées dont 39 unités qui sont reliées par des relations terminologiques de type *APourTête/APourModifieur* et 2 unités qui sont regroupées car l'ingénieur de la connaissance a détecté qu'elles sont synonymes en parcourant leurs occurrences respectives (les unités *assignement of the coil* et *first assessment*

Propriété	Valeur
Label	yield strength of the steel
Type sémantique	
Marqueur	yield strength of the steel
Occurrence	During the process the mechanical properties, such as the yield strength of the steel are also changed due to the thermal cycle it goes through.
Statut	validé

Table 6.10 – Propriétés de l'unité *yield strength of the steel*.

Propriété	Valeur
Label	target
Type sémantique	
Marqueur	target
Occurrence	The order sets a target yield strength for the coil, as well as upper and lower tolerances ; if at any point the yield strength is outside of this range there exists a mechanical defect.
Statut	validé

Table 6.11 – Propriétés de l'unité *target*.

of the assignement).

Exemple des unités terminologiques ambigues

La normalisation d'une unité polysémique consiste en l'identification des différents sens pertinents à normaliser et à l'association de chaque sens à un termino-concept distinct. Dans le cas d'usage de Arcelor Mittal, l'unité terminologique *assignment* dénote deux sens pertinents pour le domaine. L'ambiguité a été identifiée en explorant le sous graphe de l'unité terminologique *assignment*. La figure 6.2 décrit le sous graphe correspondant. L'unité termino-

Propriété	Valeur
Label	target yield strength
Type sémantique	
Marqueur	target yield strength
Statut	validé
Occurrence	The order sets a target yield strength for the coil, as well as upper and lower tolerances ; if at any point the yield strength is outside of this range there exists a mechanical defect.

TABLE 6.12 – Propriétés de l'unité *target yield strength*.

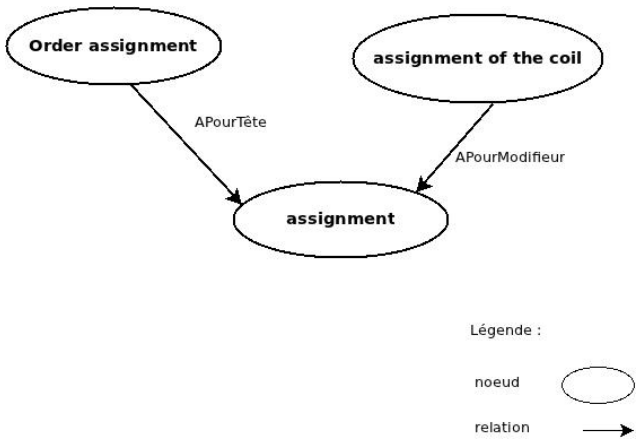

FIGURE 6.2 – Le sous graphe correspondant à l'unité terminologique *assignment*.

logique *assignment* est reliée par une relation de type *APourTête* avec l'unité *order assignement* et par une relation de type *APourModifieur* avec l'unité *assignment of the coil*. Afin d'identifier le ou les sens pertinents à normaliser, l'ingénieur de la connaissance visualise les occurrences des unités terminologiques *assignment*, *order assignement* et *assignment of the coil* (voir tableaux 6.13, 6.14 et 6.15).

L'ingénieur identifie que les deux sens de l'unité *assignement* : affectation de la commande (*order assignement*) et acceptation du produit suite à son analyse (*assignment of the coil*) sont pertinents à normaliser pour le cas d'usage de Arcelor Mittal. Il crée deux termino-concepts distincts. Le premier

Propriété	Valeur
Label	assignment
Type sémantique	Autres types
Marqueur	assignment, Assignment
Occurrence	Order Assignment at Galvanisation Line. When the line finishes the processing of each coil a decision must be made regarding the assignment of the coil : whether it is suitable for the order or not. The procedure then is to evaluate the defects present in the coil to reach a decision : keep the assignment or scrap the coil. The evaluation of the defects can result in one of three outcomes : The existing defects , if any , can be accepted within the parameters of the order , and the assignment is kept. If nevertheless this is the result , the coil must be inspected by an expert and the decision on the assignment be made manually. As long as the defect does affect more than 150 meters of the coil (with a maximum of 100 consecutive meters) , the assignment is kept. Otherwise , if the source data for the yield strength is a model the assignment is set on hold until the lab tests results are available ; if the problem persists when the data from the tests is available the coil is sent to scrapping. Based on this information , a first assessment of the assignment is performed.
Statut	validé

Table 6.13 – Propriétés de l'unité *assignment*.

termino-concept ***Order Assignement*** décrit le premier sens de l'assignement et le deuxième termino-concept ***Assignement Of The Coil*** décrit le deuxième sens. Comme l'affectation du produit fait partie du processus d'affectation de la commande, l'ingénieur de la connaissance crée une relation termino-conceptuelle de type ***faitPartieDe*** entre les termino-concepts créés.

Propriété	Valeur
Label	Order Assignement
Type sémantique	Autres types
Marqueur	Order Assignement
Occurrence	Order Assignment at Galvanisation Line.
Statut	validé

TABLE 6.14 – Propriétés de l'unité *order assignement*.

Propriété	Valeur
Label	assignment of the coil
Type sémantique	
Marqueur	assignment of the coil
Occurrence	When the line finishes the processing of each coil a decision must be made regarding the assignment of the coil : whether it is suitable for the order or not.
Statut	validé

TABLE 6.15 – Propriétés de l'unité *assignment of the coil*.

Deux liens de correspondance sont créés respectivement entre les unités *order assignement*, *assignment of the coil* et les termino-concepts ***Order Assignement*** et ***Assignment Of The Coil***.

Exemple de correspondance entre des unités terminologiques et un type de relation termino-conceptuelle

Dans le cas d'usage de Arcelor Mittal, nous avons détecté une unité terminologique qui est normalisée en un type de relation termino-conceptuelle. En effet, l'unité terminologique *result* décrit un type d'une relation spécialisée dans le texte. Le sous graphe de cette unité terminologique est décrit la figure 6.3. Le type de relation terminologique est identifié par l'étude d'une des occur-

rences de l'unité *result* « The evaluation of the defects can result in one of three outcomes : the existing defects, if any, can be accepted within the parameters of the order, and the assignment is kept.». L'ingénieur de la connaissance identifie dans cette phrase que l'unité *result* est un type de relation terminologique qui relit les unités *evaluation of the defect* et *assignment*. L'unité *result* est normalisée en un type de relation termino-conceptuelle *results* ajouté à l'ensemble des types de relations termino-conceptuelles de $TypeRTC$.

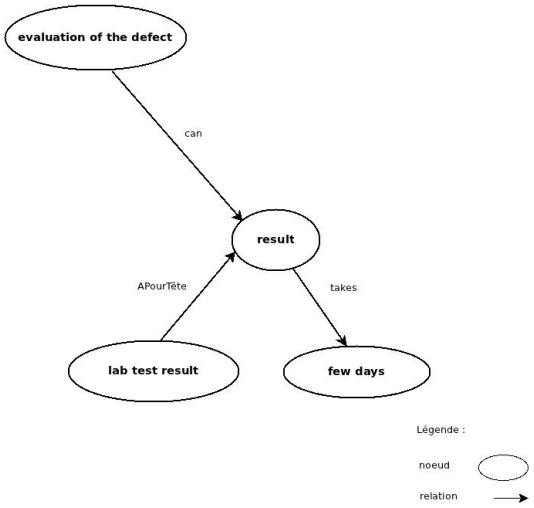

FIGURE 6.3 – Le sous graphe correspondant à l'unité terminologique *result*.

Exemple des unités terminologiques ayant des poids élevés mais qui ne sont pas normalisées

L'ingénieur de la connaissance normalise que les unités dénotant des notions pertinentes pour le domaine et suivant l'application visée. La pondération des unités terminologiques permet de guider l'ingénieur de la connaissance dans la sélection des unités qui sont potentiellement pertinentes à normaliser. Mais parmi la liste des unités terminologiques LW_{UT} ordonnée suivent le poids $W(UT)$, il existe des unités terminologiques qui ne sont pas pertinentes. Prenons par exemple les unités *few day*, *long time* et *estimation of the value* qui se trouvent en tête de la liste ordonnée mais qui ne sont pas normalisées. Elles ont comme statut « invalidé » et se retrouvent en bas de la liste des unités terminologiques [5]. Dans le cas d'usage de Arcelor Mittal, 5 unités ter-

5. Les unités terminologiques invalidées ont un poids égale à zéro.

minologiques ayant des poids élevés ont un statut *invalidé* et ne sont pas donc normalisées.

Exemple des unités terminologiques ayant des poids faibles mais qui sont normalisées

L'indicateur de contrôle de normalisation $ICN(UT)$ permet de vérifier s'il existe des unités ayant des poids faibles mais qui sont au voisinage des unités déjà normalisées. Cet indicateur permet de mettre l'accent sur des unités terminologiques ayant des poids faibles au début de la normalisation. En effet, l'indicateur de contrôle de normalisation réordonne la liste des unités terminologiques en favorisant les unités qui sont reliées à d'autres unités normalisées (*cf* section 5.5.4.2).

A la fin de la première itération du processus de normalisation, nous calculons l'indicateur de contrôle de normalisation dans le but de ré-ordonner les unités terminologiques suivant leur voisinage à des unités déjà normalisées. Nous remarquons, que dans la liste, il y a certaines unités terminologiques qui se trouvent au dessus d'autres unités déjà normalisées mais qui ne sont pas encore normalisées. Il s'avère que ces unités avaient des poids faibles au début du processus de normalisation. En appliquant notre indicateur de contrôle, nous avons réordonné la liste d'unités terminologiques LW_{UT}. Le tableau 6.16 décrit un extrait de la liste des unités terminologiques ordonnée suivant l'indicateur de contrôle de normalisation et qui avaient des poids faibles au début du processus de normalisation.

Unité terminologique	Ordre
zinc	1
defect	2
steel	3
order	4
lab test	5
range	6

TABLE 6.16 – Liste des unités terminologiques ordonnée par l'indice de normalisation.

Prenons l'exemple de l'unité terminologique *mechanical property* qui possède un poids faible et qui n'était pas sélectionnée par l'ingénieur de la connaissance. Après la création du termino-concept correspondant à l'unité *yield strength of the steel*, l'ingénieur de la connaissance étudie le sous graphe de cette unité. Cette unité est reliée par une relation de type *Hyperonymie* avec l'unité *mechanical property* ayant un poids faible. La sélection de l'unité *mechanical property* permet de visualiser ses propriétés (voir tableau 6.17) dans le but de normaliser ou pas la relation terminologique (*yield strength of the steel*, *Hyperonymie*, *mechanical property*).

Propriété	**Valeur**
Label	mechanical property
Type sémantique	
Marqueur	mechanical property, mechanical properties
Occurrence	The order sets the targets for the process : the mechanical properties of the finished product , the thickness of zinc coating. During the process the mechanical properties , such as the yield strength of the steel are also changed due to the thermal cycle it goes through. At this stage process parameters are known , as well as the results from the sensors placed in the line and the models for mechanical properties.
Statut	validé

Table 6.17 – Propriétés de l'unité *mechanical property*.

Malgré que l'unité a un poids faible, l'ingénieur de la connaissance décide que le sens dénoté par cette unité est pertinent et crée le termino-concept correspondant. Il normalise la relation d'hyperonymie qui lit ces deux unités. Il crée deux termino-concepts correspondants aux unités source et destination de

la relation et un type de relation termino-conceptuelle correspondant au type de relation terminologique. Il crée une relation termino-conceptuelle de type ***Généricité/Spécificité*** entre les deux termino-concepts ***Yield Strength Of The Steel*** et ***Mechanical Property***. Il crée aussi un lien de correspondance entre les triplets (*mechanical property*, *Hyperonymie*, *yield strength of the steel*) et (***Mechanical Property***, ***Généricité/Spécificité***, ***Yield Strength Of The Steel***).

Dans le cas d'usage de Arcelor Mittal, 6 unités terminologiques ayant des poids faibles ont été normalisées.

Exemple des unités invalidées ayant des poids faibles

Nous avons défini un poids qui permet de donner une pertinence aux unités terminologiques extraites afin de guider l'ingénieur de la connaissance dans la détection et la sélection des unités à normaliser. Mais dans la liste ordonnée par notre mesure sémantique $W(UT)$, il y a certaines unités terminologiques qui ne sont pas pertinentes pour le domaine. Ces unités ne dénotent pas de sens pertinent pour l'application de vérification de la conformité du produit aux paramètres des commandes. Par exemple les unités terminologiques *first, few, day, strength, estimation, value, finished, et coating* ont été invalidées par l'ingénieur de la connaissance. Dans le cas d'usage de Arcelor Mittal, 18 unités ont un statut invalidé.

6.4.3 Réseau termino-conceptuel

A la fin du processus de normalisation, nous avons construit un réseau termino-conceptuel comportant 26 termino-concepts et 15 relations associatives (voir tableau 6.18). Nous avons créé 7 termino-concepts qui ne sont pas

Réseau Termino-conceptuel	#Termino-concepts	#Relations Termino-conceptuelles
Arcelor Mittal	26	15

TABLE 6.18 – Le réseau termino-conceptuel obtenu en termes de nombre de termino-concepts et de relations associatives.

créés à partir d'unités terminologiques comme par exemple le termino-concept

Assignee qui décrit les différents affectations possibles d'un produit fabriqué ou encore le termino-concept ***Phenomenon*** qui décrit un phénomène détecté dans un produit. La figure 6.4 décrit un extrait du réseau termino-conceptuel créé.

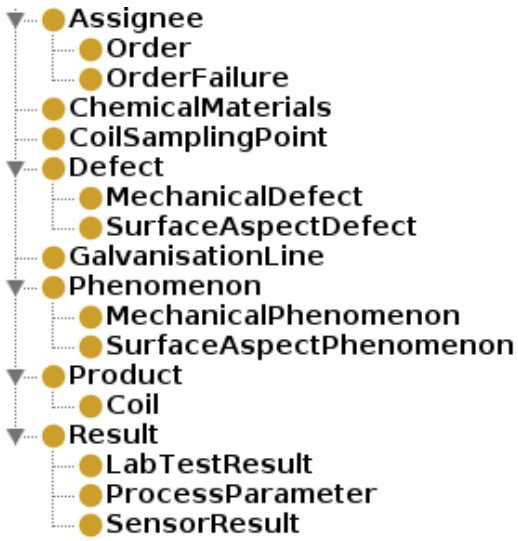

Figure 6.4 – Un extrait du réseau termino-conceptuel du cas d'usage de Arcelor Mittal.

Nous avons calculé la couverture du réseau termino-conceptuel par rapport au corpus d'acquisition en appliquant la formule $\sigma_{UN2C} = \frac{UN}{C}$, avec UN est le nombre des unités normalisées dans le texte et C est le nombre total des unités candidates du corpus d'acquisition. Nous avons obtenu une couverture égale à 51%. Par ailleurs, nous avons essayé d'apprécier la couverture du réseau termino-conceptuel par rapport aux passages saillants afin de vérifier que le vocabulaire normalisé construit couvre bien l'application visée. Nous avons calculé cette couverture grâce à la formule $\sigma_{UN2PP} = \frac{UN}{PP}$, avec UN est le nombre des unités normalisées dans des passages saillants et PP est le nombre total des unités candidates mentionnées dans des passages saillants. Le réseau termino-conceptuel couvre de 75 % le vocabulaire mentionné dans des règles métier. Nous remarquons que le vocabulaire normalisé couvre bien le vocabulaire mentionné dans les passages réglementaires plus que sa couverture par rapport au corpus d'acquisition. Notre mesure sémantique $W(UT)$ aide à l'identification des unités terminologiques qui sont employées pour décrire des règles métier dans le corpus comme le poids défini prend en compte

la fréquence des unités terminologiques dans des passages réglementaires.

6.5 Evaluation de la méthode GraphOnto

L'évaluation d'une méthode de construction de ressources termino-ontologiques ou ontologiques n'est pas une tâche bien définie dans le domaine de l'ingénierie des connaissances (Bourigault & Aussenac-Gilles, 2003). Il faut faire une distinction entre l'évaluation d'un outil de construction d'ontologies et l'évaluation des ontologies construites suivant un domaine et une application visée. Généralement l'évaluation d'une méthode consiste à collecter les appréciations des utilisateurs cibles à l'aide d'un questionnaire établi autour des fonctionnalités définies dans une méthode (Presutti, 2008). Mais le résultat de l'évaluation des avis des utilisateurs est souvent biaisé par leurs connaissances préalables du domaine et leur expertise dans le domaine de l'ingénierie des connaissances.

Une autre manière d'évaluer consiste à évaluer la ressource construite par rapport à une référence ou à l'application visée. L'ontologie créée est évaluée, par exemple, par un expert du domaine qui vérifie si cette conceptualisation répond aux besoins de l'application cible ou pas. Il est cependant difficile de mesurer l'importance de l'aide que procure une méthode de construction d'ontologies à l'ingénieur de la connaissance. En effet, la ressource était construite dans un contexte précis et il est parfois difficile de spécifier les conditions requises pour définir un cadre expérimental d'évaluation qui prend en considération plusieurs tests pour rendre compte de tous les points de vue de cette évaluation (par ex. tâche, besoin métier, coût d'exécution de requêtes). De plus, la construction de ressources sémantiques dépend de la méthode utilisée et du point de vue adopté par l'ingénieur de la connaissance, de ses choix de modélisation et de ses connaissances du domaine.

Sans prétendre ajouter une solution globale au problème de l'évaluation de l'acquisition de connaissances, nous avons cherché à évaluer certaines de nos propositions les plus précises : la pondération des unités terminologiques et la prise en compte des entités nommées durant le processus de normalisation. Nous proposons, pour chaque évaluation, un protocole d'évaluation qui compare le résultat obtenu par rapport à une référence du domaine en question. Nous évaluons aussi, pour les deux cas d'usage American Airlines et

Audi, les réseaux termino-conceptuels créés par rapport à l'application visée, la construction d'un vocabulaire normalisé permettant d'écrire des règles dans les textes réglementaires.

6.5.1 Evaluation de l'intérêt des entités nommées pour la construction d'ontologies

L'une des hypothèses de travail définie dans notre méthode GraphOnto est la considération des entités nommées comme des unités terminologiques appartenant au réseau terminologique servant à construire le réseau termino-conceptuel. Les entités nommées jouent le rôle de marqueurs de domaine car elles font référence à des entités du domaine. Cet aspect référentiel a fait que ce type d'unité linguistique a de l'intérêt pour la création d'un vocabulaire normalisé de domaine.

Dans le cas des textes réglementaires, nous avons remarqué que les entités nommées sont des unités linguistiques particulières qui posent des contraintes pour la description ou l'exécution des règles métier. En effet, dans le cas d'usage de American Airlines, certaines entités nommées ayant comme types sémantiques DATE ou POURCENTAGE contraignent l'exécution de certaines règles métier. Prenons l'exemple suivant tiré du corpus de American Airlines où une entité nommée de type sémantique DATE joue le rôle d'une date butoir pour la validité des miles dans un compte d'un adhérant et des mentions d'entités nommées de type sémantique POURCENTAGE qui spécifient la valeur du bonus gagné suivant le statut d'un adhérant : *If your account has no qualifying activity in any 18-month period, all miles in the account will expire except for those miles earned prior to July 1, 1989...The bonus earned is based on your elite status, as follows : AAdvantage Executive Platinum 100%, AAdvantage Platinum 100% and AAdvantage Gold 25%.*

Durant le travail de normalisation du réseau terminologique, à partir des entités nommées et de leurs types sémantiques, l'ingénieur de la connaissance crée des termino-concepts et des types de relations termino-conceptuelles. Nous évaluons l'intérêt de considérer des entités nommées au même niveau que les termes, c'est-à-dire au début du processus de normalisation, par rapport à la construction d'ontologies lexicalisées de domaine. Nous nous sommes intéressés aussi à la formalisation des termino-concepts qui correspondent à

des entités nommées et à l'enrichissement d'ontologies existantes en tenant en compte des entités nommées décrites dans un corpus d'acquisition.

6.5.1.1 Protocole d'évaluation

Isoler l'apport spécifique des entités nommées est difficile parce qu'elles ne peuvent pas être totalement dissociées des termes dans le travail de normalisation mais nous essayons néanmoins de le cerner pour les cas d'usage de AA et Audi. Pour évaluer notre approche, nous comparons les résultats que nous obtenons à une « référence » proposée par un expert du domaine. Selon les cas, l'expert a réellement construit une ontologie et nous prenons la liste des concepts comme référence ou bien il a formalisé une liste de règles métier qui nous a permis d'établir un vocabulaire conceptuel, ce qui correspond à une liste de termino-concepts. Pour comparer un résultat à une référence, nous avons utilisé les mesures de précision, de rappel et de F-mesure qui servent couramment en recherche d'information pour évaluer les résultats d'un système avec une référence. Ces mesures se définissent comme suit :

$$\textit{Précision} = \frac{\text{UTP}}{\text{UT}} \; \textit{Rappel} = \frac{\text{UTP}}{\text{UP}}$$
$$\textit{F-mesure} = \frac{2*\text{Précision*Rappel}}{\text{Précision+Rappel}}$$

où UP, UT ou UTP sont respectivement le nombre d'unités pertinentes (par ex. figurant dans la référence), le nombre d'unités trouvées (par ex. figurant dans le résultat) et le nombre d'unités à la fois trouvées et pertinentes.

6.5.1.2 Cas d'usage American Airlines

Dans le cas d'usage de American Airlines (AA), nous avons créé deux ontologies lexicalisées. Une première ontologie AA_1 a été construite en considérant que les termes extraits à partir du corpus d'acquisition. Puis, une deuxième ontologie AA_2 a été construite sur la base de l'enrichissement de l'ontologie AA_1 en considérant pour la construction des concepts et des instances les entités nommées extraites du corpus d'acquisition. Dans le cas de AA, la référence (AA_{Ref}) est une liste de termino-concepts qui décrit le vocabulaire conceptuel des règles métier. Nous avons construit deux ontologies, à partir des termes seuls (AA_1) et en prenant les entités nommées en compte dans la conceptualisation (AA_2). L'évaluation consiste donc à comparer les valeurs obtenues après l'analyse des entités nommées à l'ensemble de départ.

L'analyse du corpus AA donne une liste de 973 termes candidats. Après élimination des termes bruités et regroupement des variantes, nous avons obtenu une nouvelle liste de 680 termes qui nous a permis de créer une première ontologie nommée AA_1. Dans un deuxième temps, nous avons pris les entités nommées en considération. 105 entités nommées ont été extraites du corpus AA. Nous avons conceptualisé certaines entités nommées comme des concepts. L'entité nommée *Central America* qui, au-delà de sa caractéristique référentielle d'un territoire regroupant des villes (Costa Rica, El Salvador), indique un ensemble d'aéroports spécifiques qui jouent un rôle important dans l'attribution des miles. Le sens d'aéroport d'Amérique centrale renvoie à une notion qui n'avait pas émergée lors de la première analyse de la liste initiale d'unités terminologiques. Nous avons donc créé le concept **Central America** et nous lui avons associé des instances (les aéroports).

La chaîne de reconnaissance des entités nommées *Annie* a extrait par ailleurs un certain nombre de noms de compagnies aériennes, comme *American Connection*, *American Eagle*, *Japan Airways*, *Alaska Air Group*, etc. Ces compagnies aériennes participent au programme de fidélité AA. Nous avons ajouté ces entités nommées comme des instances à un concept **AAdvantage_Airline_Participant** déjà existant dans l'ontologie AA_1. Le concept **AAdvantage_Airline_Participant** que nous avions déjà créé à partir du terme *AAdvantage participant* a comme instances toutes les compagnies aériennes qui participent au même programme de fidélité (ex. *American Eagle*, *Japan airways*). Les types sémantiques associés aux entités nommées ont en outre permis de créer des concepts. Le type ORGANISATION associé aux compagnies aériennes a ainsi donné le concept **Organisation**, père du concept **AAdvantage_Airline_Participant**.

La détection des entités nommées contribue enfin à une meilleure compréhension du domaine et à la création de nouveaux concepts. Par exemple, la reconnaissance des entités nommées *Sapphire* et *Ruby*, que nous avions considérées comme du bruit dans la première analyse terminologique du corpus, a permis de détecter des catégories de statuts de voyageurs avec des règles d'attribution de miles et de bonus différentes. L'exploration des contextes de ces deux entités nommées a permis de mieux comprendre la réglementation et d'intégrer les concepts suivants à l'ontologie initiale (figure 6.5) :

– **Elite_Member** regroupe tous les statuts qu'un membre peut avoir ;

– **Benefit** décrit les avantages dont un membre peut bénéficier suivant son statut ;
– **Numerical_Quantity** correspond aux différents montants de bonus (points, segments ou miles) que peuvent gagner des voyageurs privilégiés lors de leurs voyages ;
– **Account** désigne un compte d'un membre contenant les bonus accumulés.

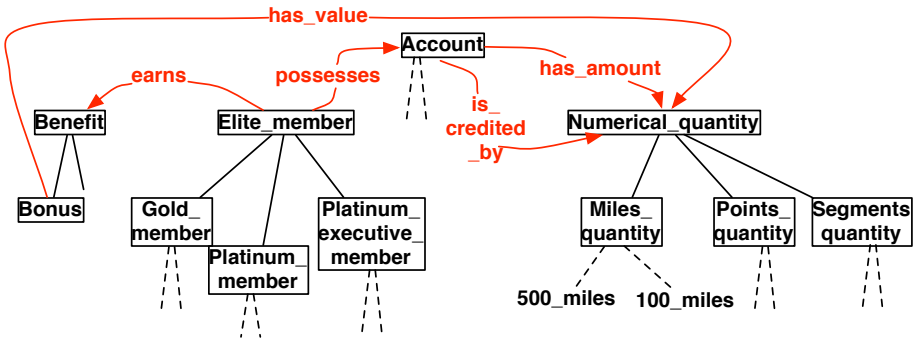

FIGURE 6.5 – Extrait de l'ontologie AA

La détection des entités nommées a permis d'identifier des concepts importants pour la modélisation du domaine. Par exemple, selon les cas, un membre ayant le statut Ruby peut gagner pour un vol aller-retour 25 000 miles, 25 000 points ou 30 segments, ce qui a conduit à introduire des propriétés (**earns**, **possesses**, **has_value**, **has_amount**) et des restrictions de valeur de ces relations conceptuelles dans l'ontologie créée. Par exemple un membre ayant comme statut **Platinum_Member** est relié avec la relation **earns** avec le concept **Benefit** telle qu'une restriction sur le nombre de miles gagnés qui est égale à 50 000 miles. La figure 6.6 décrit un extrait de l'ontologie créée.

Par rapport à l'ontologie AA_1 qui contient 130 concepts, 7 nouveaux concepts ont été ajoutés dans AA_2, 15 concepts existants ont été redéfinis et 45 instances ont été ajoutées. Nous avons considéré comme bruit les entités nommées de villes car elles ne sont pas pertinentes pour l'application visée mais toutes les autres entités nommées détectées (60% du total) ont été ajoutées d'une manière ou d'une autre à l'ontologie. Puis, nous avons procédé à une comparaison de chacune des ontologies créées AA_1 et AA_2 avec l'ontologie de référence AA_{Ref}. Le résultat obtenu pour le cas d'usage est présenté dans

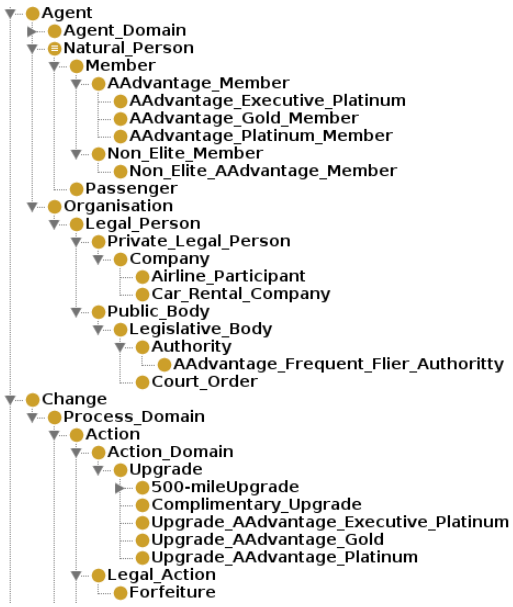

FIGURE 6.6 – Extrait de l'ontologie AA.

le tableau 6.19.

Ressources construites		**Références**	**Mesures**	
Ontologies	**Listes extraites**		**Précision**	**Rappel**
Ontologie AA_1	$LTCAA_1$	LTC AA_{Ref}	82,8%	67,5%
Ontologie AA_2	$LTCAA_2$		83%	72%

TABLE 6.19 – Evaluation de l'ontologie construite au regard de la référence : mesures de précision et rappel. A partir de l'ontologie sont extraites des listes de termino-concepts (LTC) qui sont comparées avec une référence.

Nous remarquons qu'il existe peu d'écart entre la liste enrichie $LTCAA_2$ et celle de départ $LTCAA_1$ au niveau des valeurs de précision et rappel. La précision reste stable, de 82,8% à 83%, et le rappel augmente légèrement, de 67,5% à 72%. Ceci s'explique par le nombre relativement faible d'entités nommées dans le corpus (105 entités nommées extraites au total). L'analyse détaillée des deux ontologies montre néanmoins que l'exploration des entités nommées dans le texte a permis de restructurer l'ontologie initiale (11 % de l'ontologie initiale a été restructurée) et de l'enrichir avec de nouveaux concepts et relations. Nous pouvons dire que la considération des entités nommées du-

rant la phase de conceptualisation a donc contribué à identifier des éléments pertinents du domaine (concepts) et à peupler partiellement l'ontologie. Beaucoup d'entités nommées extraites du texte se sont révélées pertinentes (60 % des entités nommées extraites sont conceptualisées). Le tableau 6.20 décrit l'ontologie enrichie AA_2.

Ontologie	#Concepts	#Instances	#Relations associatives
AA_2	210	45	74

TABLE 6.20 – Le nombre de concepts, instances et relations de l'ontologie AA_2 du cas d'usage de AA.

6.5.1.3 Cas d'usage Audi

Dans le cas d'usage de Audi, nous avons créé une ébauche d'ontologie en démarrant la phase de conceptualisation par la considération des entités nommées dans le texte et des termes qui sont au voisinage de ce type d'unité linguistique. Dans cette expérimentation, nous mettons l'accent sur le fait que les termes qui sont au voisinage des entités nommées (apparaissant dans une même phrase) sont potentiellement pertinents pour le domaine à modéliser. En effet, nous considérons que les entités nommées sont des marqueurs de domaine qui confèrent une valeur sémantique particulière à leurs contextes et appuient la pertinence des termes avoisinants. La détection de certaines mentions d'entités nommées qui sont relatives au domaine, dans le cas d'usage de Audi, a permis de démarrer la conceptualisation en contribuant à mettre l'accent sur des termes pertinents du domaine. Nous avons démarré le processus de conceptualisation de l'ontologie de Audi durant une seule itération en considérant d'abord les entités nommées durant l'analyse terminologique. Puis, en partant des entités nommées et en explorant leurs contextes, nous avons analysé les termes qui sont au voisinage de ces entités nommées comme c'est décrit dans la figure 6.7. En pratique, l'ingénieur de la connaissance n'explore pas le réseau terminologique de manière aussi systématique. L'expérience décrite ici a donc été conduite de manière artificielle pour des fins d'évaluation.

Nous comparons la liste des termino-concepts créés à une liste référence

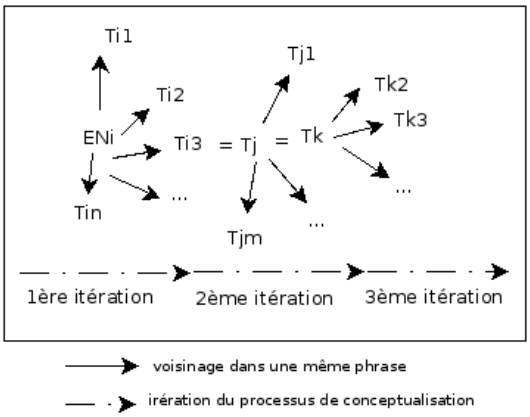

FIGURE 6.7 – Exploration des termes au voisinage des entités nommées. Pour une entité nommée EN_i, l'ingénieur de la connaissance explore les termes voisins directs (de T_i à T_n) durant la première itération du processus de construction d'ontologies puis les voisins à l'ordre 2 au cours de la deuxième itération, etc.

composée par le vocabulaire termino-conceptuel qui a été utilisé pour décrire des règles métier dans le texte. La référence est une liste de concepts extraite de l'ontologie fournie par l'expert. Nous comparons le résultat obtenu $LCAudi$ à la référence $Audi_{Ref}$.

Nous avons commencé le processus de conceptualisation en ne considérant au départ que les entités ayant comme type sémantique PERSONNE, DATE, POURCENTAGE et AUTRES TYPES. Nous avions en effet remarqué que ces entités nommées figurent souvent dans des passages réglementaires et qu'elles sont importantes à modéliser parce qu'elles introduisent des valeurs particulières qui jouent des rôles clefs dans les règles métier. Par exemple, certains chiffres en pourcentage contraignent des opérations de tests pour la ceinture de sécurité. Prenons l'exemple suivant tiré du cas d'usage de Audi où la valeur *65 + 5 per cent* indique le taux d'humidité que doit avoir l'atmosphère pour protéger la sangle (strap) « *The strap shall be kept for a minimum of 24 hours in an atmosphere having a temperature of 20 + 5 C and a relative humidity of 65 + 5 per cent.* ».

Grâce à l'exploration des contextes des entités nommées, nous avons identifié les noms des tests (*Calibration test, Dynamic test, Corrosion test*) qui permettent de vérifier si les équipements de voiture respectent les normes ou pas. Les entités nommées de type POURCENTAGE ont permis de créer des propriétés de concepts de domaine qui servent à vérifier si les équipements

d'une voiture satisfont les critères de sécurité. Ce type d'entité nommée est mentionné dans le corpus avec des valeurs spécifiques que les méthodes de test doivent vérifier, ce qui a conduit à créer les propriétés correspondantes : **length of a strap** (longueur de la sangle), **hasTemperature**, **hasHumidity** et **hasSalt** (a pour sel).

Les entités nommées de type AUTRES TYPES [6] sont aussi des unités spécifiques au domaine. Prenons comme exemple, les entités nommées apparaissant sous la forme de littéraux dans le texte : *M1, N1*, identifient des catégories spécifiques de véhicules. Ces entités nommées sont importantes pour la conceptualisation du domaine car il existe un ensemble de règles et procédures qui peuvent être paramétrées (température, durée, etc.) selon la catégorie du véhicule. De même, les entités nommées telles que *Point A, Point C* font référence à des positions exactes dans la ceinture de sécurité. Ces positions doivent être modélisées puisque les tests et procédures décrits dans la directive vérifient des paramètres qui dépendent de ces positions. Dans le cas d'Audi, nous avons construit une ébauche d'une ontologie composée de 53 concepts de domaine en une seule itération du processus de conceptualisation et en nous appuyant uniquement sur les entités nommées et les termes figurant dans leur voisinage. Cette expérimentation a été décrite dans un article scientifique (Omrane *et al.*, 2011a). La figure 6.8 décrit un extrait de l'ontologie créée. Le tableau 6.21 décrit le nombre de concepts, instances et relations de l'ontologie Audi.

Ontologie	#Concepts	#Instances	#Relations associatives
Audi	77	31	19

TABLE 6.21 – Le nombre de concepts, instances et relations de l'ontologie Audi du cas d'usage de Audi.

On obtient dans ce cas d'usage une précision de 72,2% mais un rappel de 67,5%. Le vocabulaire conceptualisé dans l'ontologie de Audi correspond globalement à celui mentionné dans les passages réglementaires. Mais l'ontologie créée ne couvre pas tout le vocabulaire utilisé pour décrire des règles métier. Cette constatation est prévisible comme l'ébauche de l'ontologie a été

6. AUTRES TYPES est un type sémantique associé aux entités nommées auxquelles aucun type sémantique n'a pu être attribué.

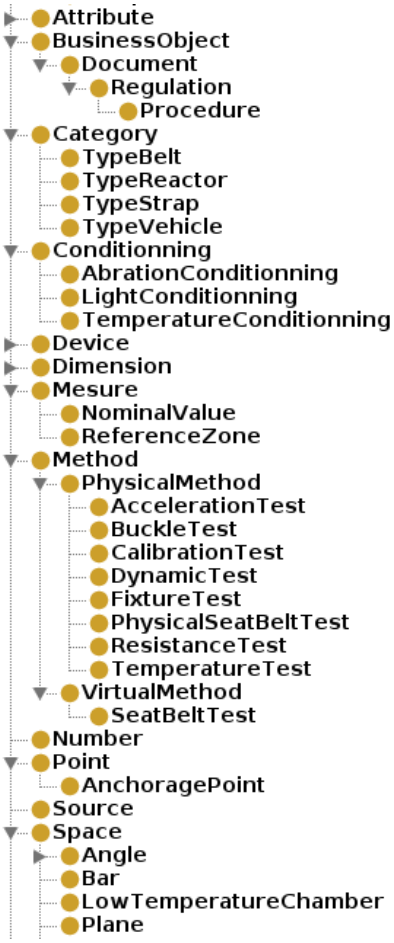

Figure 6.8 – Extrait de l'ontologie Audi.

créée durant une seule itération du processus de conceptualisation. Il faut ré-itérer le processus de construction pour détecter d'autres unités terminologiques qui sont susceptibles de dénoter des notions pertinentes au domaine. L'ontologie que nous avons construit est formée de concepts « centraux » du domaine. En effet, nous avons créé certains concepts du domaine qui représentent les concepts jouant le rôle de concepts pères pour d'autres concepts dans l'ontologie de référence. Prenons comme exemple le concept **Device** créé dans l'ontologie de Audi et qui correspond au concept père des concepts **Motor, Reactor, ChildRestrainSystem** de l'ontologie de référence. Ces chiffres montrent l'intérêt de l'utilisation des entités nommées dans le démarrage du processus de conceptualisation : on obtient une ébauche d'ontologie déjà proche de celle de l'expert. Le résultat obtenu pour le cas d'usage présenté

figure dans la tableau 6.22.

Résultats		Références	Mesures	
Ontologie *Audi*	*LCAudi*	LC $Audi_{Ref}$	72,2%	67,5%

TABLE 6.22 – Evaluation de l'ontologie construite au regard de la référence : mesures de précision et rappel. A partir de l'ontologie sont extraites des listes de concepts (LC) ou de termino-concepts (LTC) qui sont comparées avec une référence.

6.5.1.4 Filtrage des entités nommées en fonction de leur type sémantique

Nous avons vu précédemment, dans les cas d'usage American Airlines, Audi et Arcelor Mittal, que les corpus relatifs à ces cas d'usage sont pauvres en entités nommées. A ce moment là, nous utilisons les entités nommées pour filtrer les termes extraits durant l'analyse terminologique. Cela ne peut pas se faire lorsqu'un corpus est riche en entités nommées. Une autre approche a donc été mise au point pour ce type de corpus riche en entités nommées. Pour tester notre proposition, nous avons présenté le corpus du Golf, qui est un corpus riche en entités nommées, afin de tester l'intérêt de filtrer les entités nommées en fonction de leur type sémantique pour la construction d'ontologies. Dans le cas où nous avons un corpus factuel qui comprend un grand nombre d'entités nommées, nous proposons de procéder en filtrant en fonction du type sémantique. Nous ne conservons que les entités nommées dont le type a été *a priori* jugé pertinent pour l'application visée. Nous avons testé notre approche sur le cas d'usage du jeu de Golf. Nous avons retenu les entités nommées de type DATE, PERSONNE, AUTRES TYPES.

Les entités nommées de type sémantique DATE (par ex. 2007, 2008) décrivent les dates de création des règles du jeu. Pour la gestion des règles du jeu, il est important de garder l'historique de la création de ces règles pour d'éventuels interrogations sur la base de règles. Dans le corpus du Golf, les entités nommées ayant comme type sémantique PERSONNE correspondent à des noms de personnes qui participent dans l'organisation du jeu ou comme joueurs. Nous avons remarqué que les entités nommées de type sémantique AUTRES TYPES font généralement référence à des objets du domaine « jeu de

Golf ». Prenons par exemple, les entités nommées *Open Championship, Fédération Française, Stock Play* et *Match Play* qui décrivent soit des organismes jouant le rôle d'acteurs dans la définition des règles du jeu soit des types de jeu qui sont associés à des règles différentes.

A partir de certaines entités nommées, nous avons créé des termino-concepts comme par exemple les termino-concepts ***Strock play, Match play, Foursome, Threesome*** qui décrivent des catégories de jeu de Golf. Pour chacune de ces catégories, un ensemble de règles sont définies pour contrôler le déroulement du jeu. Nous avons aussi créé un termino-concept qui correspond au type sémantique DATE relatif aux unités terminologiques *2008, 2010*. Le termino-concept ***Date*** décrit les dates de création ou de modification des règles du jeu de Golf. Nous avons créé 17 termino-concepts au total à partir d'entités nommées. Cette expérimentation reste modeste par sa taille mais notre objectif était de tester le filtrage des unités terminologiques en fonction de leur type sémantique par rapport à l'application visée : construction du vocabulaire termino-conceptuel utilisé pour décrire des règles du jeu de Golf. Nous avons considéré les entités nommées de type DATE, PERSONNE, AUTRES TYPES pour la construction du vocabulaire termino-conceptuel de jeu de Golf (70 % des entités nommées extraites au total).

Ce cas d'usage reste un cas d'école pour tester les outils utilisés pour traiter un corpus écrit en français et valider la démarche consistant à amorcer la construction d'ontologies en considérant, d'abord, certains types sémantiques, pour les entités nommées relevant de ces types et enfin les termes qui sont dans leur voisinage.

Dans ces deux expérimentations relatives au domaine réglementaire, nous avons montré que l'identification des entités nommées et leur utilisation comme marqueurs durant la phase de conceptualisation guide l'ingénieur de la connaissance dans la détection et la conceptualisation d'éléments pertinents du domaine. Dans le cas d'usage de AA, la détection des entités nommées a permis d'identifier des connaissances importantes à modéliser dans l'ontologie de domaine comme par exemple les différents statuts que peut avoir un membre adhérant au programme de fidélité ou encore les zones géographiques qui définissent le nombre de miles attribués durant des voyages. Dans l'ontologie AA, les entités nommées sont conceptualisées soit en des concepts ou bien des instances de concepts. Leurs types sémantiques peuvent être considérés

comme étant des concepts dans le cas où il s'agit de connaissances pertinentes à modéliser comme par exemple les types sémantiques ORGANISATION et DATE correspondant à des concepts de l'ontologie et représentant respectivement l'ensemble des compagnies aériennes et des dates spécifiques qui sont utilisées pour l'écriture des règles métier.

Dans le cas d'usage de Audi, démarrer la conceptualisation en considérant les entités nommées a permis d'identifier des notions clefs autour de la réglementation des ceintures de sécurité des véhicules comme par exemple les noms des tests de sécurité appliquées aux équipements de véhicules ou les différentes catégories de véhicules. De plus, l'étude du voisinage de ces entités nommées a guidé l'ingénieur de la connaissance dans l'identification des termes qui correspondent au vocabulaire conceptuel utilisé pour décrire des règles métier. Prenons comme exemple les termes *temperature conditionning* et *light conditionning* qui ont été détectés au voisinage des entités nommées et qui décrivent l'ensemble des valeurs de température et luminosité utilisées comme des paramètres dans les méthodes de tests.

Un autre avantage de la détection des entités pour la construction d'ontologies est relatif à leur utilisation par un ingénieur de la connaissance. En effet, généralement le nombre des entités nommées extraites à partir d'un corpus d'acquisition constitue une liste qui peut traitée humainement. La liste des entités nommées produite par les outils est généralement plus réduite que celle des termes (105 entités nommées contre 973 termes pour le cas d'usage de AA et 90 contre 1003 termes pour le cas d'usage de Audi), il est plus facile de s'appuyer sur les entités nommées pour démarrer la conceptualisation et construire une première ébauche d'ontologie. On peut ensuite itérer en explorant les contextes des unités textuelles sélectionnées précédemment et en prenant en compte les nouveaux termes figurant dans leur voisinage. Notre contribution a fait l'objet d'articles scientifiques (Omrane *et al.*, 2011b,d). Notre démarche n' a pas été prise en compte dans notre méthodologie GRAPHONTO. Mais cette stratégie d'exploration des entités nommées est considérée comme une perspective de notre travail.

6.5.2 Evaluation de la méthode de pondération des unités terminologiques

Dans le chapitre 5, nous avons défini la mesure sémantique

$$W(ut) = \frac{Freq_{voisEN}(ut) + Freq_{voisPP}(ut) + D(ut)}{FreqTotale(ut)}$$

qui prend en compte plusieurs critères relatifs aux caractéristiques liées au corpus et au domaine pour la pondération des unités terminologiques dans le but de mettre l'accent sur celles qui sont potentiellement pertinentes à normaliser et ainsi guider l'ingénieur de la connaissance dans la détection des connaissances du domaine. Notre contribution vient du fait que l'exploration de longues listes de termes se fait souvent par ordre alphabétique (pour regrouper les termes apparentés) ou par fréquence (pour éviter de passer trop de temps sur les hapax[7] toujours très nombreux). Nous cherchons à repérer les termes décrivant le mieux le domaine sous-jacent parce qu'ils reflètent les concepts centraux[8]. Dans cette section, nous cherchons à évaluer l'intérêt de la pondération des termes que nous avons proposé comme critère de tri alternatif pour l'analyse d'une longue liste de termes. Nous évaluons notre mesure sémantique $w(UT)$ en comparant la liste des termes pondérée avec le vocabulaire terminologique qui a servi pour la création des réseaux terminologiques pour les cas d'usage AA et Audi.

Nous nous sommes particulièrement intéressés à l'évaluation du critère de voisinage des unités terminologiques figurant dans le même contexte que des entités nommées qui relèvent du domaine. En effet, une fois l'application visée choisie, nous pouvons nous appuyer sur des entités nommées qui relèvent du domaine à modéliser et sont mentionnées dans les textes pour repérer les termes qui sont eux-mêmes les plus pertinents pour ce domaine. Ces entités nommées peuvent ainsi aider à repérer les termes les plus pertinents pour un domaine donné. Ce critère de voisinage est défini par **PDomaine(**ut**)** comme suit :

$$PDomaine(ut) = \frac{Freq_{voisEN}(ut)}{FreqTotale(ut)}$$

7. Termes n'apparaissant qu'une seule fois dans un corpus.

8. Dans une ontologie de domaine, les concepts centraux correspondent aux concepts de domaine c'est à dire ceux qui décrivent des notions propres à un domaine précis.

où $Freq_{voisEN}(ut)$ est le nombre total de relations de voisinage dans lesquelles entrent les occurrences de *ut* et $FreqTotale(ut)$ sa fréquence totale (nombre d'occurrences). Le poids d'une unité terminologique est donc le nombre moyen de relations de voisinage dans lesquelles entrent ses occurrences. Prenons l'exemple suivant tiré d'un cas d'usage de AA où nous cherchons à calculer le poids de l'unité terminologique *mileage credit* :

> You may request mileage credit for past, eligible transactions up to 12 months from the transaction date. Any claim for uncredited mileage must be received by American Airlines within 12 months after the mileage credit was earned. No mileage credit will be awarded for canceled flights or if you are accommodated on another airline.

En appliquant le critère de vosinage, nous obtenons *PDomaine(mileage credit)* $= \frac{3}{3}$. L'unité terminologique *mileage credit* a 3 occurrences dans le texte. La première (non reproduite ici) n'a aucune entité nommée dans son voisinage. La seconde apparaît dans le texte avec une entité nommée (*12 months*) et la troisième avec deux (*American Airlines* et *12 months*).

6.5.2.1 Protocole expérimental

Nous avons testé notre mesure de pondération appliquée aux termes extraits par un outil de TAL sur les deux corpus AA et Audi. Pour chacun des cas d'usage AA et Audi, nous prenons en entrée deux listes d'unités terminologiques fournies par des outils de TAL : une liste LTC de termes candidats extraits par l'outil YaTeA et une liste LEN d'entités nommées extraites par la chaîne de traitement Annie de la plateforme Gate. Puis, il s'agit tout d'abord d'éliminer le bruit et de filtrer les termes les plus pertinents pour le domaine considéré. Nous calculons pour cela une liste de termes filtrés (LTF). Enfin, nous calculons une liste des termes pondérés (LTW) et nous comparons le résultat obtenu avec la liste de termes de référence Ref_T.

Nous avons extrait une liste de termes filtrés LTF pour pouvoir évaluer, d'abord, l'idée de la considération des termes qui sont au voisinage des entités nommées. Puis, nous appliquons notre mesure de pondération sur cette liste filtrée pour ordonner les termes suivant un ordre décroissant. Nous appliquons un filtrage qui s'appuie sur le fait qu'un terme t est pertinent si et seulement si

l'une de ses occurrences figure au voisinage d'une occurrence d'entité nommée :

$$Pert(TC) = \begin{cases} 1 & \text{si } \exists t, EN, e / occ(t, TC) \\ & \wedge occ(e, EN) \wedge vois(t, e) \\ 0 & \text{sinon} \end{cases}$$

où TC est un terme candidat, EN une entité nommée et où $occ(x, X)$ et $vois(x, y)$ indiquent respectivement que x est une occurrence de X et que x et et y cooccurrent dans la même phrase.

En appliquant cette méthode de filtrage sur les listes LTC et LEN, nous obtenons d'abord une liste de termes filtrés (LTF) qui contient les termes de la liste initiale figurant au moins une fois au voisinage d'une entité nommée. Cette liste de termes filtrés contient tous les termes de $Pert(TC) = 1$. Le tableau 6.23 présente les résultats obtenus.

Corpus	LTC	LEN	LTF
AA	680	105	437
Audi	1003	90	436

Table 6.23 – Nombres de termes candidats, d'entités nommées et de termes filtrés pour les deux cas d'usage.

D'après les résultats obtenus, on peut noter que l'écart entre le nombre de termes candidats filtrés est moindre dans le cas de AA que pour le corpus Audi. Cela tient au fait qu'un nettoyage préalable a été fait sur la liste des termes candidats de AA pour éliminer les termes mal formés [9] et mieux apprécier l'apport spécifique de notre filtrage. On observe que le filtrage permet de réduire la liste des candidats termes, même quand celle ci a été nettoyée (AA) et qu'il compense l'absence de nettoyage préalable (Audi).

Pour évaluer ces résultats, nous utilisons les mesures de précision, de rappel et de F-mesure (*cf* section 6.5.1.1). Nous comparons chacune des listes LTF (liste filtrée) et LTC (liste ordonnée par fréquence) avec la liste Ref_T (liste de référence). Les résultats figurent dans le tableau 6.24. Les résultats sont

9. Quelle que soit la qualité des extracteurs de termes, les outils généralistes ne peuvent pas prendre en compte les particuliarités des textes (présence de chiffres, de tableaux, etc.) sur lesquels ils sont appliqués et il reste toujours des scories.

Cas	Mesures	LTF	LTC
AA	Précision	71,6%	51%
	Rappel	89,1%	100%
	F-mesure	79,40%	67,55%
Audi	Précision	56,8%	31,4%
	Rappel	78,7%	100%
	F-mesure	65,98%	47,79%

TABLE 6.24 – Impact du filtrage sur les mesures de précision, rappel et F-mesure.

dans l'ensemble moins bons dans le cas d'Audi mais le filtrage des termes améliore significativement la précision dans les deux cas d'usage. Pour le cas de AA, la précision a augmenté de 51% à 71,6%. Le cas d'usage de Audi, la précision s'est améliorée de 31,4% à 56,8%. Mais, le rappel a baissé pour le cas d'usage de Audi de 100% à 78,7%. Pour le cas d'usage de AA, cette baisse est légère (de 100% à 89,1%). Le fait que le rappel soit moindre sur les listes de termes filtrées montre que le filtrage n'est pas parfait (certains termes pertinents ne figurent pas au voisinage d'entités nommées [10]), mais le gain de plus de 10 points de F-mesure sur la liste des termes filtrés montre que le filtrage est globalement positif. Il l'est même d'autant plus que la liste des termes candidats a été peu nettoyée au départ, comme dans le cas d'Audi. Le filtrage est une première manière grossière mais simple pour nettoyer une liste de termes candidats. Notre but est d'évaluer notre méthode de pondération. Dans la section suivante, nous décrivons l'évaluation de la pondération des unités terminologiques.

6.5.2.2 Evaluation de la pondération

Une fois nous avons évalué le filtrage d'une liste de termes candidats par rapport à une liste de termes, nous évaluons dans cette section la pondération des unités terminologiques en la comparant avec une liste de termes ordonnée par la fréquence, toujours par rapport à une liste de termes référence. Nous appliquons notre mesure sémantique $W(UT)$ définie dans le chapitre 5 mais

10. Les valeurs de 100% de rappel s'expliquent du fait que les références ont été construites à partir des termes candidats extraits sans ajout d'autres termes.

en considérant seulement le critère de voisinage des unités terminologiques avec des entités nommées $Freq_{voisEN}(ut)$ comme suit :

$$W(ut) = \frac{Freq_{vois}(ut)}{FreqTotale(ut)}$$

où $Freq_{vois}(ut)$ est le nombre total de relations de voisinage dans lesquelles entrent les occurrences de ut et $FreqTotale(ut)$ sa fréquence totale (nombre d'occurrences). Le poids d'une unité terminologique est donc le nombre moyen de relations de voisinage dans lesquelles entrent ses occurrences. Nous appliquons ce poids sur des unités terminologiques représentées par des termes. Nous obtenons une liste de termes candidats *pondérés* en fonction du nombre d'entités nommées figurant dans leur voisinage.

Nous évaluons les listes de termes pondérés par rapport aux mêmes références que précédemment. Il s'agit d'apprécier notre capacité à placer des termes pertinents en haut de classement. Nous regardons donc comment la précision évolue quand on considère un nombre croissant de termes bien classés dans les listes ordonnées pour les deux cas d'usage AA et Audi. Le tableau 6.25 montre les résultats obtenus pour les 100 termes les mieux classés, puis en fixant le seuil aux rangs 200 et 400 et enfin en considérant toute la liste des termes filtrés.

Cas	**r=100**	**r=200**	**r=400**	**LTF**
AA	78%	75%	71,5%	71,1%
Audi	45%	52,5%	60%	56,8%

Table 6.25 – Evaluation de la précision en fonction du nombre de termes retenus, le seuil étant fixé en fonction du rang ($r = 100, r = 200, r = 400$).

Les résultats obtenus sont contrastés. Dans le cas d'AA, la précision diminue régulièrement quand le rang augmente (de 78% à 71,5%). Cette constatation montre que plus on descend dans la liste pondérée plus les unités terminologiques qui y figurent ne font pas partie de la liste de référence. Ce qui prouve que la pondération des unités terminologiques met l'accent sur les unités qui sont pertinentes et que plus le poids est faible plus les unités considérées ne font pas partie du vocabulaire lié au domaine.

Mais, dans le cas d'Audi, elle augmente jusqu'au rang 400 avant de s'infléchir au-delà. Le tri proposé est donc globalement pertinent dans le premier cas et moins dans le second. Une analyse détaillée montre que les entités nommées numériques du corpus Audi contribuent à surpondérer des termes dénotant des propriétés de concepts (*length* dans *...maximum length of 840 mm* , par exemple.) par rapport aux termes dénotant les concepts eux-mêmes (*strap*) qui sont filtrés eux-aussi mais avec un poids inférieur (*the length of strap... shall be... as close as possible to 450 mm...*). Finalement, le voisinage des unités terminologiques fortement pondérées aide à la détection des unités qui seront normalisées. Notre mesure de pondération donne un autre point de vue d'une liste de termes candidats triée initialement par fréquence en mettent l'accent sur des termes qui vont représenter des futures propriétés de concepts ou qui aident à la détection de futurs concepts dans l'ontologie de domaine.

6.5.2.3 Bilan

Le voisinage des entités nommées apparaît comme un critère pertinent à prendre en compte pour la détection des termes d'un domaine. Il permet de réduire la part de bruit dans les résultats des outils de TAL et de débroussailler le travail de validation. Il faut seulement prévoir de pouvoir « récupérer » par d'autres méthodes des termes pertinents qui auraient été éliminés lors du filtrage. Comme critère de tri, le voisinage des entités nommées est à utiliser avec précaution mais il donne un éclairage complémentaire de la fréquence sur une liste de termes. Nous illustrons ce point sur quelques termes de la référence du cas d'usage Audi (voir tableau 6.26) qui n'ont pas le même rang dans les deux approches. On a ainsi deux critères de tri qui sont bruités mais pertinents et qui mettent en avant des termes différents. Cette expérimentation a fait l'objet d'un article scientifique (Omrane *et al.*, 2011c).

6.6 Conclusion

Dans ce chapitre, nous avons essayé de tester et valider deux de nos contributions dans cette thèse : la considération des entités nommées dans le processus de construction d'ontologies de domaine et la pondération des unités terminologiques. Nous avons utilisé des corpus d'acquisition qui sont de na-

Terme	Rang Poids	Rang Freq.
size class	1	36
Frontal impact test	3	36
diameter	4	33
buckle	53	18
seat	52	11
belt	45	1

TABLE 6.26 – Exemples de termes pertinents et leurs rangs.

ture différente mais qui sont relatifs au domaine réglementaire. Nous avons pu montrer que la détection des entités nommées permet l'identification de connaissances pertinentes pour le domaine. Nous avons vu que la pondération des unités terminologiques suivant des critères relatifs au corpus et à l'application visée permet de mettre l'accent sur des unités qui sont représentatives des notions d'un domaine spécifique.

Les possibilités d'évolution de notre méthode et les pistes pertinentes à explorer dans le domaine de l'ingénierie des connaissances sont présentées dans le chapitre suivant.

Chapitre 7

Conclusion et perspectives

Sommaire

Dans cette thèse, nous nous sommes intéressées au problème de l'acquisition des connaissances d'un domaine qui représente le cœur de la recherche en ingénierie des connaissances. Nous avons constaté dans l'état de l'art que plusieurs méthodologies et méthodes ont essayé de proposer des solutions à ce problème mais que peu de ces approches se sont intéressées spécifiquement à l'élicitation des connaissances à partir de textes. Afin de répondre à cette problématique, nous avons proposé une méthodologie de normalisation du réseau terminologique extrait à partir d'un corpus d'acquisition en un thésaurus qui décrit un vocabulaire normalisé et structuré du domaine. Notre méthode vient guider l'ingénieur de la connaissance dans le travail de normalisation qui est considéré comme une étape à la fois délicate et importante parce qu'elle assure le passage du niveau linguistique vers le niveau conceptuel. Dans ce chapitre, nous présentons les résultats de nos travaux et les perspectives à venir.

7.1 Conclusion générale

Nous avons présenté un état de l'art portant sur les indices textuels utilisés ainsi que les méthodologies et méthodes qui s'appuient sur ces indices pour construire des ressources terminologiques et ontologiques. L'analyse de l'état de l'art montre que beaucoup d'indices textuels sont exploités pour la construction d'ontologies à partir de textes mais qu'il est difficile de combiner ces indices entre eux.

Nous avons aussi conclu qu'il n'existe pas à ce jour de méthodologie de construction d'ontologies qui bénéficie d'un consensus général. Une méthodologie ou une méthode est choisie en fonction de l'objectif pour lequel l'ontologie doit être construite, les ressources utilisées et les choix techniques pour cette construction. Avec TERMINAE, nous nous sommes appuyées sur une approche terminologique qui nous a paru solide pour construire des ontologies de domaine à partir de textes. Cette approche repose en effet sur une analyse précise du rôle des termes et des relations sémantiques pour la construction d'ontologies. Nous avons cependant montré qu'on peut intégrer d'autres indices textuels dans le processus d'acquisition terminologique, notamment les entités nommées. Nous avons considéré les entités nommées dès le début du processus de construction d'ontologies, en les exploitant au même niveau que les termes et nous avons montré qu'elles servent pour la création des termino-concepts au niveau termino-conceptuel de la méthode TERMINAE. Evidemment la méthodologie de normalisation que nous proposons pourrait exploiter d'autres indices, comme les classes sémantiques par exemple ou exploiter mieux les résultats des outils de TAL, pour la détection des relations spécialisées. Ce sont des perspectives intéressantes pour la suite de nos travaux.

L'analyse de l'état de l'art montre aussi et surtout l'importance de la méthodologie en acquisition de connaissances. Cela s'explique dès lors que l'acquisition est vue comme un processus interactif et itératif. Devant la masse et la diversité des indices textuels extraits par des outils de TAL, il a fallu développer une réflexion méthodologique sur le travail d'acquisition de connaissances. La définition d'une méthode de construction d'ontologies a comme objectif de guider l'ingénieur de la connaissance durant le processus de construction d'ontologies. La plupart des méthodes de construction d'ontologies à partir de textes sollicitent l'intervention de l'ingénieur de la connaissance à plusieurs niveaux.

Pourtant, les méthodologies proposées dans l'état de l'art restent souvent à gros grain. Par exemple, les travaux du groupe TIA et TERMINAE mettaient surtout l'accent sur l'importance des éléments textuels comme données de base, sur le rôle d'un niveau de connaissance intermédiaire, termino-conceptuel, entre les niveaux terminologique et conceptuel et sur le fait que le processus global d'acquisition ne peut se faire sans l'intervention d'un expert.

Il fallait aller plus loin et définir une méthode plus précise pour assister

l'ingénieur de la connaissance dans son travail. En nous appuyant sur ces premiers résultats, nous avons commencé par formaliser les différentes structures de connaissances sur lesquelles repose le processus d'acquisition d'ontologies. Leur formalisation permet d'identifier la nature des connaissances manipulées dans chacun des niveaux de connaissance afin de définir les opérations appropriées qui permettent de les exploiter tout en respectant l'ensemble des contraintes définies pour ces structures de connaissances. De plus, nous avons défini des liens de correspondance entre ces structures de connaissances pour articuler entre eux les différents niveaux de connaissances qui entrent dans la construction d'ontologies de domaine. Ces liens de correspondance sont précieux parce qu'ils permettent de construire les niveaux de connaissances les uns à partir des autres et de relier les concepts définis au niveau conceptuel avec les mots qui y font référence dans les textes.

Nous avons analysé en détail la première étape du processus de construction d'ontologies au sein des approches terminologiques, celle qui permet de transformer un réseau terminologique en un réseau termino-conceptuel. Nous avons proposé une méthode détaillée pour le travail de normalisation terminologique qui s'appuie sur les différentes structures de connaissances manipulées aux niveaux terminologique et termino-conceptuel. Nous avons montré comment combiner différents indices textuels, à savoir les termes, les entités nommées et les relations terminologiques, pour la construction d'un réseau termino-conceptuel qui sert de vocabulaire normalisé pour la construction d'ontologies de domaine ou comme une ressource sémantique pour l'annotation de documents. Nous avons considéré que l'ingénieur de la connaissance joue un rôle central dans le travail de normalisation. Notre méthode fait intervenir l'ingénieur de la connaissance le plus longtemps possible tout en l'accompagnant dans le travail de normalisation.

7.2 Perspectives

Ces travaux ont abouti à la mise en place d'une méthode de normalisation, nommée GraphOnto, d'un réseau terminologique en un réseau termino-conceptuel mais de nombreuses perspectives restent ouvertes.

De nouvelles expérimentations pourraient être réalisées. Dans cette thèse, nous avons testé et évalué notre mesure sémantique de pondération des unités

terminologiques. En l'état actuel, nous ne prenons en considération que le critère de voisinage des termes et des entités nommées. Nous devons améliorer cette mesure de pondération des unités terminologiques en prenant en compte les autres critères de pertinence. Il peut s'agir par exemple du repérage des unités terminologiques dans des passages règlementaires et du nombre de relations terminologiques qui relient ces unités. Il faudrait tester ces mesures sur d'autres corpus réglementaires.

La plateforme de construction d'ontologies existe et nous avons défini une méthode de normalisation mais l'enchaînement du processus de normalisation n'est pas encore implémenté dans cette plateforme. Il faut prévoir d'intégrer des modules assurant le parcours, la normalisation du réseau terminologique, la mise à jour du réseau termino-conceptuel et le contrôle du processus de normalisation au sein de l'outil TERMINAE. Il faut aussi proposer des améliorations à l'ergonomie des fiches terminologiques pour rendre explicite les informations liées à l'analyse du réseau terminologique et à la création du réseau termino-conceptuel.

Une des contributions majeures de cette thèse est la formalisation des structures de connaissances manipulées dans chacun des niveaux terminologique, termino-conceptuel et conceptuel : cela permet d'expliciter des connaissances définies dans chaque niveau ainsi que les liens de correspondance qui permettent de créer des unités appartenant à niveau de connaissance à partir des éléments décrits dans le niveau précédent et de garder une traçabilité entre ces structures de connaissance. Dans le cas où les unités reliées par ces liens de correspondance évoluent, c'est-à-dire qu'elles sont enrichies par l'ingénieur, il faut se poser la question de leur interdépendance qui les relie. Autre lien de correspondance défini entre des unités termino-conceptuelles et des unités conceptuelles, nous pensons ajouter un lien de correspondance qui lie des termino-concepts à des « modules conceptuels ». Nous définissons un module conceptuel comme un patron conceptuel composé d'un ensemble de concepts reliés par des relations conceptuelles. A ce stade, nous proposons de définir un lien entre une ou plusieurs unités termino-conceptuelles à une ou plusieurs unités conceptuelles. Dans le cas où l'unité conceptuelle évolue (par ex. du fait de l'enrichissement de l'ontologie) en un ensemble de concepts reliés par des relations conceptuelles (par ex. spécialisation d'un concept en plusieurs concepts fils reliés par des relations conceptuelles à d'autres concepts d'une

ontologie), nous envisageons de définir un lien de correspondance entre des structures de connaissances différentes.

Une autre perspective vise à améliorer les indicateurs de contrôle de normalisation. Dans cette thèse, nous avons défini des indicateurs qui permettent de vérifier la progression du travail de normalisation et de s'assurer qu'il ne reste plus d'unités terminologiques pertinentes pour le domaine et non encore normalisées. Ces indicateurs permettent aussi de s'assurer de la couverture du vocabulaire normalisé par rapport à celui utilisé dans des passages réglementaires afin de vérifier que la ressource construite répond aux besoins de l'application visée. Il nous semble intéressant de définir un nouvel indicateur qui vérifie s'il existe au sein du réseau termino-conceptuel des « anomalies» liées par exemple à la non conformité de la description formelle des éléments du réseau termino-conceptuel avec la norme du vocabulaire contrôlé (par ex. SKOS) utilisé pour leur création ou relatives à la structure du réseau termino-conceptuel. En effet, durant le travail de normalisation l'ingénieur de la connaissance a une vision partielle du réseau termino-conceptuel en cours de construction. Il peut donc créer des termino-concepts partageant le même label préféré. Il peut aussi créer plusieurs termino-concepts non reliés à d'autres termino-concepts par des relations de type *Généricité/Spécificité*.

Les développements et perspectives évoqués ici ne doivent pas occulter le fait qu'une méthode ne vaut que parce qu'elle est utilisée. Tout ce travail de thèse a été guidé et nourri par les cas d'usage de ONTORULE. Les nouvelles expériences faites avec TERMINAE dans le cadre d'autres projets vont naturellement permettre d'éprouver et d'enrichir la méthode que nous avons proposée.

Bibliographie

(2011). *Mesures de similarité distributionnelle entre termes*. 22

Alessandro L., Simonetta M., Vito P. & Giulia V. (2007). NLP-based ontology learning from legal texts. a case study. In P. Casanovas, M. A. Biasiotti, E. Francesconi & M.-T. Sagri, Eds., *LOAIT*, volume 321, p. 113–129. 5, 16, 44, 51

Amardeilh F., Laublet P. & Minel J. L. (2005). Document annotation and ontology population from linguistic extractions. In *Proceedings of the 3rd international conference on Knowledge capture*, p. 161–168 : ACM. 7, 19, 20

Aubin S. & Hamon T. (2006). Improving term extraction with terminological resources. In T. Salakoski, F. Ginter, S. Pyysalo & T. Pahikkala, Eds., *Advances in Natural Language Processing*, p. 380–387. Springer Berlin/ Heidelberg. 107

Aussenac-Gilles N., Després S. & Szulman S. (2008). The TERMINAE method and platform for ontology engineering from texts. p. 199–223. 3, 5, 6, 16, 44, 50, 51, 57, 64, 65

Aussenac-Gilles N. & Jacques M.-P. (2008). Designing and evaluating patterns for relation acquisition from texts with caméléon. *Terminology, Pattern-Based approaches to Semantic Relations*, p. 45–73. 25

Aussenac-Gilles N. & Séguéla P. (2000). Les relations sémantiques : du linguistique au formel. *Cahiers de Grammaire. Sémantique et Corpus. Textes réunis par A. Condamines*, **25**. 20, 33, 45

Bachimont B. (2000). Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. *Ingénierie des Connaissances : Evolutions récentes et nouveaux défis*, p. 1–16. 52

Bachimont B. (2004). *Arts et sciences du numérique : Ingénierie des connaissances et critique de la raison computationnelle*. Mémoire d'habilitation à diriger des recherches, Université de Technologie de Compiégne. 40

Basili R., Pazienza M. T. & Velardi P. (1996). An empirical symbolic approach to natural language processing. *Artificial Intelligence*, p. 59–99. 26

Bendaoud R., Napoli A. & Toussaint Y. (2008). Formal concept analysis : A unified framework for building and refining ontologies. In A. Gangemi & J. Euzenat, Eds., *Knowledge Engineering : Practice and Patterns*, volume 5268, p. 156–171. 47, 52, 53

Bendaoud R., Rouane Hacene M., Toussaint Y., Delecroix B. & Napoli A. (2007). Construction d'une ontologie à partir d'un corpus de textes avec l'ACF. In F. Trichet, Ed., *Actes des 18èmes Journées francophones d'Ingénierie des Connaissances (IC2007)*, p. 121–133. 26

Bisson G. & Nedellec C. (2001). Aide à la conception de méthodes de classification pour la construction d'ontologies : l'atelier mo'k. In *Extraction et la Gestion des Connaissances*, p. 213–225. 26

Bisson G., Nedellec C. & Cañamero D. (2000). Designing clustering methods for ontology building - the mo'k workbench. In *In Proceedings of the ECAI Ontology Learning Workshop*, p. 13–19. 5

Blaz F., Marko G. & Dunja M. (2006). Semi-automatic data-driven ontology construction system. In *Proceedings of the 9th international multi-conference information society*. 18, 52

Bob G. S., Wielinga B. & Jansweijer W. (1995). The KACTUS view on the 'o' word. In *In IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing*, p. 159–168. 43, 54

Bontcheva K. & Cunningham H. (2003). The semantic web : A new opportunity and challenge for human language technology. In *Proceedings of Workshop on Human Language Technology for the Semantic Web and Web Services, 2nd International Semantic Web Conference*. 18

Bourigualt D. & Aussenac-Gilles N. (2003). Construction d'ontologies à partir de textes. In *Actes de la Conférence Traitement Automatique du Langage Naturel (TALN'2003)*, p. 27–47, France. 198

P. Buitelaar, P. Cimiano & B. Magnini, Eds. (2005). *Ontology Design and Population*. IOS Press. 18

Cea G. A. D., Gómez-Pérez A., Montiel-Ponsoda E. & Suárez-Figueroa M. C. (2008). Natural language-based approach for helping in the reuse of ontology design patterns. p. 32–47. xv, 56, 57

Charlet J., Aussenac-Gilles N., Pierra G., Nada N., Szulman S. & Teguiak H. (2008). DAFOE : Une plateforme multi-méthodes et multi-modèles pour le développement d'ontologies de domaine. In D. Benslimane, C. Roche & S. Spaccapietra, Eds., *Journées Francophones sur les Ontologies (JFO)*, p. 1–12. 3, 16, 40, 83

Charlet J., Zacklad M., Kassel G. & Bourigault D. (2000). Ingénierie des connaissances : recherches et perspectives. In *Ingénierie des connaissances : évolutions récentes et nouveaux défis*. 2

Church K. W. & Hanks P. (1989). Word association norms, mutual information, and lexicography. In *27th Meeting of the Association for Computational Linguistics*, p. 76–83. 26

Cimiano P. (2006). *Ontology Learning and Population from Text : Algorithms, Evaluation and Applications*. 112

Cimiano P. (2007). *Ontology Learning and Population from text*. Algorithms, Evaluation and Application. ISO Press. 107

Cimiano P., Hotho A. & Staab S. (2004). Clustering ontologies from text. In *Proceedings of the 4th international conference on language resources and evaluation (LREC)*, p. 1721–1724. 47

Cimiano P., Hotho A. & Staab S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. *J. Artif. Intell. Res. (JAIR)*, **24**, 305–339. 5

Cimiano P., Hotho A. & Staab S. (2011). Learning concept hierarchies from text corpora using formal concept analysis. *Computing Research Repository (CoRR2011)*. 26, 47, 53

Cimiano P. & Völker J. (2005). Text2onto - a framework for ontology learning and data-driven change discovery. In A. Montoyo, R. Munoz & E. Metais, Eds., *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)*, p. 227–238. 2, 5, 16, 44

Crampes M., Ranwez S. & Villerd J. (2008). Cartographie sémantique auto-organisée d'un référentiel de connaissances partagé. In *19èmes journées francophones d'Ingénierie des Connaissances (IC2008)*, p. 161–172. 69

Dahab M. Y., Hassan H. A. & Rafea A. (2008). TextOntoEx : Automatic ontology construction from natural english text. *Expert Syst. Appl.*, p. 1474–1480. 46, 50, 51

Daille B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. PhD thesis, Université Paris 7. 29

Daille B. (2003). Terminology mining. In *Information Extraction in the Web Era*, p. 29–44. Springer Berlin Heidelberg. 29

Didier B. & Frérot C. (2005). Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique. *Traitement Automatique des Langues*, **47**, 141–154. 33

Drouin P. (2003). Term extraction using non-technical corpora as a point of leverage. In *Terminology*, p. 99–117. 28, 29

Drouin P. & Langlais P. (2006). Evaluation du potentiel terminologique de candidats termes. In *Actes des 8e Journées internationales d'analyse statistique des données textuelles (JADT-2006)*, p. 379–388. 29

Drymonas E., Zervanou K. & Petrakis E. G. M. (2010). Unsupervised ontology acquisition from plain texts : the ontogain system. In *Proceedings*

of the Natural language processing and information systems, and 15th international conference on Applications of natural language to information systems, p. 277–287. 2, 16, 45, 46, 48, 49, 53

Ehrmann M. (2008). *Les Entités Nommées, de la linguistique au TAL : statut théorique et méhodes désambiguïsation*. PhD thesis, Université de Paris VII. 17, 18, 30, 113

Faure D. & Nédellec C. (1999). Knowledge acquisition of predicate argument structures from technical texts using machine learning : the system asium. In *Proceedings of the 11th International Conference on Knowledge Engineering and Knowledge Management*, p. 329–334. 21, 33, 47

C. Fellbaum, Ed. (1998). *WordNet An Electronic Lexical Database*. The MIT Press. 69

Fernandez-Lopez M., Gomez-Perez A. & Juristo N. (1997). METHONTOLOGY : from ontological art towards ontological engineering. In *Proceedings of the AAAI97 Spring Symposium*, p. 33–40, Stanford, USA. 2, 43, 54

Fort K., Ehrmann M. & Nazarenko A. (2009). Towards a methodology for named entities annotation. In *Proceedings of the Third Linguistic Annotation Workshop*, p. 142–145 : Association for Computational Linguistics. 18

Frantzi K. & Ananiadou S. (1997). Automatic term recognition using contextual cues. In *Proceedings of 3rd DELOS Workshop*. 28

Fuchs B., Huchard M. & Napoli A. (2010). Une étude sur la mise en forme de patrons de conception pour les ontologies avec l'analyse formelle de concepts. In J.-C. R. Eric Cariou, Ed., *Langages et Modèles à Objets (LMO)*, p. 83–98. 21, 56

Gangemi A. & Presutti V. (2009). *Handbook on Ontologies*. Steffen Staab and Ruder Studer, Springer. 56

Gómez-Pérez A., Fernández-López M. & Corcho O. (2004). *Ontological Engineering*. 43, 54

Gómez-Pérez A., Fernández-López M. & Corcho O. (2007). *Ontological Engineering : with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. (Advanced Information and Knowledge Processing).* 54

Gruber T. R. (1993). *Toward principles for the design of ontologies used for knowledge sharing.In In Formal Ontology in Conceptual Analysis and Knowledge Representation.* Available as stanford knowledge systems laboratory report, Kluwer Academic. 1, 39, 40

Guarino N. (1998). Formal ontology and information systems. p. 3–15. 41, 42

Hamon T. (2000). *Variation sémantique en corpus spécialisés : acquisition de relations de synonymie à partir de ressources lexicales.* PhD thesis, Université Paris-Nord. 108, 118

Harris Z. (1954). Distributional structure. *Word.* 25, 46

Hearst M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, p. 539–545. 20, 24

Ibekwe-SanJuan F. (2005). Inclusion lexicale et proximité sémantique entre termes. *Terminologie et Intelligence Artificielle (TIA 2005.* 16

Kassel G. (2002). OntoSpec : une méthode de spécification semi-informelle d'ontologies. In *13èmes journées francophones d'inégnierie des connaissances IC.* 2, 55

Kassel G. (2009). Vers une ontologie formelle des artefacts. In *IC 2009 : Actes des 20es Journées Francophones d'Ingénierie des Connaissances*, p. 121–132. 55

Kayser D. (1997). *La représentation des connaissances.* Collection informatique. Hermés. 69

Lerat P. (2009). La combinatoire des termes. exemple : nectar de fruits. In *Hermes. Journal of Language and Communication Studies*, p. 211–232. 15

Lopes L., Vieira R., Finatto M. J. B. & Martins D. (2010). Extracting compound terms from domain corpora. *J. Braz. Comp. Soc.*, **16**, 247–259. 51

Maedche A. & Staab S. (2001). Ontology learning for the semantic web. *IEEE Intelligent Systems*, **16**, 72–79. 2, 38, 46

Magnini B., Pianta E., Popescu O. & Speranza M. (2006). Ontology population from textual mentions : Task definition and benchmark. In *Proceedings of the OLP2 workshop on Ontology Population and Learning*, Sidney, Australia. 18

Malaisé V. (2005). *Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles à partir de corpus textuels*. These, Université Paris-Diderot - Paris VII. 45

Manzano-Macho D., Gómez-Pérez A. & Borrajo D. (2008). Unsupervised and domain independent ontology learning : Combining heterogeneous sources of evidence. In *LREC* : European Language Resources Association. 18, 49, 51, 117, 119

Maynard D. & Ananiadou S. (1999). Term extraction using a similarity-based approach. In D. Bourigault, C. Jacquemin & M.-C. Lhomme, Eds., *Recent Advances in Computational Terminology*, p. 261–278. John Benjamins. 28, 112

Mondary T., Nazarenko A., Zargayouna H. & Barreaux S. (2012). The Quaero evaluation initiative on term extraction. In *in Proceedings of the 8th International Conference on Language Resources and Evaluation - LREC2012*. 29, 107

Morin E. (1999). Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique. *TAL. Traitement automatique des langues*, p. 143–166. 20

Morita T., Fukuta N., Izumi N. & Yamaguchi T. (2008). Doddle-owl : Interactive domain ontology development with open source software in java. In *IEICE Transactions on Information and Systems, Special Section on Knowledge-Based Software Engineering*, p. 945–958. 16

Nadeau D. & Sekine S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigaciones*, **30**. 113

Navigli R., Velardi P., Cucchiarelli A. & Neri F. (2004). Quantitative and qualitative evaluation of the ontolearn ontology learning system. In *COLING*. 45, 52

Navigli R., Velardi P. & Faralli S. (2011). A graph-based algorithm for inducing lexical taxonomies from scratch. In *IJCAI*, p. 1872–1877. 16

Nazarenko A. (2004). Primitives logiques et réseaux sémantiques : quel bilan pour la terminologie ? *Revue d'Intelligence Artificielle*. 69

Nazarenko A., Guissé A., Lévy F., Omrane N. & Szulman S. (2011). Integrating Written Policies in Business Rule Management Systems. In *Rule-Based reasoning, Programming, and Applications*, volume 6826 of *Lecture Notes in Computer Science*, p. 99–113, Barcelona, Espagne : Springerlink. 181

Nédellec C., Nazarenko A. & Bossy R. (2009). Information extraction. In S. Staab & R. Studer, Eds., *Handbook on Ontologies in Information Systems*, chapter 31. Springer Verlag. 18

Nobata C., Collier N. & Tsujii J. (2000). Comparison between tagged corpora for the named entity task. In *Proceedings of the Association for Computational Linguistics (ACL 2000) Workshop on Comparing Corpora*, p. 20–26 : Kilgarri, A. and Berber, T. 19

Novácek V. (2012). Distributional framework for emergent knowledge acquisition and its application to automated document annotation. *Computing Research Repository (CoRR2012)*. 7, 45, 52

Noy N. F., Fergerson R. W. & Musen M. M. (2000). The knowledge model of protégé-2000 : Combining interoperability and flexibility. In *Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management : Methods, Models, and Tools (EKAW 2000)*, volume 1937 of Lecture Notes in Artificial Intelligence (LNAI), p. 17–32, Juan-les-Pins, France : Springer. 43

Noy N. F. & Mcguinness D. L. (2001). *Ontology Development 101 : A Guide to Creating Your First Ontology*. Rapport interne. 49

Omrane N., Nazarenko A., Rosina P., Szulman S. & Westphal C. (2011a). Lexicalized ontology for a business rules management platform : An automotive use case. In Springer, Ed., *Rule-Based Modeling and Computing in the Semantic Web*, LNCS 7018, p. 179–192, Florida, États-Unis. 206

Omrane N., Nazarenko A. & Szulman S. (2010). Combining terms and named entities for modeling domain ontologies from texts. In *The 17 th International conference on Knowledge Engineering and Knowledge Management*, p. 3–5, Lisbon, Portugal. 182

Omrane N., Nazarenko A. & Szulman S. (2011b). From Linguistics to Ontologies The Role of Named Entities in the Conceptualisation Process. In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development*, p. 249–254, Paris, France. 210

Omrane N., Nazarenko A. & Szulman S. (2011c). Le poids des entités nommées dans le filtrage des termes d'un domaine. In *9th International Conference on Terminology and Artificial Intelligence*, p. 80–86, Paris, France. 216

Omrane N., Nazarenko A. & Szulman S. (2011d). Les entités nommées : des clés linguistiques pour la conceptualisation. In *22èmes journées francophones d'ingénierie des connaissances (IC2011)*, p. 435–450, Chambéry, France. 210

Poibeau T. (2005). Sur le statut référentiel des entités nommées. In *Proceedings of TALN'05*, p. 173–182. 17, 30

Presutti A. G. V. (2008). Content ontology design patterns as practical building blocks for web ontologies. In *Conceptual Modeling - ER 2008*, volume 5231 : Lecture Notes in Computer Science. 56, 198

Quillian M. (1968). *Semantic Memory*. In M. Minsky (ed.), Semantic Information Processing, MIT Press. 69

Rastier F. (2005). *Enjeux épistémologiques de la linguistique de corpus*, p. 31–45. 12, 14, 23

Rebeyrolle J. & Tanguy L. (2000). *Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires*, chapter Cahiers de Grammaire, p. 153–174. 20, 24

Schmid H. (1995). Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*, p. 47–50. 107

Seppala S. (2012). *Contraintes sur la sélection des informations dans les définitions terminographiques : vers des modèles relationnels génériques pertinents*. PhD thesis, Université de Genève. 24

Smith B. (2001). Fois introduction : Ontology-towards a new synthesis. In *Proceedings of the international conference on Formal Ontology in Information Systems - Volume 2001*, New York, NY, USA. 40

Smith B. (2003). *Blackwell Guid to Philosophiy of Computing and Information*, volume chapter Ontology. Oxford : Blackwell. 39

Smrz P. & Nováček V. (2006). Ontology acquisition for automatic building of scientific portals. In *SOFSEM*, p. 493–500 : Springer-Verlag. 45, 52

Srikant R. & Agrawal R. (1995). Mining generalized association rules. In *Very Large Data Bases VLDB 1995*, p. 407–419 : Morgan Kaufmann Publishers Inc. 46

Studer R., Benjamins R. & Fensel D. (1998). Knowledge engineering : Principles and methods, data and knowledge engineering. *Data and Knowledge Engineering*, **25**, 161–197. 39

Sure Y., Erdmann M., Angele J., Staab S., Studer R. & Wenke D. (2002). Ontoedit : Collaborative ontology engineering for the semantic web. **2342**, 221–235. 43, 54

Szulman S., Biébow B. & Aussenac-Gilles N. (2002). Structuration de terminologie à l'aide d'outils de tal avec terminae. *Traitement Automatique des Langues*, p. 103–128. 3

Tanev H. & Magnini B. (2008). Weakly supervised approaches for ontology population. In *Proceeding of the 2008 conference on Ontology Learning and Population : Bridging the Gap between Text and Knowledge*, p. 129–143, Amsterdam, The Netherlands, The Netherlands : IOS Press. 18, 112

Tjong Kim Sang E. F. & De Meulder F. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, p. 142–147 : Walter Daelemans and Miles Osborne. 17

Uschold M. & King M. (1995). Towards a methodology for building ontologies. In *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing at the International Joint Conference on Artificial Intelligence*. 40, 54

Velardi P., Navigli R., Cucchiarelli A. & Neri F. (2006). Evaluation of OntoLearn, a methodology for automatic population of domain ontologies. In P. Buitelaar, P. Cimiano & B. Magnini, Eds., *Ontology Learning from Text : Methods, Applications and Evaluation*. 2, 5, 43, 112

Wilson W., Wei L. & Mohammed B. (2007). Determining termhood for learning domain ontologies using domain prevalence and tendency. In *Proceedings of the sixth Australasian conference on Data mining and analytics - Volume 70*, AusDM'07, p. 47–54, Darlinghurst, Australia, Australia : Australian Computer Society, Inc. 28

Xu F., Kurz D., Piskorski J. & Schmeier S. (2002). A domain adaptive approach to automatic acquisition of domain relevant terms and their relations with bootstrapping. In *Proceedings of the 3rd International Conference on Language Resources an Evaluation (LREC'02), May 29-31* : Las Palmas, Canary Islands, Spain. 16, 20