Contribution à la modélisation d'une conscience culturelle artificielle émique par les ontologies

Jean Petit

École Doctorale Sciences Technologie Santé (547)

Université de Reims Champagne-Ardenne

Thèse

pour obtenir le grade de

Docteur de l'Université de Reims Champagne-Ardenne

Discipline : Informatique

présentée et soutenue publiquement par

Jean Petit

le 27 juin 2017

Thèse dirigée par M. Francis Rousseaux

Contribution à la Modélisation d'une Conscience Culturelle Artificielle Émique par les Ontologies

---

Jury

M. Arnaud Martin,	Professeur,	IUT de Lannion,	Président
M. Francis Rousseaux,	Professeur,	Université de Reims Champagne-Ardenne,	Directeur de thèse
M. Gilles Kassel,	Professeur,	Université de Picardie,	Rapporteur
M. Eddie Soulier,	Professeur,	Université Technologique de Troyes,	Examinateur
M. Hacène Fouchal,	Professeur,	Université de Reims Champagne-Ardenne,	Examinateur
Mme Colette Faucher,	Professeur,	Université d'Aix-Marseille,	Examinatrice

École Doctorale Sciences Technologie Santé (547)

Université de Reims Champagne-Ardenne

Thèse

pour obtenir le grade de

Docteur de l'Université de Reims Champagne-Ardenne

Discipline : Informatique

présentée et soutenue publiquement par

Jean Petit

le 27 juin 2017

Thèse dirigée par M. Francis Rousseaux

Contribution à la Modélisation d'une Conscience Culturelle Artificielle Émique par les Ontologies

---

Jury

M. Arnaud Martin,	Professeur,	IUT de Lannion,	Président
M. Francis Rousseaux,	Professeur,	Université de Reims Champagne-Ardenne,	Directeur de thèse
M. Gilles Kassel,	Professeur,	Université de Picardie,	Rapporteur
M. Eddie Soulier,	Professeur,	Université Technologique de Troyes,	Examinateur
M. Hacène Fouchal,	Professeur,	Université de Reims Champagne-Ardenne,	Examinateur
Mme Colette Faucher,	Professeur,	Université d'Aix-Marseille,	Examinatrice

Remerciements

Pour commencer, j'exprime ici toute ma reconnaissance envers mon directeur de thèse, Francis Rousseaux, pour l'opportunité de cette thèse CIFRE et son travail de supervision.

Je veux exprimer ma gratitude à l'entreprise d'accueil Capgemini Technology Services pour m'avoir permis d'effectuer ma thèse dans de bonnes conditions. Mes pensées vont en particulier vers Catherine Lhermet qui a su me soutenir et me motiver, ainsi que vers Jean Brunet qui a participé à mon épanouissement professionnel.

Merci aux rapporteurs Gilles Kassel et Arnaud Martin pour le temps qu'ils ont consacré à la relecture de ma thèse, leurs commentaires et leurs conseils. Je souhaite aussi remercier les autres membres du jury, Eddie Soulier, Hacène Fouchal et Colette Faucher, ainsi que la représentante de Capgemini, Suzanne Angeli.

Mes remerciements vont aussi à mon collègue thésard Jean-Charles Risch, avec qui j'ai partagé la majeure partie et les meilleurs moments de ma thèse, notamment la conférence IJCAI en Argentine.

Je souhaite remercier les chercheurs et chercheuses qui se sont investis dans ma thèse. À Jean-Charles Boisson, je veux sincèrement le remercier pour s'être intéressé à ma thèse lors de cette « Journée des doctorants » et s'être investi avec sérieux et rigueur dans mon suivi. Nos diverses réunions via Skype ont énormément contribué à l'avancement et à la qualité de mes travaux. À Eunika Laurent-Mercier, je veux lui exprimer toute mon affection. Ses compétences scientifiques ont enrichi mes recherches et son enthousiasme a grandement participé à ma confiance en moi. À Gulgun Kayakutlu, je veux lui dire merci pour nos échanges transculturels.

Finalement, je souhaite remercier mes proches qui m'ont soutenu à leur manière durant ces trois années de thèse. Tout d'abord ma copine, Sophie Dziengelewski, mon ami de longue date, Yann Vanmansart, et ma famille : Thibault, Nicolas, Samuel, Laurent et Brigitte Petit.

Résumé en français

Depuis l'expansion du web, de nombreuses applications cherchent à répondre aux besoins d'utilisateurs ou de machines aux origines culturelles variées. De ce contexte de diversité culturelle émergent de nombreux conflits liés à des conceptions du monde différentes. Proposer des services adaptés requiert l'intégration au sein du système d'une forme de conscience culturelle. Une conscience culturelle artificielle est composée de représentations et de médiations culturelles formelles offrant au système les moyens pour interpréter les cultures représentées et déterminer leurs différences. Jusqu'à présent les représentations utilisées dans le développement des systèmes culturellement conscients sont issues de modèles universels ou « étiques ». Ces modèles grossiers, bien qu'ils soient adaptés, limitent la compréhension possible des cultures représentées. Par conséquent ils constituent un goulot d'étranglement dans le développement des systèmes culturellement conscients.

Cette thèse explore le développement d'une conscience culturelle artificielle plus fine sur la base de modèles culturels dits « émiques », c'est-à-dire spécifiques à chaque culture. J'étudie la construction, la formalisation et la médiation des représentations culturelles émiques. Mes contributions principales sont la conception et la validation, d'une part, d'un nouveau processus ethnographique semi-automatique de construction de modèles émiques via la fouille de textes et, d'autre part, d'une conscience culturelle artificielle émique fondée sur l'alignement d'ontologies culturelles issues de ces modèles.

Mots clefs : système culturellement conscient, conscience culturelle artificielle émique, modèle culturel prototypique, relation sémantique, ontologie culturelle, médiation culturelle.

English abstract

With the growing web, a number of applications seek to meet the needs of users or machines having diverse cultural backgrounds. From this context of cultural diversity arises conflicts linked to different world conceptions. Offering adaptated services requires the integration of a form of cultural awareness in the system. An artificial cultural awareness is composed of formal cultural representations and mediations providing the system with the means to interpret the represented cultures and to determine their differences. So far the representations used for the development of culturally-aware systems come from universal or "etic" models. Those coarse-grained models, even though they are adapted, limit the possible understanding of the represented cultures. As a consequence they constitute a bottleneck for the development of culturally-aware systems.

This thesis investigates the development of a finer-grained artificial cultural awareness based on cultural models called "emic", in other words specific to each culture. I study the construction, the formalisation and the mediation of these emic cultural representations. My main contributions are the design and validation of, on the one hand, a new semi-automatic ethnographic process for building emic models through text-mining, on the other hand, an emic artificial cultural awareness based on the mapping of cultural ontologies coming from those models.

Keywords : culturally-aware system, emic artificial cultural awareness, prototypical cultural model, semantic relation, cultural ontology, cultural mediation.

Sommaire

Liste des figures

Liste des tableaux

Acronymes

CAS	Culturally-Aware System – Système Culturellement Conscient
TEL	Technology Enhanced Learning – Apprentissage Assisté par la Technologie
HCI	Human-Computer Interaction – Interaction Homme-Machine
CATS	Culturally-Aware Tutoring System – Système de Tutorat Culturellement Conscient
HRAF	Human Relations Area Files – Fichiers Géographiques des Relations Humaines
OCM	Outline of Cultural Materials – Schéma des Matériaux Culturels
OWC	Outline of World Cultures – Schéma des Cultures Mondiales
ACA	Artificial Cultural Awareness – Conscience Culturelle Artificielle
UOC	Upper Ontology of Culture – Ontologie Supérieure de la Culture
culTEL	culturally-aware Technology Enhanced Learning – Apprentissage Assisté par la Technologie et culturellement conscient
PCM	Prototypical Cultural Model – Modèle Culturel Prototypique
CCT	Cultural Consensus Theory – Théorie du Consensus Culturel
KDP	Knowledge Discovery Process – Processus de Découverte de Connaissances
ML	Machine Learning – Apprentissage Automatique
TF	Term Frequency – Fréquence des Termes
IDF	Inverse Document Frequency – Fréquence de Documents Inversée
PMI	Pointwise Mutual Information – Information Mutuelle par Points

NGD	Normalized Google Distance – Distance de Google Normalisée
HAL	Hyperspace Analogue to Language – Hyperespace Analogue à la Langue
LSA	Latent Semantic Analysis – Analyse Sémantique Latente
VSM	Vector Space Model – Machine à Vecteurs de Supports
BLESS	Baroni and Lenci Evaluation of Semantic Spaces – Évaluation des Espaces Sémantiques par Baroni and Lenci
XML	eXtensible Markup Language – Langage de Balisage Extensible
GEXF	Graph Exchange XML Format – Format XML d'Échange de Graphe
URI	Uniform Resource Identifier – Identifiant Uniforme de Ressource
RDF	Resource Description Framework – Cadre de Description des Ressources
RDFs	RDF scheme – schéma RDF
OWL	Web Ontology Language – Langage Ontologique Web
SPARQL	SPARQL Protocol and RDF Query Language – Protocole SPARQL et Langage de Requête RDF
SQL	Structured Query Language – Langage de Requête Structurée
BDD	Base De Données

Chapitre 1

Introduction

1.1 Introduction

Le contexte

Depuis les années 2000 avec l'expansion rapide du web, la culture se digitalise et les systèmes informatiques sont confrontés à sa diversité. Que ce soient des services de recherche d'information tels que Google, des sites e-commerce comme E-bay ou d'autres applications mondialement utilisées, tous sont concernés par ce phénomène. Néanmoins, ces systèmes experts empreints d'une culture sont amenés à communiquer pour satisfaire des utilisateurs ou machines qui en possédent une toute autre.

La problématique

Les systèmes informatiques sont naturellement culturellement incompétents puisqu'ils ne sont pas conçus pour interagir avec d'autres cultures. L'inaptitude à prendre en considération les situations interculturelles est vecteur d'un certain nombre de risques allant de l'incompréhension au malentendu, jusqu'à l'insulte. Le développement des systèmes informatiques doit donc se faire selon la perspective d'un relativisme culturel.

Le développement de systèmes culturellement conscients est nécessaire dans un monde multiculturel que ce soit pour des raisons éthiques ou économiques. En effet, on peut considérer qu'il est normal que les services informatiques publiques d'un pays soient adaptés aux principales cultures qui le composent. D'un point de vue économique, ces systèmes interculturellement compétents vont pouvoir atteindre un nombre plus

important d'utilisateurs et donc repousser les frontières de leur marché. Ainsi, le développement de systèmes culturellement conscients est d'intérêt à la fois pour les producteurs et pour les consommateurs de services informatiques.

La littérature

Pour l'instant, le développement d'une conscience culturelle artificielle est basé sur des modèles formels universels ou « étiques ». Cette approche (1) offre une base interculturelle intrinsèque pour réaliser des médiations et permet (2) de décliner aisément des représentations culturelles par instanciation (notamment la quantification). De nombreux travaux ont validé l'utilisation de ces représentations dans le développement de systèmes culturellement conscients. Cependant, bien que cette approche soit adaptée, les modèles universels ne permettent pas une représentation détaillée des cultures, ce qui limite considérablement la compréhension qu'ils peuvent en avoir (Mohammed et Mohan, 2013a). La conscience culturelle artificielle formée à partir de ces modèles est pauvre et son utilité pour les systèmes informatiques est restreinte.

L'objectif

L'objectif principal de cette thèse est de contribuer à la modélisation d'une conscience culturelle artificielle élaborée. C'est pourquoi je propose une nouvelle approche, ancrée sur des représentations spécifiques ou « émiques » propres à chaque culture, qui invite à :

— revoir leur processus ethnographique d'acquisition dans une perspective automatique ;

— repenser les médiations dans un contexte de conceptualisations culturellement hétérogènes.

La méthodologie

Le développement d'une conscience culturelle artificielle émique est pluridisciplinaire. La création des représentations culturelles prend racine en anthropologie cognitive. L'automatisation du processus ethnographique qui permet leur production est réalisée à partir des travaux en fouille de textes. La formalisation de ces représentations est possible grâce aux progrès en ingénierie des connaissances, en particulier en ingénierie des ontologies. La production des médiations à partir de représentations hétérogènes est, quant à elle, issue des recherches en alignement d'ontologies.

Les limitations

Le langage est à la fois une production culturelle et une partie du système culturel. L'analyse de différentes cultures est résolument multilingue. Pourtant, afin de simplifier cette thèse déjà complexe, cet aspect a volontairement été ignoré. C'est pourquoi, bien qu'adaptable à différentes langues, le travail présenté est monolingue.

La structure de la thèse

Cette thèse a une structure en trois chapitres principaux. Le premier a vocation à traiter l'ensemble des recherches associées, d'une part, au développement d'une conscience culturelle artificielle, et, d'autre part, à la solution que je propose. Le second chapitre traduit cette solution en développant les processus nécessaires à sa réalisation. Le troisième chapitre teste et critique la solution apportée par le biais d'une expérience. Mes contributions sont finalement résumées dans une conclusion qui s'ouvre vers des pistes d'études et d'améliorations.

1.2 Préambule

À ce jour, il n'y a aucun accord sur ce qu'est exactement la culture à part qu'elle s'apprend et se partage (Bennardo et De Munck, 2014). Kroeber et Kluckhohn (1952) ont recensé pas moins de 164 définitions. Faute de pouvoir avoir une définition consensuelle de la culture, Spencer-Oatey et Franklin (2012) ont essayé d'identifier des caractéristiques qui lui sont propres en classant de nombreuses citations provenant de divers chercheurs décrivant cette dernière. Dans ce préambule je cherche, sans essayer de donner de définition définitive, de préciser la notion de culture.

1.2.1 Les trois types de manifestations de la culture

En anthropologie la culture est définie par Tylor (1871) comme « ce tout complexe qui inclut la connaissance, la croyance, l'art, la loi, la morale, les coutumes, et toutes autres aptitudes et habitudes acquises par un membre de la société[1] ». Cette définition aborde trois manifestations différentes de la culture :

1. **Des motifs comportementaux** – La culture s'observe dans les comportements propres aux individus d'une société. Certaines situations impliquent systématiquement le même comportement culturellement défini. Par exemple en France, saluer quelqu'un que l'on rencontre pour la première fois se fait presque toujours par une poignée de main.
2. **La Culture Matérielle** – Les productions matérielles d'une société peuvent partager des motifs qui sont d'ordre culturel. C'est le cas pour les motifs genrés avec les habits bleus pour les garçons et roses pour les filles. Cependant, c'est le langage qui est certainement la plus importante production culturelle dans chaque société.
3. **Des idées partagées** – La culture n'est pas composée uniquement de motifs visibles (Schein, 1984, 1990). Elle est présente à différents niveaux de profondeur (Spencer-Oatey et Franklin, 2012), son origine se situant dans la tête sous la forme d'idées, de connaissances ou encore de croyances partagées. La partie visible de la culture résulte de l'expression de cette culture invisible. Par exemple, si l'on serre la main d'une personne que l'on rencontre, c'est parce que ce geste porte l'idée d'une

1. "that complex whole which includes knowledge, belief, art, law, morals, custom, and any other capabilities and habits acquired by man as a member of society".

forme de politesse et de respect. Si des personnes fabriquent des habits genrés, c'est parce qu'elles pensent qu'il y a fondamentalement une différence entre garçons et filles, et que le rose et le bleu véhiculent respectivement des valeurs féminines et masculines.

1.2.2 Un phénomène psychologique et social

La culture est autant une construction sociale que psychologique (Matsumoto, 1994). C'est une construction sociale puisqu'elle s'apprend et se partage, mais c'est aussi une construction psychologique puisqu'elle existe dans la tête des individus. C'est sa distribution à la fois psychologique et sociale qui fait que la notion de culture est floue (Spencer-Oatey et Franklin, 2012). Son partage non uniforme se manifeste notamment par la présence de variations intraculturelles au sein de grands groupes (Pelto et Pelto, 1975). Parce que la culture n'est pas partagée par tous, elle émerge par le biais du consensus.

1.2.3 Un système intégré

« Les différentes parties de la culture sont toutes, jusqu'à un certain degré, interdépendantes[2]. » (Spencer-Oatey et Franklin, 2012) Les cultures sont des « systèmes intégrés constitués de nombreuses parties interconnectées, si bien que des changements sur une partie de la culture vont certainement amener des changements dans les autres[3] » (Ferraro, 1998). La culture est perçue comme un système relativement stable ou invariant (Mathews, 2005). Cependant, sur le long terme, ses composants sont amenés à évoluer. Par exemple, le contexte social peut changer, ce qui va engendrer des changements culturels. C'est pourquoi la culture elle-même est sujette à des changements graduels.

2. "The various parts of a culture are all, to some degree, interrelated."
3. "integrated systems with a number of interconnected parts, so that a change in one part of the culture is likely to bring about changes in other parts".

Chapitre 2

Recherche

2.1 Introduction

En ce moment, le coeur du problème pour la modélisation d'une conscience culturelle artificielle se situe dans les représentations universelles grossières des cultures, sur lesquelles cette dernière se fonde pour acquérir une forme de conscience culturelle. Ma solution s'inspire des représentions culturelles émiques développées en anthropologie cognitive qui, après leur formalisation, vont offrir aux machines une compréhension plus profonde des cultures. Ce bouleversement en matière d'approche à de lourdes conséquences à la fois sur le développement d'une conscience culturelle artificielle mais aussi sur le processus ethnographique même qui conduit à la construction de représentations culturelles émiques. En effet, pour que ce genre de consciences culturelles artificielles puissent être mises en pratique, cela requiert une acquisition automatisée de ces représentations que ce soit au niveau de leur construction ou de leur formalisation. Ce changement d'approche invite aussi à repenser la création des médiations entre représentations culturelles qui jusqu'à présent étaient intrinsèquement interculturelles puisque dérivant de modèles universels. Ce nouveau défi s'inscrit dans des problématiques de recherches en alignement d'ontologies où la médiation de représentations hétérogènes se fait par le biais de leurs concepts et de leurs relations qui, eux, sont universels.

2.2 Une conscience culturelle artificielle sur une base étique

L'objectif premier de cette thèse est de contribuer à la modélisation d'une Conscience Culturelle Artificielle (*Artificial Cultural Awareness* – ACA). Le développement d'une ACA consiste en la capacité d'une Intelligence Artificielle (*Artificial Intelligence* – AI) à simuler une forme de conscience culturelle.

Cette partie a pour objectifs principaux (1) de comprendre où en est le développement actuel d'une forme de conscience culturelle artificielle au sein des Systèmes Culturellement Conscients (*Culturally-Aware Systems* – CASs) et (2) de cibler les limitations et proposer des solutions. Pour y parvenir, je vais tout d'abord commencer par un bref historique de l'intérêt porté aux CASs. Je vais ensuite présenter les principes fondamentaux d'une ACA. Je poursuivrai par un tour d'horizon des divers systèmes dits culturellement conscients. Puis, je vais étudier plus particulièrement le développement d'une forme de conscience culturelle dans ces systèmes. Après avoir souligné le problème de granularité de l'approche *étique* actuelle, j'expliquerai comment sa contrepartie *émique* peut apporter une réponse au problème et je présenterai aussi les nouveaux défis qui l'accompagnent.

2.2.1 Historique

Le sujet des CASs est récent et principalement étudié au sein de deux communautés de chercheurs : l'Apprentissage Assisté par la Technologie (*Technology Enhanced Learning* – TEL) et l'Interaction Homme-Machine (Human-Computer Interaction – HCI). Le premier atelier sur les Systèmes de Tutorat Culturellement Conscients (*Culturally-Aware Tutoring Systems* – CATS) s'est tenu en 2008 au Canada. Depuis, de nombreuses conférences ont visité ce sujet telles que EC-TEL, AIED, FLAIRS et *ACM Symposium on Applied Computing* (Mohammed et Mohan, 2013b).

2.2.2 Les principes fondamentaux d'une ACA

Dans l'enseignement, Baker explique que la conscience culturelle « accentue le besoin pour les élèves de devenir conscients des normes, croyances et comportements de leur propre culture et des autres cultures[1] » (Baker, 2011). La conscience culturelle se

1. "stress the need for learners to become aware of the culturally based norms, beliefs, and behaviours of their own culture and other cultures".

caractérise alors par la capacité à avoir conscience d'abord de sa propre culture et ensuite de celle des autres (Tomalin et Stempleski, 2013).

La conscience culturelle critique est composée d'une « compréhension relative de la nature des normes culturelles qui mène à la capacité d'évaluer, de manière critique et sur la base de critères explicites, les perspectives, pratiques et produits de sa propre culture et de son propre pays et de celle et celui des autres [2] » (Byram, 1997). Cette conscience se traduit par la capacité à être « conscient des similarités et différences entre groupes culturels [3] » (NCCC, 2004).

Pour résumer, les éléments constitutifs d'une conscience culturelle sont :

1. les représentations culturelles décrivant des groupes culturels ;
2. les médiations culturelles soulignant les similarités et différences entre ces groupes.

Cependant il faut comprendre que ce n'est pas le fait de posséder ces éléments qui donne accès à une conscience culturelle, mais la capacité à les produire. Une ACA doit être capable de réaliser ces deux tâches [4] de création et de médiation de représentations culturelles. Pour y parvenir, une ACA doit naturellement être à même de formaliser ces représentations, sans quoi elle ne peut accéder à une forme de compréhension des cultures, compréhension nécessaire à leur médiation. Pour résumer, une ACA doit être capable de :

1. créer des représentations culturelles ;
2. formaliser ces représentations ;
3. produire des médiations.

2.2.3 Des systèmes dits « culturellement conscients »

Blanchard *et al.* (2010) définissent les systèmes culturellement conscients comme « n'importe quel système où des informations en relation avec la culture ont eu des impacts sur sa conception, processus internes ou d'exécution, structures, et/ou objectifs [5] ». Cette définition reflète la diversité des systèmes dits culturellement conscients.

2. "understanding of the relative nature of cultural norms which leads to an ability to evaluate, critically and on the basis of explicit criteria, perspectives, practices and products in one's own and other cultures and countries".
3. "conscious of similarities and differences among cultural groups".
4. Ici « tâche » est employé au sens informatique du terme, c'est-à-dire comme la spécification d'un programme qui mime une compétence précise d'un être humain.
5. "any system where culture-related information has had some impact on its design, runtime or internal processes, structures, and/or objectives".

2.2.3.1 Les systèmes de gestion de données culturelles

Ces systèmes concernent le plus souvent des domaines relatifs à la culture – comme l'anthropologie, la géographie ou encore l'histoire – qui sont donc amenés à manipuler des données culturelles. Leur objectif est de récupérer, structurer et organiser ces données afin de pouvoir y accéder aisément.

Pour illustrer mes propos, il suffit de regarder l'application web associée à l'organisation *Human Relations Area Files* (HRAF). Cette organisation fondée en 1949 à l'université de Yale a pour domaine d'étude l'anthropologie culturelle. Son objectif est de faciliter les comparaisons entre cultures. HRAF possède deux collections électroniques : *eHRAF World Cultures* et *eHRAF Archaeology*. La première collection en particulier contient des données ethnographiques qui couvrent tous les aspects de la vie culturelle et sociale de 385 sociétés. Deux systèmes de classification permettent d'indexer ces données culturelles au sein d'une base de données : OCM (*Outline of Cultural Materials*) et OWC (*Outline of World Cultures*).

Dans les systèmes de gestion de données culturelles les données sont associées à un groupe culturel le plus souvent identifié par une zone géographique. Il n'y a donc pas de conscience à proprement parler de la part du système. Le système est conçu pour prendre en considération l'aspect social de la culture.

2.2.3.2 Les systèmes enculturés

L'enculturation est un terme qui désigne le processus par lequel un groupe va transmettre à l'enfant, dès sa conception, des éléments culturels. Attribué aux systèmes informatiques, ce terme est utilisé pour définir les systèmes dont la conception répond aux besoins relatifs à un groupe culturel (Blanchard *et al.*, 2010). Il est intéressant de noter que si tout système conçu est naturellement enculturé, puisque créé par au moins une personne faisant partie d'un groupe culturel, l'appellation *systèmes enculturés* dans la littérature fait uniquement référence à ceux où cet acte est volontaire.

L'enculturation est le plus souvent intégrée manuellement par des experts qui donnent naissance aux systèmes : des ingénieurs qui possèdent des connaissances sur la culture pour laquelle est destiné le système. Comme pour les systèmes de gestion de données culturelles, il n'y a pas de conscience culturelle de la part du système. Il y a uniquement de la part de l'ingénieur une gestion des manifestations culturelles dans l'application qu'il conçoit, qu'elles soient d'ordre matériel (l'interface) ou

comportemental (les méthodes et fonctions).

Le coût important en ressources (experts et temps) a mené à des travaux pour automatiser l'enculturation, ce qui a conduit aux systèmes s'adaptant à la culture.

2.2.3.3 Les systèmes d'adaptation

Ces systèmes abordent d'une part l'automatisation de l'enculturation et d'autre part la gestion de la diversité culturelle par l'adaptation. La méthode utilisée pour automatiser l'enculturation consiste à formaliser un ensemble de connaissances culturelles déclinables et à y adjoindre des règles ou fonctions qui peuvent s'exprimer sur la conception du système. L'adaptation est l'opération qui met en adéquation la culture de l'utilisateur avec l'enculturation du système. Le système doit ainsi être capable, après identification de l'origine culturelle d'un utilisateur, de l'associer à un modèle culturel interprétable et transposable pour produire une enculturation cohérente. L'identification se fait généralement soit en demandant à l'utilisateur de sélectionner sa culture soit par géo-localisation (notamment se servant de l'adresse IP).

Il existe de nombreux systèmes possédant un mécanisme d'adaptation. Par exemple, Reinecke et Bernstein (2009) ainsi que Mohammed et Mohan (2011) présentent tous deux une conception d'applications web dont l'enculturation est produite en appliquant des règles issues du modèle culturel d'Hofstede et Hofstede (2001). En partie basée sur le même modèle, Blanchard *et al.* (2009) décrit une architecture de système multi-agents qui adapte le contenu multimédia en fonction de profils culturels. Un travail intéressant réalisé par Mascarenhas et Paiva (2010) est la création de cultures synthétiques virtuelles qui prend en plus en considération dans le modèle culturel les symboles et les rituels. La tâche d'adaptation implique une certaine généricité du modèle culturel qui doit pouvoir se décliner relativement aisément pour un nombre conséquent de cultures.

Les systèmes d'adaptation font le lien entre un groupe social, une représentation formelle de leur culture et l'exploitation de cette représentation pour produire des éléments culturels. Les systèmes d'adaptation possèdent une conscience culturelle embryonnaire puisque des représentations culturelles formalisées sont utilisées.

2.2.3.4 Les systèmes interculturels

Ces systèmes se basent sur l'identification d'éléments culturels pertinents d'une culture à une autre. Les systèmes d'éducation interculturelle constituent une partie des

CATS dont l'objectif principal est le développement de connaissances et compétences interculturelles (Blanchard et Mizoguchi, 2014). Contrairement à ces derniers qui possèdent une vision d'apprentissage à long terme, les systèmes de collaboration interculturelle cherchent à fournir des recommandations suscitées par des situations culturelles particulières (Blanchard *et al.*, 2010).

Un exemple de système d'éducation interculturelle est le système de simulation immersive présenté par Johnson (2010b). L'utilisateur participe à des formations qui suivent la méthodologie de culture située (Situated Culture Methodology – SCM) (Johnson, 2010a). Dans le cas présent, cette méthodologie permet de déterminer les informations culturelles pertinentes pour des militaires américains en mission en Afghanistan. Cependant, ces derniers acquièrent seuls les compétences culturelles au travers de scénarios où ils interagissent via des dialogues et des gestes avec des personnages non-joueurs (Non-Player Characters – NPC) enculturés.

Les systèmes interculturels sont construits sur la base de modèles culturels permettant d'exposer les différences entre plusieurs cultures. Comme les systèmes d'adaptation, ils formalisent des représentations culturelles mais s'en distinguent en allant plus loin dans une forme de conscience culturelle puisqu'ils abordent leur médiation. Bien que ces systèmes soient les plus aboutis au regard d'une conscience culturelle, ils ne possèdent aucunement les capacités requises de création, de formalisation et de médiation de représentations culturelles.

2.2.4 Des représentations étiques

Une conscience culturelle est basée sur un ensemble de représentations issues d'une conceptualisation particulière de la culture. Il faut distinguer parmi les conceptualisations utilisées dans les culTELs (les systèmes culturels d'apprentissage assisté par la technologie) celles qui sont légitimes de celles qui sont illégitimes, produites selon une approche *folk* (Mohammed et Mohan, 2013a). « Les approches *folk* prennent racine dans les descriptions et perceptions subjectives et personnelles du contexte culturel qui est utilisé pour représenter des caractéristiques culturelles[6]. » (Mohammed et Mohan, 2013a) Les conceptualisations légitimes proviennent, elles, d'approches scientifiquement valides qui sont largement débattues et explorées en anthropologie et dans toutes les disciplines connexes.

6. "Folk approaches stem from subjective, personal descriptions and perceptions of cultural contexts which are used to represent cultural features."

Les représentations culturelles utilisées dans le développement des CASs sont basées principalement sur des systèmes de valeurs éprouvés tels que Hosfstede (Hofstede et Hofstede, 2001; Hofstede et Bond, 1984), GLOBE (House *et al.*, 2004) ou encore celui proposé par Schwartz (1994). Leur principe est d'identifier un certain nombre de dimensions formant *le noyau stable des cultures* (Blanchard et Mizoguchi, 2014). Les systèmes de valeurs constituent des modèles universels dits « étiques ».

Le système de valeurs de culture nationale créé par Hofstede est de loin le modèle le plus fréquemment utilisé pour la conception de systèmes culturellement conscients (Khashman et Large, 2013; Mascarenhas et Paiva, 2010; Reinecke et Bernstein, 2009; Marcus et Gould, 2000; Chandramouli *et al.*, 2008; Mohammed et Blanchard, 2015). Ce système a été produit en majeure partie en analysant les données de 50 pays et 3 régions dans le monde entre 1967 et 1982. Composé initialement de 5 dimensions, ce modèle a été étendu à 6 (Hofstede et Hofstede, 2001; Hofstede et Bond, 1984) :

1. **Évitement de l'incertitude** – Cette dimension exprime à quel point une société est mal à l'aise dans son rapport avec l'inconnu (le futur).
2. **Distance au pouvoir** – Ce critère définit le degré d'acceptation d'un pouvoir inégalement distribué.
3. **Collectivisme/Individualisme**. L'individualisme est la propension des individus à prendre soin d'eux-même et plus largement de leur famille proche, alors que le collectivisme représente la préférence de ces derniers au communautarisme.
4. **Masculinité/Féminité** – Cette dimension repose sur une vision sexiste des sociétés avec celles qui privilégient l'héroïsme, les résultats et les récompenses matérielles et les autres qui favorisent la coopération, la modestie, l'attention portée aux plus faibles et la qualité de vie.
5. **Orientation à long terme/Orientation normative à court terme**. Cette dimension fait état de la relation d'une société avec son passé, son présent et son futur.
6. **Indulgence/Retenue** – L'indulgence caractérise les sociétés qui se récompensent en se faisant plaisir (en s'amusant). Au contraire, la retenue s'applique aux sociétés qui régulent strictement les récompenses par des normes sociales.

Le tableau 2.1 est un exemple des représentations culturelles pour la France et le Canada selon le système de valeurs d'Hofstede. Chaque dimension a été quantifiée pour produire une représentation de ces deux pays. On observe que la France est décrite avec

Pays	DaP	Indi.	Mas.	EdI	OaLT	Indu.
France	68	71	43	86	63	48
Canada	39	80	52	48	36	68

TABLE 2.1 – Représentations culturelles de la France et du Canada selon le système de valeurs d'Hofstede (DaP : Distance au Pouvoir, Indi. : Individualisme, Mas. : Masculinité, EdI : Évitement de l'Incertitude, OaLT : Orientation à Long Terme, et Indu. : Indulgence)

un niveau assez élevé de Distance au Pouvoir, ce qui caractérise une société inégalitaire avec une hiérarchie présente que ce soit dans les répartitions géographiques ou dans les relations professionnelles. Le Canada, avec un score de 39, traduit quant à lui une société plus égalitaire. La dimension d'Individualisme est importante pour les deux pays, ce qui signifie que les membres de ces sociétés se fient davatange à eux-mêmes qu'aux collectifs auxquels ils appartiennent. Avec une Masculinité proche de la moyenne (43 pour la France et 52 pour le Canada), ces deux pays sont composés de personnes aspirant à un équilibre entre la poursuite de l'excellence et la recherche d'une bonne qualité de vie. La France a un très haut score d'Évitement de l'Incertitude, ce qui se caractérise par une société où tout est planifié, normé et régulé. Le Canada a une valeur moyenne de 48, ce qui indique une relativement bonne capacité à accepter des imprévus tels que de nouvelles idées. L'Orientation à Long Terme pour la France est de 63, décrivant ainsi une approche pragmatique permettant à la société d'évoluer. Le Canada, avec un score de 36, correspond à une société plus attachée aux traditions. Pour finir, concernant l'Indulgence, la France et le Canada, avec des valeurs de 48 et 68, représentent des sociétés où les récompenses sont, pour la première, d'ordre à la fois personnelles et socialement établies, et, pour la deuxième, principalement personnelles.

L'avantage des modèles étiques est qu'ils sont adaptés pour le développement d'une conscience culturelle artificielle puisqu'ils vont permettre de produire facilement des représentations culturelles pour différents groupes culturels, par exemple en quantifiant les dimensions par des scores qui sont représentatifs de ces groupes. Ainsi, un unique modèle étique va pouvoir se décliner pour représenter différentes cultures car il est fondé sur des concepts partagés. L'exemple sur la représentation de la France et du Canada avec le système de valeurs d'Hofstede permet d'exposer la facilité avec laquelle il est possible de décrire différentes sociétés sur la base de quelques critères supposément partagés.

2.2.5 Des médiations culturelles intrinsèques triviales

La production de médiations culturelles indicatives des similarités et des différences entre groupes culturels est un sujet absent des études faites sur la production de CASs. Les modèles étiques étant par définition interculturels, des médiations culturelles sont inévitables dès lors que des représentations culturelles en sont issues.

Si je reprends l'exemple des représentations de la France et du Canada selon le système de valeurs de culture nationale d'Hofstede, le tableau 2.1 offre une grille de lecture des continuités et discontinuités culturelles entre ces deux pays. La France et le Canada ont des médiations culturelles de similitude en ce qui concerne l'Individualisme et la Masculinité, et de divergence sur leur Distance au Pouvoir, l'Évitement de l'Incertitude, l'Orientation à Long Terme et l'Indulgence.

Les représentations culturelles issues de modèles étiques, en plus de fournir une compréhension des différentes cultures, permettent de manière triviale de prendre conscience des similarités et différences entre ces dernières. C'est pourquoi ce sont ces représentations qui pré-dominent dans le développement des CASs.

2.2.6 Vers une conscience culturelle émique

J'ai montré dans cette partie que les modèles étiques sont adaptés au développement d'une ACA. En effet, ces modèles permettent de produire facilement un grand nombre de représentations culturelles. De plus, ces représentations possèdent intrinsèquement les fondations pour leur médiation. La formalisation de ces représentations en un format interprétable par une machine va ensuite permettre à cette dernière d'accéder à une compréhension des cultures représentées ainsi qu'à leurs similarités et différences, en d'autres termes à une forme de conscience culturelle.

Cependant, comme l'indique Mohammed et Mohan (2013a) dans leur recherche sur les culTELs, ces modèles culturels étiques posent un problème de granularité. En effet, les représentations culturelles fondées sur ces modèles sont à la fois trop globales et grossières. Par exemple, le modèle d'Hofstede est applicable au niveau de pays alors qu'une ACA a besoin de représentations à l'échelle de communautés. Pour continuer avec le système de valeurs d'Hofstede, il n'est pas non plus assez détaillé pour qu'une machine puisse accéder à une compréhension même relative des cultures représentées à l'aide de ce modèle. Le fait de savoir que la France et le Canada ont un haut niveau d'Individualisme n'apporte aucune indication sur la cause qui est certainement spécifique aux deux pays. Ainsi, l'utilisation

de modèles étiques conduit à une compréhension limitée des cultures par les CASs.

Les représentations culturelles sont le résultat d'un processus ethnographique qui se définit comme le processus de collecte, de stockage et de recherche de motifs pour décrire une culture. La distinction entre les approches étique et émique a été initialement proposée par Pike (1954). L'approche étique prend place dans une optique ethnologique qui a donc pour objectif l'analyse comparative des cultures humaines. L'approche étique cherche ainsi à trouver des universaux grâce à une vue extérieure des cultures. À l'inverse, l'approche émique cherche à identifier les spécificités propres à chaque culture par une vue de l'intérieur. Cette approche mène à la production de représentations sur-mesure, détaillées et uniques qui permettent d'accéder à une réelle compréhension des cultures décrites. Les représentations émiques sont donc susceptibles d'apporter une solution au problème de granularité rencontré dans le développement des ACAs.

Imaginer une conscience culturelle sur la base de représentations émiques permet potentiellement de repousser les limitations actuelles, mais introduit aussi de nouveaux défis :

1. **Le coût de la création des représentations culturelles émiques** – Il est vivement souhaitable qu'une ACA puisse traiter un grand nombre de cultures différentes. C'est pourquoi elle doit pouvoir disposer d'un nombre important de représentations culturelles. Si leur accès et leur création sont aisés dans une démarche étique, ce n'est pas le cas pour les représentations émiques qui nécessitent un travail ethnographique plus conséquent.
2. **La complexité de la production de médiations culturelles** – Alors que la production de médiations est triviale pour des représentations culturelles issues d'un modèle étique, elles ne le sont plus pour les représentations émiques qui sont intrinsèquement uniques. La production de médiations à partir de représentations émiques se présente donc comme un nouveau défi à relever.

2.3 Les modèles culturels prototypiques

La contribution majeure que je souhaite apporter à la modélisation d'une ACA porte sur la base émique de sa conscience culturelle. Cette conscience implique naturellement de pouvoir accéder à un nombre conséquent de représentations culturelles émiques. Les principaux objectifs de cette partie sont (1) de définir ce que sont les Modèles Prototypiques Culturels (Prototypical Cultural Models – PCMs) et (2) de comprendre le processus ethnographique qui mène à leur construction afin de pointer et de proposer une solution au problème de l'approche actuelle. Je commence par un rapide historique sur l'anthropologie cognitive. Je poursuis en définissant les PCMs en partant des modèles mentaux pour aller jusqu'aux modèles culturels [7] dont ils sont issus. Ensuite, je présente les trois techniques qui composent le processus ethnographique de découverte de PCMs que sont : l'échantillonnage, l'explicitation des modèles mentaux et l'analyse de consensus. Je continue en exposant le déroulement du processus classique de création d'un PCM. Je finis par expliquer l'incompatibilité du coût en temps de l'approche actuelle avec le besoin de mise à l'échelle, incompatibilité qui appelle nécessairement à l'automatisation du processus.

2.3.1 Historique

L'anthropologie cognitive est une discipline qui est apparue dans les années 1950-1960 sous différents noms : ethnosémantique, ethnoscience, ethnolinguistique, et nouvelle ethnographie. Cette discipline aborde la manière dont un groupe culturel perçoit le monde en faisant le lien entre la cognition et la culture (Sturtevant, 1964). L'anthropologie cognitive est purement descriptive et cherche à représenter les conceptualisations culturelles de natifs – appelées des modèles *folk*. L'anthropologie cognitive embrasse une perspective complètement émique puisque les « ethnographes doivent découvrir les principes organisateurs d'une culture – le monde sémantique des natifs – tout en évitant l'imposition de leur propres catégories sémantiques sur ce qu'ils percoivent [8] » (Corsaro et Heise, 1990).

L'importance grandissante des théories sur les modèles mentaux introduits par Craik (1943) et sur le consensus sont à l'origine de l'apparition dans le milieu des années 1980 des

7. Dans cette partie la notion de *modèle culturel* est à distinguer de l'utilisation précédemment faite qui était délibérément relativement abstraite.

8. "[e]thnographers must discover the organizing principles of a culture – the semantic world of the natives – while avoiding the imposition of their own semantic categories on what they perceive".

modèles cognitifs culturels (Blount, 2013) qui sont à l'origine des représentations émiques les plus abouties.

2.3.2 Les modèles mentaux

Jones *et al.* (2011) décrit un modèle mental comme « une représentation simplifiée de la réalité qui permet aux gens d'interagir avec le monde [9] ». Des concepts similaires ont été développés en intelligence artificielle, en sciences cognitives et en linguistique : les *frames* (Bateson, 1955; Lees et Chomsky, 1957), scripts (Schank et Abelson, 1977), schemata (Rumelhart, 1975).

2.3.2.1 Des conceptualisations partielles

Les modèles mentaux sont des conceptualisations d'une taille limitée. La conceptualisation est un procédé cognitif fondamental de schématisation et de catégorisation (Sharifian, 2003). La schématisation implique la sélection discriminante de certains aspects pour décrire un tout (Talmy, 1983). La catégorisation regroupe des entités distinctes pour être traitées de manière similaire (Rosch et Lloyd, 1978). Bien qu'une conceptualisation soit l'abstraction d'un phénomène par ses concepts principaux et ses relations importantes (Studer *et al.*, 1998), elle est le plus souvent représentée schématiquement par un ensemble de mots interconnectés sous la forme d'un réseau sémantique.

« La taille maximale d'un modèle mental est déterminée par la capacité de la mémoire de travail, l'espace de travail mental sur lequel les gens stockent temporairement de l'information [10]. » (Doyle, 1998) La mémoire de travail ne peut traiter parallèlement qu'un nombre fixe d'informations. Ce nombre défini à 7 plus ou moins 2 par Miller (1956) limite la taille du réseau conceptuel d'un modèle mental.

2.3.2.1.1 Les concepts

Le triangle sémiotique d'Ogden *et al.* (1923) décrit les relations entre concepts, objets et symboles (voir la figure 2.1). Un concept fait référence à un objet concret ou abstrait et le concept lui-même est exprimé au travers d'un symbole qui représente un objet.

9. "a simplified representation of reality that allows people to interact with the world".
10. "The maximum size of a mental model is thought to be determined by the capacity of working memory, the mental workbench on which people store information temporarily while thinking about it."

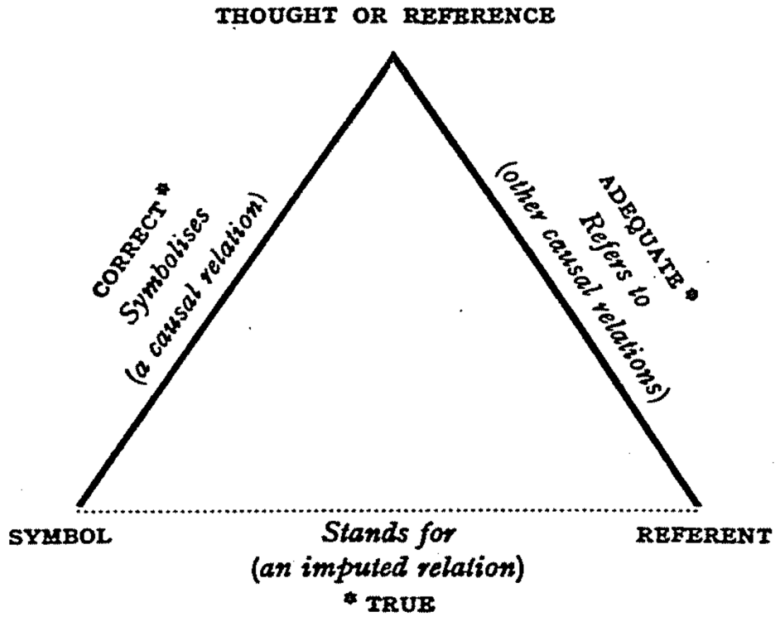

FIGURE 2.1 – Le triangle sémiotique Ogden *et al.* (1923) dans sa version originale. Pour la source, se référer à Wikipédia : https://fr.wikipedia.org/wiki/Signe_linguistique

2.3.2.1.2 Les relations lexico-sémantiques

Les relations constitutives d'une conceptualisation sont majoritairement sémantiques, c'est-à-dire qu'elles ont du sens. Généralement, ces relations sont représentées par des tuples mettant en relation un minimum de deux concepts et une classe sémantique. Parmi les relations sémantiques, certaines ont une importance toute particulière puisqu'elles sont empreintes d'une logique fondamentale régissant l'organisation structurelle des conceptualisations. Ce sont les relations lexico-sémantiques. Elles sont définies par des triplets (c_1, r, c_2) où c_1 et c_2 sont des concepts et r une classe lexico-sémantique.

Les relations lexico-sémantiques sont des relations paradigmatiques et non pas syntagmatiques. Un syntagme est un groupe d'éléments mis en séquence de manière grammaticalement cohérente alors que les relations paradigmatiques se situent entre les éléments qui appartiennent à une même catégorie. La figure 2.2 montre la différence entre les deux types de relations.

Les relations lexico-sémantiques expriment des fonctions cognitives fondamentales. Leur universalité est confirmée par leur présence dans toutes les cultures (Khoo et Na, 2006). Elles constituent des notations interculturelles adéquates puisqu'elles sont étiques (Wierzbicka, 2006).

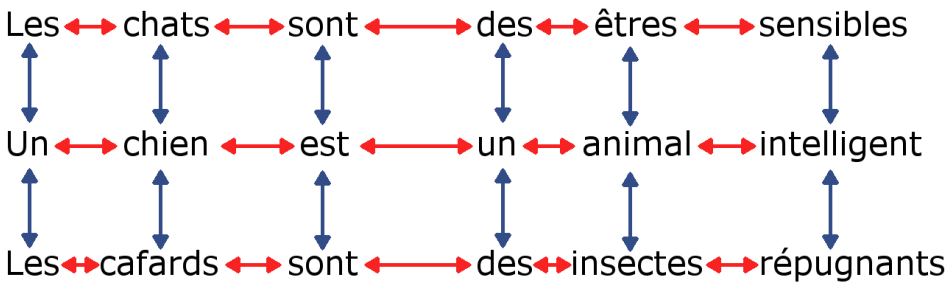

FIGURE 2.2 – Exemple de relations syntagmatiques (en rouge) et paradigmatiques (en bleu)

Les relations lexico-sémantiques ont ainsi un certain nombre de propriétés identifiées par les linguistes (Khoo et Na, 2006; Murphy, 2003). Khoo et Na (2006) présente 8 propriétés dont 7 générales sont exposées ci-dessous :

1. **Productive** – Les relations lexico-sémantiques sont des relations logiques, il est donc possible d'en produire des nouvelles à partir d'existantes (Sowa, 1983; Cruse, 2011).
2. **Variable** – Le sens des mots présents dans les relations varie en fonction des autres mots avec lesquels ils sont en relation.
3. **Prototypique** – Certaines paires de termes sont meilleures que d'autres pour illustrer une relation lexico-sémantique. Par exemple, *(chien, animal)* est un meilleur exemple que *(vison, animal)* pour la classe sémantique *est un*.
4. **Semi-sémantique** – L'existence des relations lexico-sémantiques est déterminée par des propriétés non sémantiques telles que la proximité cognitive des termes qui les composent. La semi-sémanticité fait référence à la dimension humaine (psychologique et cognitive) des relations lexico-sémantiques.
5. **Innombrable** – Le nombre de relations lexico-sémantiques n'est pas défini. C'est un groupe dynamique.
6. **Prévisible** – Ce sont des relations qui suivent certains motifs et règles, c'est ce qui permet notamment de les identifier dans du texte (Hearst, 1992).
7. **Universelle** – Ce sont des relations étiques qui caractérisent des fonctions cognitives fondamentales.

Il n'y pas d'accord sur une liste exhaustive de classes sémantiques. Cependant, la tâche 2 du SemEval-2012 (Jurgens *et al.*, 2012) présente une base intéressante constituée

de 79 classes divisées en 10 catégories. Toutefois, les classes sémantiques les plus communes sont l'hyperonymie, la méronymie, la causalité, la synonymie et l'antonymie. Les relations d'hyperonymie constituent le squelette de la majeure partie des conceptualisations. Elles soutiennent les structures taxonomiques en représentant des relations de type classe-inclusion comme *(policier, est un, être vivant)*. Les méronymes expriment des relations de type partie-tout ou membre-tout tel que *(victime, partie de, crime)*. Tout comme les deux types de relations précédentes, la causalité est une relation transitive. *(accident, cause des, dégats)* en est un exemple. La synonymie et l'antonymie sont des relations symétriques qui définissent respectivement l'équivalence et l'opposition avec pour exemples *(voiture, équivaut à, auto)* et *(légal, opposé à, illégal)*.

2.3.2.2 Des abstractions imparfaites

Les modèles mentaux sont des abstractions partielles et imparfaites. Elles ne représentent que ce qui est considéré vrai. Ces représentations peuvent potentiellement contenir des erreurs (Moray, 1997) et même des propositions contradictoires (Byrne, 2005).

2.3.2.3 Des modèles dynamiques

Les modèles mentaux sont dynamiques. De nouvelles connaissances sont produites à partir des connaissances premières (Wiig, 2004), soit par raisonnement soit par apprentissage. Le raisonnement est achevé grâce à la structure combinatoire des modèles mentaux (Rutherford et Wilson, 1989). L'apprentissage émerge des expériences qui résultent de l'interprétation des données externes.

La connaissance ainsi créée est soit assimilée soit accommodée (Piaget et Cook, 1952). L'accommodation fait référence au processus qui permet d'intégrer indirectement la nouvelle connaissance en adaptant la structure du modèle mental. L'assimilation, quant à elle, est le procédé qui intègre directement la nouvelle connaissance sans que des modifications n'apparaissent.

2.3.2.4 La réalité cognitive des modèles mentaux

Il n'y a pas d'accord sur l'endroit où sont stockés les modèles mentaux (Jones *et al.*, 2011), dans la mémoire à : long terme (Craik, 1943), court terme (Johnson-Laird, 1983) ou les deux (Nersessian, 2002). Ces propriétés cognitives semblent être intimement liées à

celles de la connaissance. La partie tacite (inconsciente) des connaissances qui composent les modèles mentaux serait alors stockée dans la mémoire à long terme tandis que la partie explicite (consciente), elle, serait stockée dans la mémoire à court terme.

2.3.2.5 Les modèles mentaux comme schémas d'interprétation

En se basant sur des travaux sur la pensée analogique, Gentner et Gentner (1982) et Collins et Gentner (1987) ont prouvé que les modèles mentaux pouvaient être utilisés comme des cadres d'inférence. Ils permettent l'interprétation des données externes en agissant comme cadres d'interprétation (Chi, 2008; Johnson-Laird, 1983). Tsuchiya (1993) explique le mécanisme de cette interprétation : « Bien que les termes "donnée", "information" et "connaissance" sont souvent utilisés de manière interchangeable, il existe une distinction claire entre eux. Quand du sens est donné à la donnée au travers du framework d'interprétation, elle devient information, et quand le sens de l'information est lu au travers du framework d'interprétation, elle devient connaissance[11]. »

Jones a observé que « les gens tendent à filtrer les nouvelles informations en fonction soit de leur congruence, soit de leur compréhension existante[12] » (Jones *et al.*, 2011). Cette tendance à chercher les informations qui soutiennent ce qui est considéré comme vrai est un phénomène appelé le « biais de confirmation » (Bank, 2015). Ce phénomène reflète l'imperfection des modèles mentaux dans l'interprétation.

2.3.3 Les modèles cognitifs culturels

Les modèles mentaux vont des idiosyncratiques aux universels. Les modèles idiosyncratiques sont personnels, formés par les expériences propres à chacun. Les modèles universels sont quant à eux liés à des facultés innées telles que les propriétés du système visuel (Bennardo et De Munck, 2014). Les modèles culturels se situent entre ces deux extrêmes. Ce sont des modèles mentaux collectifs épousant la théorie cognitive de la culture.

11. "Although terms "datum", "information", and "knowledge" are often used interchangeably, there exists a clear distinction among them. When datum is sense-given through interpretative framework, it becomes information, and when information is sense-read through interpretative framework, it becomes knowledge."
12. "people tend to filter new information according to its congruence or otherwise with their existing understandings".

2.3.3.1 La théorie cognitive de la culture

Pour D'andrade (2001), il y a maintenant un accord en anthropologie cognitive sur le fait que la culture est composée de symboles, de concepts et ultimement de sens. D'après Bennardo et De Munck (2014), ce consensus était déjà présent parmi les sociologues et socio-anthropologues (Bartlett, 1932; Kroeber et Parsons, 1958; Turner, 1967). La théorie cognitive de la culture situe la culture dans la tête d'individus en tant que système de connaissances socialement apprises et partagées (Goodenough, 1981). Ainsi, « les bâtiments, les comportements, les films, les artefacts préhistoriques et tout le reste "en dehors" ne *sont* pas culture[13] » (Bennardo et De Munck, 2014), ce sont des expressions culturelles.

La connaissance est distribuée. Tous les individus appartenant à une population possèdent leur propre jeu de connaissances plus ou moins propagées au sein des communautés qu'ils fréquentent. Il est alors très peu probable qu'une connaissance soit partagée par l'ensemble d'une communauté. Ainsi, ce qui constitue les bases de la culture c'est la connaissance socialement enracinée et consensuellement partagée.

2.3.3.2 Des modèles construits et partagés socialement

Les modèles culturels émergent d'un « processus social d'élaboration, de communication et de dissémination d'un système de connaissance[14] » (Wagner et Hayes, 2005). Ils sont définis comme des modèles mentaux pris pour acquis et distribués, socialement construits et inter-subjectivement partagés par un groupe social (Gee, 2008; Holland et Quinn, 1987; D'Andrade, 1987). Cependant, les modèles culturels sont distribués. C'est pourquoi, bien que chaque personne appartenant au même groupe possède des modèles culturels similaires, ils n'en restent pas moins légèrement différents.

2.3.3.3 Des modèles matures et stables

Les modèles culturels se construisent dans la durée. C'est une conséquence du processus social. Ainsi, ces modèles possèdent une certaine maturité. C'est ce que confirment les travaux de Mathevet *et al.* (2011) sur la gestion de l'eau dans la réserve de la biosphère de la Camargue. L'analyse des modèles mentaux des membres du comité de gestion révèle que les nouveaux arrivants ont des modèles mentaux plus hétérogènes

13. "buildings, behaviors, movies, prehistoric artifacts, and anything else "out there" [disqualify] from *being* culture".
14. "a historical social process of elaboration, communication and dissemination of knowledge systems".

que ceux des membres formant le noyau homogène du comité. Ces modèles mentaux collectifs, de par leur partage, convergent vers des modèles stables et durables (Berninger *et al.*, 2009; Mathieu *et al.*, 2005).

2.3.3.4 Des modèles mentaux tacites

La connaissance est personnelle (Polanyi, 1958). Elle est profondément ancrée dans l'inconscient dans un état tacite. Pour que la connaissance devienne explicite, elle doit devenir objet de pensée (Alexander *et al.*, 1991). Les modèles culturels sont composés de connaissances culturelles (D'Andrade, 1995; Eraut, 2000; Gee, 2008). Ces connaissances sont tacites, c'est-à-dire qu'elles n'ont pas besoin d'être explicitées pour être communiquées car elles sont partagées de manière intersubjective. C'est pourquoi les modèles culturels sont particulièrement tacites.

2.3.4 Une représentation émique des modèles culturels

Le processus ethnographique de découverte de modèles culturels est ce qui va produire les Modèles Culturels Prototypiques (Cultural Prototypical Models – PCMs). Ce sont des représentations abstraites consensuelles qui forment des « modèles culturels standards ou prototypiques qui sont partagés mais qui ne sont tenus par [personne] [15] » (Bennardo et De Munck, 2014). Le processus ethnographique est basé sur trois techniques : l'échantillonnage, l'explicitation et l'analyse de consensus.

2.3.5 L'échantillonnage

L'échantillonnage constitue la première partie de tout processus ethnographique. Cette tâche est basée sur l'idée que la culture est socialement construite. L'objectif est de capturer un nombre représentatif d'individus appartenant à un même groupe social et donc étant susceptibles de partager une même culture. Généralement, cette tâche est achevée en identifiant une communauté. Une communauté est définie par Wasserman et Faust (1994) comme un ensemble d'individus possédant des relations sur le long terme, fortes, directes, intenses, fréquentes et positives.

15. ""prototypical" model[s] that [are] shared but not held by [anyone]".

2.3.5.1 L'identification d'une communauté

De manière générale les anthropologues sélectionnent leurs échantillons au travers de critères sociaux partagés tels que les genres, les religions, les professions ou les lieux – lieux de travail (Mathieu *et al.*, 2009), villes (Young, 1980) ou régions (Vuillot *et al.*, 2016). La force de cette méthode est qu'elle est facile à utiliser. Cependant, sa faiblesse est qu'elle ne peut garantir que les individus sélectionnés appartiennent réellement à une même communauté car elle est peu précise.

Des techniques plus performantes mais aussi plus coûteuses, issues des sciences sociales, permettent d'identifier avec plus de précision des individus appartenant à une même communauté. C'est le cas par exemple des algorithmes de détection de communauté qui viennent de l'analyse de réseaux sociaux.

2.3.5.2 La sélection du bon nombre d'individus

Après avoir identifié la communauté d'intérêt, il est nécessaire de s'assurer que la taille de l'échantillon soit bonne, c'est-à-dire qu'elle maximise la représentativité du groupe étudié tout en minimisant les futurs coûts liés au traitement. Ainsi, la taille de l'échantillon doit permettre de garantir la validité culturelle des résultats qui seront produits en étant de la plus petite taille possible.

La sélection peut se faire de deux manières différentes : soit exploratoire, soit confirmatoire. L'approche exploratoire cherche à sélectionner directement un nombre suffisant d'individus pour conduire l'analyse. L'approche de confirmation permet, à partir des résultats d'une première analyse, de déterminer le nombre d'individus à reprendre si nécessaire.

Pour déterminer la taille des échantillons, Weller (1987) se base sur l'hypothèse de la dépendance des cas. Contrairement à l'hypothèse d'indépendance des cas en statistique où les individus statistiques ont une distribution aléatoire/indépendante, les individus culturels ont une distribution naturellement convergente/dépendante. Sur la base de cette hypothèse, il a démontré que la formule de la prophétie de Spear-Brown peut être appliquée aux échantillons. Cette formule peut mettre en relation le nombre d'individus, le niveau de consensus existant au sein d'un groupe et la fiabilité des connaissances culturelles obtenues.

En suivant une approche exploratoire basée sur cette formule, la taille optimale pour l'échantillon se situe entre 29 et 48 individus (Weller, 2007; Gravlee, 2011). Le nombre

minimum d'individus pour composer ce dernier peut-être de 3 si leur niveau moyen d'accord est de 0.9, ce qui permet d'atteindre un niveau de confiance de 0.96 et de validité de 0.99.

La sélection aléatoire n'étant pas obligatoire dans un contexte culturel, l'échantillonnage peut se faire en fonction de certaines préférences de la part des ethnographes. Néanmoins, si les individus peuvent être sélectionnés par commodité, ils ne doivent pas être dépendants les uns des autres (Bennardo et De Munck, 2014). En effet, si la réduction de la taille de l'échantillon s'effectue sur des critères sociaux, il y a un fort risque d'obtenir au final une sous-communauté qui ne serait alors plus représentative de la communauté initialement identifiée.

2.3.6 L'explicitation des modèles mentaux

L'explicitation est le processus permettant d'expliciter des connaissances tacites ou encore d'extérioriser des structures de connaissances internes. Le but de l'explicitation est de rendre tangible la connaissance des individus.

2.3.6.1 Les différentes approches d'explicitation

Shadbolt et Smart (2015) présentent un très grand nombre de techniques d'explicitation. Différentes classifications ont été proposées en fonction du contexte ou du domaine dans lequel étaient étudiées ces techniques (Gavrilova et Andreeva, 2012; Burge, 2001; Milton, 2008; Jones *et al.*, 2011).

Par exemple, la taxonomie que propose Gavrilova et Andreeva (2012) est réalisée dans une perspective de gestion des connaissances. Leur classification se construit au niveau de la relation entre les analystes et les experts. L'explicitation peut prendre trois formes différentes : elle peut être guidée par les experts, orientée par les analystes ou être collaborative. Les brainstormings et les tables rondes constituent des techniques collectives d'explicitation.

Dans le cas de l'explicitation des modèles mentaux, seules les techniques d'explicitation individuelles sont utiles. Jones *et al.* (2011) proposent une classification compatible avec la dimension individuelle en distinguant deux catégories d'explicitation des connaissances. Dans une explicitation directe, les connaissances sont explicitées sans intermédiaire alors que dans une explicitation indirecte, elles sont acquises par l'analyse de données collectées.

2.3.6.1.1 L'explicitation directe

Dans l'explicitation directe, le participant formalise directement ses propres connaissances sans intermédiaire. Néanmoins, il est souvent assisté dans cette tâche.

La méthode 3CM est la technique la plus représentative de cette catégorie. Basée sur une entrevue, elle a pour but d'expliciter spatialement la carte cognitive du participant en le faisant écrire des mots et leurs relations au sein d'un domaine. Cette technique a deux variations. Dans l'implémentation ouverte le participant doit construire sa carte à partir de rien alors que dans l'implémentation structurée il est contraint d'utiliser une liste de mots qu'on lui a fourni préalablement.

2.3.6.1.2 L'explicitation indirecte

Dans l'explicitation indirecte, les connaissances émergent de l'analyse de données non structurées produites par un individu. C'est donc un processus en deux étapes, la première se concentrant sur la collecte des données et la deuxième sur leur analyse.

Les approches pour collecter des données sont au nombre de trois :

1. **L'entrevue** – Il y a plusieurs types d'entrevues : informelles, non structurées, semi-structurées et structurées. Les entrevues informelles se caractérisent par un manque de contrôle lors de leur déroulement. C'est le genre de discussions que l'on peut avoir par exemple dans un bar. Dans les entrevues non structurées, il faut suivre un plan clair mais avec des questions ouvertes. Les entrevues semi-structurées suivent un guide : une liste de questions avec un sujet précis à aborder. Les entrevues structurées sont quant à elles méticuleusement préparées dans les moindres détails. Plus l'entrevue est structurée, plus le risque de biais est important, mais aussi plus rapide et meilleure est l'explicitation.
2. **L'observation** – L'observation peut s'effectuer de différentes manières, soit « complètement participant », soit « participant-observateur », soit « complètement observateur ». Être « complètement participant » signifie que l'observation se fait de l'intérieur en s'intégrant dans l'environnement de l'individu étudié. L'avantage de cette approche est qu'elle permet d'obtenir des observations fines. À l'inverse, être « complètement observateur » veut dire que l'on observe l'individu de l'extérieur, tel un espion, en évitant toute interaction. Bien entendu, l'espionnage est un extrême qui n'est pas éthique. Ce genre d'abus est banni des pratiques autorisées. En restant à distance, on diminue les risques

d'influencer la personne et donc de minimiser le biais produit par l'ethnographe lors de l'observation. L'observation en tant que « participant-observateur » est communément utilisée en anthropologie. C'est une approche hybride qui se fait par alternance entre « complètement participant » et « complètement observateur ».

3. **La collecte de données brutes** – Les entrevues et observations sont incontournables pour obtenir des données encore indisponibles. Si ces données sont déjà disponibles, la collecte de données brutes est une alternative de choix. En effet, les entrevues, de leur planification à leur réalisation, constituent un processus qui se compte au minimum en jours. Les observations, quant à elles, s'étendent plutôt sur des mois ou des années. De plus, dans certaines situations, ni les entrevues ni les observations ne sont possibles. Par exemple, l'archéologie permet de récolter des données qui ne peuvent plus être produites. La collecte de données existantes à partir du web offre aussi de nouvelles perspectives avec une ethnographie en temps réel. D'ailleurs, le web fournit un accès à une quantité phénoménale de données sur les nombreuses populations qui s'y connectent, que ce soit du texte, des bandes sonores ou bien des vidéos.

La majorité des données collectées ou collectables sont des données linguistiques. Leur analyse peut se faire soit de manière qualitative soit de manière quantitative. Lors d'une analyse qualitative, l'expert va traiter manuellement le contenu afin d'en extraire des connaissances riches et précises. C'est un processus coûteux en temps et sujet au biais. Les connaissances produites le sont à partir de données limitées qui ne sont donc pas statistiquement représentatives.

En revanche, une analyse quantitative traite un grand volume de données. Ainsi, les connaissances trouvées sont statistiquement représentatives. Bien que ce processus soit moins précis que dans l'analyse qualitative, il est automatique.

2.3.6.2 Les critères importants dans l'explicitation

Comme l'expliquent Jones *et al.* (2011), les techniques d'explicitation ne sont pas toutes également efficaces. Chacune possède ses points forts et ses points faibles. La qualité d'une explicitation est généralement évaluée par rapport à trois critères :

1. **Granularité** – La granularité définit le niveau de précision avec lequel la connaissance est explicitée. On qualifie la granularité de fine quand il y a un haut

degré de détails, et de grossière dans le cas contraire.

2. **Dimension tacite** – La capacité à gérer l'aspect tacite de la connaissance va de pair avec la possibilité d'expliciter les structures fondamentales qui la composent. Cet aspect est particulièrement important dans un contexte culturel où la connaissance est fortement tacite (D'Andrade, 1995; Eraut, 2000; Gee, 2008).
3. **Biais** – L'explicitation cherche à acquérir telle quelle la connaissance d'un individu. Cependant, de nombreuses interférences externes durant ce processus peuvent altérer le résultat final. Par exemple, lors d'une entrevue le participant peut se comporter différemment pour de multiples raisons en relation avec son interlocuteur. La fiabilité de la connaissance explicitée dépend ainsi de la propension à conserver le biais initial.

2.3.6.3 Les paramètres externes impactant la qualité

La qualité du processus d'explicitation ne se limite pas à prendre en considération uniquement les techniques qui le compose. En effet, le participant et l'environnement dans lequel se déroule l'explicitation sont aussi des paramètres qui peuvent influencer la granularité de la connaissance explicitée, la capacité à sonder sa dimension tacite ainsi que le biais associé.

2.3.6.3.1 L'impact de l'expérience du participant

Un participant effectue quatre tâches durant l'explicitation : comprendre la question, se rappeler des informations pertinentes, effectuer un jugement et produire une réponse structurée (Meyer et Booker, 2001). D'après McBride et Burgman (2012), ces tâches sont associées à un biais psychologique qui impacte la qualité des connaissances explicitées. Des actions durant l'explicitation peuvent être entreprises pour aborder le biais psychologique telles que motiver, structurer, conditionner, encoder et vérifier, et ainsi améliorer les résultats. Il existe aussi d'autres actions pré-existantes. En effet, en tant que processus psychologique, la capacité d'un individu à expliciter ses connaissances se travaille. C'est pourquoi s'entraîner à l'explicitation permet d'en améliorer la fiabilité et la précision (Ferrell, 1994; Renooij, 2001).

Par exemple, un professeur sera certainement plus à même d'expliciter en détail ses connaissances car il est amené à le faire régulièrement lors de ses cours. De la même manière, des personnes travaillant à l'international seront peut-être plus conscientes de

leurs propres connaissances culturelles.

Ainsi l'expérience même du participant est une variable dans le processus d'explicitation.

2.3.6.3.2 L'influence de l'environnement

Jones *et al.* (2014) ont étudié les différences entre les procédures d'explicitation situées et non situées. Une procédure non située implique d'effectuer l'explicitation dans un lieu sans rapport avec les connaissances que l'on cherche à expliciter. Au contraire, lors d'une explicitation suivant une procédure située, le participant est dans un environnement en lien avec le motif de l'explicitation.

Les résultats de l'expérience conduite par Jones *et al.* (2014) ont démontré qu'être dans un environnement adéquat permet une explicitation plus complète qui se traduit notamment par un nombre plus élevé de concepts. Néanmoins, bien que les résultats obtenus ne permettent pas d'indiquer si cette amélioration ne se fait pas au détriment d'un biais produit par l'environnement lui-même, il est difficile d'imaginer que le fait de placer une personne dans un environnement suggestif n'ait aucun impact sur les résultats.

2.3.6.4 Les différentes approches pour expliciter un modèle mental

Après leur enquête sur l'explicitation des modèles mentaux, DeChurch et Mesmer-Magnus (2010) distinguent deux approches clefs. Dans la première approche, les concepts sont fournis aux participants afin qu'ils les relient pour former un modèle mental. Dans la deuxième approche, les modèles mentaux sont produits dans leur totalité par les participants. Ces deux approches sont cohérentes avec les façons d'expliciter les structures de connaissances que l'on retrouve dans les implémentations structurées et ouvertes de la méthode 3CM (Kearney et Kaplan, 1997).

Cependant, la première approche décrite par DeChurch et Mesmer-Magnus (2010) n'est pas obligatoirement bipartite (concepts/relations). En effet, un participant peut avoir à associer des concepts qu'on lui a suggéré, mais aussi identifier les classes sémantiques d'une liste de relations. Cette approche constitue donc une démarche incrémentale d'explicitation des connaissances. En d'autres termes, on distingue deux approches : une entière/ouverte et une structurée/itérative.

2.3.7 L'analyse de consensus

En tant que théorie, l'analyse de consensus permet d'opérationnaliser la culture (Bennardo et De Munck, 2014). La Théorie du Consensus Culturel (Cultural Consensus Theory – CCT) « formalise l'intuition selon laquelle l'accord entre individus est fonction de l'étendue avec laquelle chacun connaît la "vérité" culturellement définie[16] » (Kempton *et al.*, 1996).

2.3.7.1 L'analyse de consensus en tant que méthode

La CCT est fondée sur trois hypothèses restrictives (Garro, 2000; Weller, 2007).

1. L'échantillon doit partager une **culture unique** afin de garantir que les données soient cohérentes.
2. La collecte des données des individus doit se faire de manière **indépendante**. Cette condition permet d'éviter le biais de ces dernières causé par des interactions collectives.
3. Les participants doivent posséder un **niveau de compétence similaire** pour produire des données ethnographiques homogènes.

L'analyse de consensus culturel utilise des données ethnographiques collectées au travers de questionnaires ou d'entrevues structurées. D'après Weller (2007), la manière la plus simple pour estimer la dimension culturelle des réponses est de les agréger en utilisant la moyenne pour des données quantitative et la majorité pour des données qualitatives. Des modèles analytiques plus élaborés introduits par Romney *et al.* (1986, 1987) permettent : (1) de déterminer si la variabilité observée dans les données est culturelle, (2) de mesurer la compétence culturelle que possède chaque individu et (3) de définir ce qui est culturellement valide (Weller, 2007).

1. Les valeurs propres sont des mesures indicatives de la variabilité présente dans les données de l'échantillon. Elles sont issues de l'analyse factorielle de la matrice d'accords. Cette matrice contient les accords entre chaque individu analysé. Une règle générale est qu'il y a une seule culture si le ratio pour la valeur propre la plus élevée est de 3 ou plus. Gatewood et Lowe (2008) fixent cette valeur minimum à 3.5.

16. "formalizes the insight that agreement among [individuals] is a function of the extent to which each knows the culturally defined "truth"".

2. La compétence ou fiabilité de chaque participant peut être déterminée par le degré de similarité entre les réponses d'un individu et celles des autres (Weller, 2007; Bennardo et De Munck, 2014). Un individu qui a un niveau de compétence élevé aura des réponses proches de celles culturellement acceptées.

3. La clef de réponses est le résultat final de l'analyse de consensus culturel. Elle révèle ce qui est considéré vrai, c'est-à-dire consensuellement validé. En d'autres termes, elle représente les connaissances culturelles découvertes.

2.3.8 Le processus ethnographique classique de construction des PCMs

Il existe presque autant de manières de produire des PCMs qu'il existe de combinaisons entre les choix possibles d'échantillonnage, d'explicitation et d'analyse de consensus. Par exemple, pour construire un modèle similaire à un PCM, Vuillot *et al.* (2016) utilisent un échantillonnage classique, une explicitation directe de type 3CM et une analyse de consensus par agrégation. Cependant, comme l'a remarqué Bennardo et De Munck (2014), le processus de découverte des modèles culturels composé par les ethnographes est généralement associé à trois techniques d'explicitation qui forment autant d'étapes : *free-listing*, *pile-sorting* et questionnaires.

2.3.8.1 La création d'un domaine culturel sur la base du *free-listing*

La création d'un domaine culturel est la première étape du processus traditionnel de production de PCMs. Le *free-listing* est la technique d'explicitation communément utilisée pour découvrir ce domaine (D'Andrade, 1985; Holland et Quinn, 1987; De Munck, 2013; Caulkins, 2004; Freeman *et al.*, 1981; Fryberg et Markus, 2007; Young, 1980). Cette technique consiste à demander aux participants d'expliciter les concepts liés à un domaine en écrivant une liste de mots. Les listes produites sont ensuite agrégées afin de quantifier le nombre de votes que chaque concept a obtenu. À partir des votes, il devient relativement facile de déterminer les concepts qui appartiennent au domaine culturel et ceux qui ont une dimension personnelle.

Pour mieux comprendre, voici un exemple. Imaginons que je cherche à définir le domaine culturel associé à la conception du *crime* au sein d'une force de police. Après avoir sélectionné quatre individus qui vont composer mon échantillon, je vais procéder à un *free-listing* pour chacun d'entre eux :

1. **Policier n°1** – homicide, meurtre, fuite, victime, criminel, arrestation.
2. **Policier n°2** – meurtre, délit, agression, victime, homicide, criminel, arrestation.
3. **Policier n°3** – agression, meurtre, homicide, victime, arrestation.
4. **Policier n°4** – homicide, terrorisme, meurtre, arrestation.

Ces réponses peuvent être organisées au sein d'un tableau afin que l'on puisse les quantifier.

Policier	**homi.**	**meurt.**	**fuite**	**victi.**	**crimi.**	**arrest.**	**délit**	**agress.**	**terror.**
N°1	1	1	1	1	1	1	0	0	0
N°2	1	1	0	1	1	1	1	1	0
N°3	1	1	0	1	0	1	0	1	0
N°4	1	1	0	0	0	1	0	0	1
Total	**4**	**4**	1	**3**	2	**4**	1	2	1

TABLE 2.2 – Quantification et agrégation du *free-listing* pour quatre policiers fictifs

On observe dans le tableau 2.2 que les mots *homicide*, *meurtre* et *arrestation* sont partagés par tous alors que les mots *terrorisme* et *délit* ne sont présents que pour un individu. En fixant un seuil pour ne garder que les mots ayant une majorité de votes (3 ou plus), je peux déterminer un domaine culturel qui, dans ce cas, est composé des mots : *homicide*, *meurtre*, *victime* et *arrestation*.

La technique de *free-listing* combinée à une analyse de consensus par agrégation permet d'estimer simplement un domaine culturel. Cependant, aucune information n'est recueillie à propos des relations entre les concepts de ce domaine.

2.3.8.2 La découverte des relations culturelles par *pile-sorting*

La deuxième étape pour construire un PCM consiste à identifier les relations existantes entre les concepts du domaine culturel. Le *pile-sorting* est une technique fréquemment utilisée dans l'explicitation des modèles culturels (De Munck, 2011; Maltseva et D'Andrade, 2011). Son objectif est de trouver les relations qui existent entre des objets. Lors du processus de *pile-sorting*, il est demandé aux participants de classer une liste d'objets en fonction de leur similarité. Ce tri peut se faire de manière libre ou contrainte (Roos, 1998). Dans l'approche libre, les participants classent comme bon leur semble les différents objets à leur disposition. En revanche, dans l'approche contrainte les relations qu'ils ont à évaluer leur sont imposées. En fournissant une liste de concepts à ces derniers, les similarités vont ainsi dessiner une carte cognitive reflétant la

dimension sémantique des relations. L'explicitation des relations constitutives des modèles culturels de chaque participant va permettre par analyse de consensus de découvrir des relations à propension culturelle.

Pour illustrer la découverte de relations culturelles, je vais continuer avec l'exemple des quatre policiers qui vont ranger les concepts du domaine culturel de l'exemple précédent. Les résultats finaux sont présentés sur le tableau 2.3.

	homicide	**meurtre**	**victime**	**arrestation**
homicide	-1	4	3	4
meurtre		-1	3	4
victime			-1	0
arrestation				-1

TABLE 2.3 – Exemple de l'agrégation des résultats de *pile-sorting* de quatre policiers fictifs

Le tableau 2.3 montre que les quatre policiers partagent l'idée que les mots *homicide*, *meurtre* et *arrestation* sont en relation. Une majorité pense aussi que le mot *victime* est en relation avec *homicide* et *meurtre*. En revanche, aucun ne lie *victime* avec *arrestation*.

En agrégeant les résultats des classements des concepts du domaine culturel par les membres de l'échantillon, il est possible de découvrir des relations culturelles. Cependant, ces relations sont pauvres puisqu'elles ne sont pas encore spécifiées. En effet, pour poursuivre avec l'exemple, on ne connaît pas la nature de la relation qui lie *homicide* et *meurtre*.

2.3.8.3 La finalisation du PCM sur la base de questionnaires

La troisième étape du processus de construction d'un PCM est fondée sur la compréhension des éléments culturels précédemment explicités par l'ethnographe. Ces éléments vont guider la création d'un questionnaire afin d'identifier des structures de connaissance plus fines. Les réponses des participants vont fournir des données ethnographiques parfaitement adaptées au modèle formel d'analyse de consensus culturel (Romney *et al.*, 1986) qui prend en considération des aspects particuliers des questionnaires, comme l'aptitude du participant à fournir des réponses au hasard. Ce sont ces structures consensuellement partagées qui vont mener à la construction du PCM de l'échantillon.

Les quatre policiers fictifs étant toujours disponibles, je vais faire un nouvel exemple avec eux. Suite aux résultats de découverte de relations culturelles, je décide de produire

un questionnaire pour en apprendre davantage sur ces dernières. Dans mon questionnaire très basique que je soumets aux policiers, il y a cinq propositions :

1. Les mots *homicide* et *meurtre* sont synonymes. (Vrai ou Faux)
2. Un *homicide* est un type de *meurtre*. (Vrai ou Faux)
3. Tout *meurtre* mène à une *arrestation*. (Vrai ou Faux)
4. Que ce soit un *homicide* ou un *meurtre*, tous deux impliquent une *victime*. (Vrai ou Faux)
5. Un *homicide* mène souvent à une *arrestation*. (Vrai ou Faux)

	Quest. n°1	**Quest. n°2**	**Quest. n°3**	**Quest. n°4**	**Quest. n°5**
Policier n°1	Vrai	Faux	Vrai	Vrai	Vrai
Policier n°2	Vrai	Faux	Vrai	Vrai	Vrai
Policier n°3	Faux	Faux	Vrai	Faux	Faux
Policier n°4	Vrai	Faux	Faux	Vrai	Vrai

TABLE 2.4 – Exemple de réponses offertes par quatre policiers fictifs à cinq questions qui leur sont posées

L'analyse de consensus culturel par le modèle formel de Romney *et al.* (1986) des réponses présentées sur le tableau 2.4 donne plusieurs résultats : matrice d'accords, valeurs propres, scores de compétences et clef de réponses. La matrice d'accords est consultable ci-dessous.

	Policier n°1	**Policier n°2**	**Policier n°3**	**Policier n°4**
Policier n°1	1	0.80	0.40	0.60
Policier n°2		1	0.40	0.60
Policier n°3			1	-0.20
Policier n°4				1

TABLE 2.5 – Matrice d'accords entre policiers

Cette matrice permet de calculer par analyse factorielle le ratio de la valeur propre la plus élevée. C'est cette valeur qui est indicative de la variabilité présente dans les données de l'échantillon. Dans cet exemple, le score obtenu est de 2.175. Inférieur à 3, ce score signifie que l'échantillon est composé d'un mélange d'individus appartenant certainement à différentes cultures. Les scores de compétences des individus permettent généralement de corroborer cette hypothèse.

Sur le tableau 2.6, on observe que les compétences des deux premiers policiers sont élevées. Cela révèle une cohérence entre les réponses apportées par ces derniers et celles

Policier	Score de compétences
N°1	0.935
N°2	0.935
N°3	0.304
N°4	0.576

TABLE 2.6 – Score de compétences associé à chaque policier

consensuellement acceptées par le groupe. Contrairement à eux, le policier n°3 a un score de 0.304 ce qui est indicatif de sa singularité culturelle par rapport à celle de l'échantillon. Au regard de ces nouvelles informations, il peut être recommandé de reconduire l'analyse de consensus en excluant ce genre d'individus. Pour finir, sur la base des accords et compétences de chacun, une clef de réponses est produite.

Question	Réponse consensuelle
N°1	Vrai
N°2	Faux
N°3	Vrai
N°4	Vrai
N°4	Vrai

TABLE 2.7 – Clef de réponses produite à l'issue de l'analyse de consensus culturel

La clef de réponses illustrée par le tableau 2.7 va permettre dans cet exemple de spécifier les relations culturelles préalablement obtenues. Cette dernière étape va mener à la construction d'un PCM propre au groupe culturel échantillonné.

Dans cette dernière étape, il faut retenir que le ratio de valeur propre va permettre de confirmer ou d'infirmer l'existence d'une seule culture, les scores de compétences des individus vont offrir des informations sur l'homogénéité culturelle du groupe, et la clef de réponses obtenue va guider la construction du modèle culturel prototypique final.

2.3.9 Une explicitation indirecte automatisée pour la création de PCMs

Dans cette partie j'ai commencé par définir les modèles culturels comme des modèles mentaux collectifs socialement construits et intersubjectivement partagés. Les PCMs ont été présentés comme le résultat du processus ethnographique émique de découverte de ces modèles formant des standards culturels. Ce processus est fondé sur trois techniques : l'échantillonnage qui prend en considération la dimension sociale,

l'explicitation qui extériorise les modèles mentaux des individus et l'analyse de consensus qui gère l'aspect partagé de la culture.

Il existe des variations dans les choix des ethnographes sur ces techniques qui sont conditionnés notamment par leur expérience. Si les explicitations directe et indirecte peuvent produire des résultats similaires, c'est la première qui, en pratique, est privilégiée dans la construction de PCMs. Ce choix est lié principalement à la perspective émique du processus ethnographique. En effet, cet aspect interdit toute forme d'interprétation de la part de l'ethnographe. Ainsi, l'observation ne peut être utilisée pour collecter des données. De plus, une explicitation indirecte sur les bases d'une analyse qualitative est inconcevable. La dernière possibilité est une explicitation indirecte sur les bases d'une analyse quantitative. Cependant, cette possibilité n'est pas exploitée pour de nombreuses raisons que l'on peut imaginer, parmi lesquelles :

1. **L'inexistence d'un outil** – Tout d'abord, ce type d'explicitation nécessite le développement d'un outil accessible et simple d'utilisation pour qu'un ethnographe puisse le prendre facilement en main. De plus, cet outil doit intégrer l'explicitation dans le processus ethnographique global. Pour l'instant, il n'existe aucun outil de la sorte qui soit libre d'utilisation.
2. **La quantité de données requise** – Un autre aspect est la quantité phénoménale de données nécessaires pour l'analyse statistique qui est difficilement compatible avec ce qui peut être récolté même dans une entrevue ouverte. Ainsi, seule la collecte de données existantes, notamment via le web, reste disponible, ce qui limite considérablement les champs d'application.
3. **La fiabilité de l'explicitation** – Une dernière variable à prendre en considération dans l'explicitation est sa fiabilité. La découverte de structures de connaissances à partir de données textuelles est d'une précision variable, mais évidemment bien inférieure à celle produite lors d'une explicitation directe.

Si l'utilisation d'une explicitation directe permet d'éviter certains problèmes, elle a indéniablement un coût en temps qui est incompatible avec une production à grande échelle de PCMs. Or pour qu'une ACA soit utile, elle doit justement disposer d'un grand nombre de représentations culturelles. Il faut donc que l'explicitation soit la plus automatisée possible, c'est-à-dire qu'elle soit indirecte et qu'elle inclut une analyse quantitative.

2.4 L'acquisition automatique de structures conceptuelles par la fouille de textes

Théoriquement, le développement d'une ACA émique a plus de potentiel que sa contrepartie étique. Cependant, dans la pratique, ce développement est directement conditionné par la capacité à produire des représentations culturelles émiques de manière quasi automatique. Actuellement, le problème dans la construction des PCMs se situe au niveau de l'acquisition des modèles mentaux des individus qui est principalement basée sur une explicitation directe coûteuse en temps et par conséquent incompatible avec le développement d'une ACA. L'objectif principal de cette partie est de montrer comment différentes tâches de fouille de textes peuvent permettre l'explicitation de modèles mentaux de manière similaire à celle effectuée communément dans la construction des PCMs. Dans cette optique, je commence par rapidement situer les débuts de la fouille de texte. Puis je vais détailler chaque étape du processus qui la compose. Je présente ensuite les processus automatiques fondés sur la fouille de textes qui autorisent pour la construction de PCMs la découverte de concepts importants, de relations sémantiques et de relations lexico-sémantiques. Je clos cette partie en montrant que l'objectif est atteint et en introduisant le manque de formalisme des PCMs produits.

2.4.1 Historique

La fouille de textes hérite du Processus de Découverte de Connaissances (Knowledge Discovery Process – KDP) dans les bases de données introduit par Fayyad *et al.* (1996). Le KDP est « défini comme le processus non trivial d'identification de motifs valides, nouveaux, potentiellement utiles et ultimement compréhensible dans les données [17] » (Cios *et al.*, 1998). Ce processus composé de cinq étapes est illustré sur la figure 2.3.

Contrairement à la fouille de données classique qui ne constitue qu'une étape du KDP, la fouille de textes fait référence au processus global de découverte de connaissances, ce qui est ambigu. Ce processus peut même aller au-delà si l'on se réfère à Toussaint (2011) :

« La fouille de textes est un ensemble de processus permettant, à partir d'un ensemble de ressources textuelles, de construire des connaissances pouvant être représentées dans un langage formel de représentation de connaissances et exploitées pour raisonner sur le

17. "defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data".

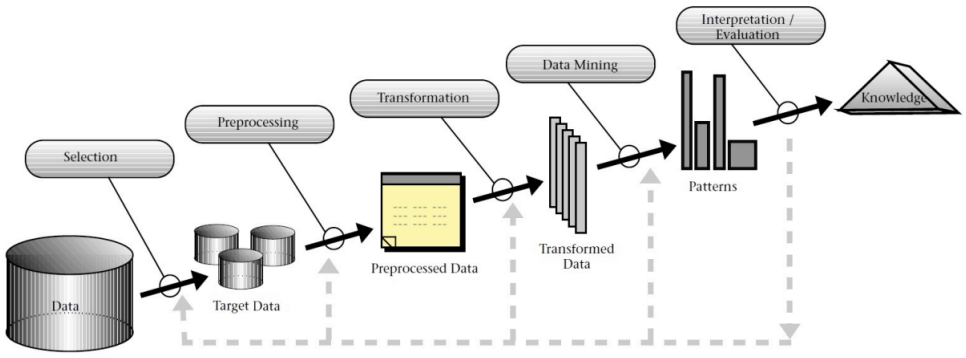

FIGURE 2.3 – KDP selon Fayyad *et al.* (1996)

contenu des textes. »

Dans cette thèse, je reste sur une conception plus usuelle de la fouille de textes qui est une adaptation du KDP pour cette source de données particulière.

2.4.2 La sélection des données textuelles

L'étape de sélection des données a pour objectif d'identifier toutes les sources de données nécessaires à la fouille. Pour la fouille de textes, cette étape se traduit par la construction d'un corpus où un ensemble de documents sont regroupés. La sélection peut nécessiter des traitements pour s'assurer de l'intérêt des textes ou des extraits de textes recueillis. L'opération la plus courante est la détection de la langue.

2.4.3 Les pré-traitements : nettoyage et enrichissement

Cette étape a pour objectif d'identifier dans les données ce qui est nécessaire à la fouille. Elle se partage entre le nettoyage des données d'une part et leur enrichissement d'autre part. Le nettoyage va chercher à réduire le bruit dans les données en éliminant celles qui sont inutiles voire nocives. L'enrichissement consiste à ajouter des informations pour réaliser des tâches de fouille plus complexes. Souvent, ces informations résultent d'une autre fouille.

Les pré-traitements pour les données textuelles sont intimement liés à la langue. Le nettoyage consiste principalement à s'assurer que les données soient effectivement textuelles. Ensuite, une pratique courante consiste à supprimer certaines données textuelles qui apparaissent en nombre mais qui sont considérées comme inutiles. Ce sont typiquement les *stopwords*.

L'enrichissement des données textuelles est réalisé par un traitement de la langue naturelle. Les opérations classiques sont la détection de phrases, la tokenisation, l'étiquetage des classes grammaticales et la lemmatisation :

1. **Détection de phrases** – Avant de pouvoir effectuer des opérations sur du texte il s'agit en premier lieu de délimiter les contours des phrases. Voici un exemple :

```
Les forces de police sont opérationnelles.
Elles se sont déployées ce matin.
```

2. **Tokenisation** – Cette tâche a pour but d'identifier les unités lexicales qui composent du texte : les mots, les signes de ponctuation, etc. Pour illustrer la tokenisation :

```
Les forces de police sont opérationnelles .
```

3. **Étiquetage des classes grammaticales** – Cet étiquetage a pour objectif d'associer une classe grammaticale à chaque token. En continuant avec l'exemple précédent, le tableau 2.8 montre le résultat de l'étiquetage :

Token	**Classe grammaticale**
Les	DT – déterminant
forces	NNS – nom commun pluriel
de	DT – déterminant
police	NN – nom commun
sont	VB – verbe
opérationnelles	JJ – adjetcif
.	PUNC – ponctuation

TABLE 2.8 – Étiquetage des classes grammaticales des tokens

4. **Lemmatisation** – Différents tokens peuvent représenter un même concept. La majorité des mots ne sont pas invariables et possèdent des déclinaisons, c'est le cas par exemple de *policier* et *policière*. Afin de pouvoir regrouper des mots qui expriment un concept similaire, il existe deux techniques appelées respectivement la racinisation et la lemmatisation. Les deux sont dépendantes de la langue. Néanmoins, la première nécessite moins de connaissances linguistiques que la deuxième. La racinisation est une troncature basée le plus souvent sur une expression régulière avec une gestion très légère des exceptions. Ainsi, les racines de *policier* et *policière* seront certainement identiques : *polic*. Le problème de

cette technique, c'est que d'une part elle ne fonctionne pas toujours et d'autre part elle va regrouper des mots avec des sens différents. Le mot *police* sera ainsi associé à *policier*, *policière* et *policer*.

La lemmatisation morphologique se base sur la classe grammaticale des mots pour offrir une gestion plus fine de leur racine appelée un « lemme ». Par exemple, le lemme des verbes est l'infinitif. Ainsi les mots *policer*, *policé* et *policèrent* vont tous être associés à *policer*. Cependant, cette lemmatisation requiert l'identification au préalable de la classe grammaticale de chaque token. Le tableau 2.9 ci-dessous illustre la lemmatisation de l'exemple précédent.

Token	**Classe grammaticale**	**Lemme**
Les	DT	Le
forces	NNS	force
de	DT	de
police	NN	police
sont	VB	être
opérationnelles	JJ	opérationnel
.	PUNC	.

TABLE 2.9 – Lemmatisation des tokens trouvés sur la base des classes grammaticales

Les opérations de pré-traitements mises bout à bout sont assez coûteuses en temps. C'est pourquoi toutes les informations produites au travers des différents pré-traitements sont généralement stockées. Une bonne pratique consiste à gérer les informations linguistiques sous la forme d'annotations afin de pouvoir les réutiliser facilement.

2.4.4 La transformation des données

La transformation a pour objectif de préparer les données pour la fouille. Bien que cette étape ne soit pas obligatoire pour la fouille qui parfois ne requiert pas de transformation particulière des données pré-traitées, elle se traduit généralement pour la fouille de textes en la quantification d'annotations et/ou de leurs attributs.

Par exemple, imaginons que je cherche à fouiller mon corpus pour découvrir les actions les plus courantes. Je vais d'abord devoir identifier, lors du pré-traitement, les formes verbales qui représentent les actions dans un texte. Ensuite je vais réduire toutes mes annotations de verbes en un unique tableau qui va répertorier le nombre de fois où chaque annotation apparaît. C'est ce tableau que je vais pouvoir fouiller pour connaître les fréquences des verbes et ainsi apprendre quelles sont les actions les plus courantes. Concrètement, si je dispose du fil d'actualité suivant :

[Le 20 juillet]	Les policiers ont arrêté des voleurs de diamants.
[Le 22 juillet]	Un homme suspecté de vendre de la drogue a été arrêté.
[Le 30 juillet]	Un contrôle de police a permis de prévenir un attentat.
[Le 2 août]	Un policier blessé lors d'une intervention.

Je vais pouvoir aisément transformer les annotations pour constituer un tableau (voir ci-dessous la référence 2.10) sur lequel je pourrai appliquer divers algorithmes.

Lemme issu d'annotation	Nombre d'occurrences
arrêter	2
vendre	1
permettre	1
prévenir	1
blesser	1

TABLE 2.10 – Quantification des lemmes d'annotations des formes verbales

2.4.5 La fouille

Cette tâche concerne l'utilisation de tout algorithme capable de découvrir des informations utiles. Dans la majorité des cas ce sont des algorithmes d'Apprentissage Automatique (*Machine Learning* – ML). Le ML est un champs d'étude de l'AI héritant des études sur la reconnaissance de motifs et sur les théories d'apprentissage computationnel. Son objectif est de permettre à un processus d'évoluer (de s'améliorer) de manière systémique par le biais d'algorithmes qui apprennent à partir d'exemples. Une illustration du processus de ML est offerte ci-dessous :

Un processus de ML est constitué grossièrement de deux phases :

1. **Gestion des caractéristiques de classification** – L'objectif de cette phase est de préparer des exemples pertinents pour la tâche d'apprentissage. À partir d'un jeu de données, des caractéristiques sont tout d'abord produites, puis celles jugées intéressantes sont sélectionnées. Les trois premières étapes du KDP permettent cette gestion des caractéristiques de classification.

2. **Apprentissage** – Prenant un ensemble d'exemples en entrée, un algorithme est utilisé pour entraîner un modèle statistique à les classer. Chaque classe est identifiée par un label. Ce modèle appelé classificateur (ou *classifier*) peut ensuite être utilisé pour prédire l'appartenance d'un exemple quelconque à une classe en

lui assignant un label.

La structure des exemples fournis au programme permet de définir deux grandes familles d'apprentissage :

1. **Supervisé** – Désigne des méthodes dont l'apprentissage est guidé en définissant explicitement l'objet de la recherche. Cela se traduit au niveau des exemples par des couples *(entrées, résultats)*.
2. **Non supervisé (ou *Clustering*)** – Fait référence à des méthodes capables de regrouper entre elles des ensembles de données sans aucune autre information.

2.4.6 L'évaluation

Les résultats obtenus lors de la fouille sont évalués à l'aide d'un *gold standard*. Un *gold standard* est un jeu de données correctes, validées si possible par une communauté d'experts, à partir desquelles les résultats produits peuvent être comparés.

2.4.6.1 Les méthodes

Une bonne pratique est de diviser les données en deux ensembles, un pour l'entraînement et un pour les tests, appelé *hold-out*. Cette séparation permet d'éviter d'apprendre sur les données dont la prédiction va être évaluée. Le plus souvent, la répartition des données réserve 70 % pour l'apprentissage et 30 % pour les tests. L'utilisation d'un ensemble de données *hold-out* permet d'obtenir une validation plus fiable. Néanmoins, le hasard peut fausser considérablement les résultats de la validation. En effet, la partie des données dédiée à l'apprentissage peut être relativement simple contrairement à celles pour le test ou vice versa.

Afin d'apporter un début de solution à ce problème, la méthode proposée consiste à construire aléatoirement k ensembles de données d'apprentissage et de *hold-out*, et à faire la moyenne des k évaluations possibles. Cette méthode est appelée « test de *hold-out* répété ». Le problème de cette méthode est qu'elle crée une disparité dans l'utilisation des données (par exemple, certaines vont être plus souvent réservées pour l'apprentissage que d'autres).

La solution à ce problème est la validation croisée à k-pli (ou *k-fold cross-validation*). Cette méthode divise le jeu de données en k parties disjointes et égales. Ensuite, par roulement, l'entraînement du modèle statistique s'effectue sur *k-1* parties et le test sur celle restante. Le résultat final est une moyenne des k évaluations. Cette méthode permet

de garantir que chaque exemple de données est utilisé exactement autant de fois pour l'apprentissage et uniquement une fois lors des tests.

2.4.6.2 Les mesures de précision, rappel et F1-score

De nombreuses mesures basées sur l'évaluation permettent de déterminer la qualité de la tâche de fouille. Pour que la tâche soit considérée comme bonne, il faut que la qualité se rapproche de celle qu'auraient obtenu des experts. Les mesures usuelles sont la précision et le rappel.

Le rappel est défini par le nombre d'exemples identifiés correctement comme appartenant à une classe c au regard du nombre d'exemples connus comme appartenant à c :

$$Rappel = \frac{Corrects_c}{Connus_c}$$

La précision est définie par le nombre d'exemples identifiés correctement comme appartenant à une classe c rapporté au nombre total d'exemples identifiés comme appartenant à c :

$$Précision = \frac{Corrects_c}{Identifiés_c}$$

Le F1-score est une moyenne harmonique de la précision et du rappel dont la formule est :

$$F1 - score = \frac{2 * Précision * Rappel}{Précision + Rappel}$$

2.4.7 Les trois tâches de fouille liées à la découverte des PCMs

Il faut rappeler que la construction de PCMs est fondée sur trois techniques d'explicitation qui constituent des tâches à part entière que sont : le *free-listing*, le *pile-sorting* et le questionnaire. L'objectif du *free-listing* est d'expliciter les concepts importants, celui du *pile-sorting* de trouver les relations à partir d'une liste de concepts et le questionnaire de pouvoir connaître la nature des relations. Pour automatiser le processus de construction de PCMs, il faut que ces objectifs puissent être réalisés par trois tâches de fouille de données. Ces tâches impliquent (1) d'identifier lors du pré-traitement les éléments d'intérêt, (2) de les transformer en une forme compatible avec la fouille et (3) de les fouiller afin d'atteindre l'objectif fixé.

2.4.7.1 Trouver les concepts importants

Dans toutes les conceptualisations, les concepts fondamentaux sont des groupes nominaux. Repérer les groupes nominaux dans un texte peut se faire de plusieurs manières plus ou moins coûteuses en temps. La plus simple, rapide mais aussi la moins précise est d'utiliser une expression régulière basée sur les classes grammaticales des tokens. La manière la plus performante, complexe et lente est basée sur de l'apprentissage supervisé du contexte dans lesquels les groupes nominaux apparaissent. La décomposition des phrases en arbres syntaxiques est actuellement la meilleure manière de produire ce contexte.

Pour pouvoir juger de l'importance des groupes nominaux présents dans un corpus, il faut les quantifier. L'idée est que leur nombre d'occurrences est indicatif d'une certaine importance. Les mesures les plus courantes sont :

1. **TF (*Term Frequency*)** – Cette mesure de base permet d'indiquer la proportion que représente un groupe nominal gn par rapport au nombre total :

$$TF(gn) = \frac{Occurrences(gn)}{NombreOccurrences}$$

2. **TF/IDF (*Term Frequency/Inverse Document Frequency*)** – Cette mesure inclut une gestion de la distribution du groupe nominal dans les documents. L'IDF permet d'identifier la rareté d'un groupe nominal :

$$IDF(gn) = \log\left(\frac{NombreDocuments}{Documents(gn)}\right)$$

$$TF/IDF(gn) = \frac{TF(gn)}{IDF(gn)}$$

La fouille des concepts importants se fait généralement sur la base d'un seuil. Ce seuil peut s'appliquer sur la mesure elle-même ou sur le nombre de groupes nominaux à conserver.

2.4.7.2 Déterminer les relations sémantiques

La découverte de relations sémantiques nécessite de travailler sur les relations entre groupes nominaux.

Cette découverte est basée sur l'hypothèse distributionnelle harrissienne (Harris, 1954) qui suppose que la proximité des mots et celle de leur sens sont corrélées. Ainsi, deux mots qui apparaissent souvent dans un contexte similaire ont de fortes chances d'être liés sémantiquement. En s'appuyant sur l'occurrence et la co-occurrence des mots, il est

alors possible de calculer des mesures d'interdépendance et de similarité traduisant leurs proximités. Différentes formules permettent de calculer cette proximité :

1. **Jaccard** – Le coefficient de Jaccard entre deux groupes nominaux gn_1 et gn_2 est calculé en divisant leur nombre de co-occurrences par leur nombre respectif d'occurrences :

$$Jaccard(gn_1, gn_2) = \frac{O(gn_1 \cap gn_2)}{O(gn_1 \cup gn_2)}$$

2. **Overlap** – L'expression de cette mesure permet de minimiser les effets liés à la comparaison de deux entités aux tailles trop différentes :

$$Overlap(gn_1, gn_2) = \frac{P(gn_1 \cap gn_2)}{min(P(gn_1), P(gn_2))}$$

3. ***Pointwise Mutual Information* (PMI)** – Cette mesure quantifie la dépendance statistique de deux entités :

$$PMI(gn_1, gn_2) = \log(\frac{P(gn_1 \cap gn_2)}{P(gn_1) * P(gn_2)})$$

Relativement récemment, les algorithmes de construction de modèles d'espace vectoriel (*Vector Space Models* – VSMs) à partir de données textuelles ont prouvé qu'il était possible de construire des vecteurs sémantiques de haute qualité (Pantel et Lin, 2002; Rapp, 2003). Les modèles les plus connus sont *Hyperspace Analogue to Language* (HAL) (Lund et Burgess, 1996), *Latent Semantic Analysis* (LSA) (Landauer et Dumais, 1997) et plus récemment *Word2Vec* (Mikolov *et al.*, 2013).

Comme pour trouver les concepts importants, la sélection des relations sémantiques se fait généralement sur la base d'un seuil.

2.4.7.3 Découvrir des relations lexico-sémantiques

Hearst (1992) développe une méthode qui permet d'exploiter des motifs syntaxiques (ou patrons syntaxiques) afin d'extraire des relations lexico-sémantiques de type hyperonyme. Elle remarque que dans certaines phrases sont présentes des séquences de mots qui permettent d'identifier des relations d'hyperonymie telles que présentées ci-dessous :

`Un voleur est un criminel .`

`Les bâtiments tels que les stations de police et centres de loisir .`

Hearst (1992) se base sur cette évidence pour lister un ensemble de motifs récurrents qui vont permettre le filtrage automatique d'hyperonymes comme illustré ci-dessous :

```
{GN} tel(s|le|les) que {((GN, )* (GN (et|ou))*) GN}
Exemple : les bâtiments tels que les stations de police et centres de loisir.
Extraction : (bâtiments, stations de police), (bâtiments, centres de loisir)
---
{(GN, )* GN} (et|ou) (l'|les|une|des)? autre(s)? {GN}
Exemple : les violences, meurtres et autres crimes.
Extraction : (crimes, violences), (crimes, meurtres)
---
{GN} inclu[a-z]* {((GN, )* (GN (et|ou))*) GN}
Exemple : les drogues incluent l'héroïne et le cannabis.
Extraction : (drogues, héroïne), (drogues, cannabis)
```

Sa méthode a été réutilisée et étendue par d'autres chercheurs, notamment sur les relations lexico-sémantiques que sont la méronymie et la causalité (Girju *et al.*, 2002, 2003, 2006; Caraballo, 2001; Cederberg et Widdows, 2003; Pantel et Ravichandran, 2004).

La transformation des informations recueillies avec la méthode d'Hearst (1992) dans une forme exploitable pour la fouille est basée sur les triplets que composent les éléments mis en relation et leur patron syntaxique *(e_1, patron, e_2)*. Des mesures statistiques de proximité lexico-sémantique peuvent être produites de manière classique ou en traitant les occurrences des couples *(e_1, patron)* et *(patron, e_2)* tel que présenté dans Costa *et al.* (2011).

L'extraction de relations lexico-sémantiques sur la base de patrons syntaxiques est généralement associée à une précision relativement basse. Pour les relations d'hyperonymie, Cederberg et Widdows (2003) ont rapporté 40 % de précision, Maynard *et al.* (2009) 48.5 % et Hearst (1998) 52 %. Dans une expérience que j'ai réalisée (Petit *et al.*, 2017) cette précision est descendue jusque 30 %. Ces variations sur la base d'une méthode similaire s'expliquent d'après Cederberg et Widdows (2003) par les différences de qualité des données textuelles traitées. En effet, Hearst (1998) utilise l'encyclopédie de Grolier, Maynard *et al.* (2009) le contenu de Wikipédia, eux-mêmes le British National Corpus et mes tests étaient effectués sur des données issues de sites web (Petit *et al.*, 2017).

La qualité de la fouille des relations lexico-sémantiques peut être améliorée en adoptant une démarche d'apprentissage. Cet apprentissage se fait par (1) la sélection de caractéristiques de classification, (2) le choix d'un algorithme d'apprentissage supervisé

et (3) d'un ensemble de relations labélisées.

2.4.7.3.1 La sélection des caractéristiques de classification

Il existe une très large gamme de caractéristiques utilisables pour l'apprentissage. Dans Hendrickx *et al.* (2009), pas moins de 16 types de caractéristiques sont utilisés. Pour n'en citer que quelques uns : classe grammaticales, racines, n-grams, chemins de dépendance, motifs syntaxiques et mesures de proximité sémantique.

2.4.7.3.2 Les algorithmes d'apprentissage supervisé

Pour ce qui est de l'apprentissage supervisé de relations lexico-sémantiques, de nombreux algorithmes différents ont déjà été testés. Si l'algorithme le plus populaire est celui de Machine à Vecteurs de Supports (*Support Vector Machine* – SVM), une bonne pratique consiste à en tester plusieurs et à garder celui qui obtient les meilleures performances (Hendrickx *et al.*, 2007; Davidov et Rappoport, 2008). Très brièvement, les algorithmes principaux sont :

1. **SVM** – C'est un algorithme supervisé de classification linéaire, binaire et non-probabilistique. Un modèle SVM est une représentation des exemples de données comme points transposés dans un espace où sont séparés par le plus grand écart possible les deux catégories. La classe de nouveaux exemples est assignée en fonction du côté dans lequel ces derniers apparaissent.
2. **Naïf bayésien** – C'est un classificateur statistique appliquant le théorème de Bayes qui considère les caractéristiques de classification comme indépendantes. L'avantage de cet algorithme est qu'il demande peu de données pour l'apprentissage.
3. **Arbre de décision** – Cet algorithme est basé sur une structure graphe en forme d'arbre inversé. L'arbre est en général construit en séparant l'ensemble des données en sous-ensembles en fonction de la valeur d'une caractéristique d'entrée. Ce processus est répété de façon récursive sur chaque sous-ensemble.
4. **Réseau de neurones** – C'est un modèle mathématique qui est inspiré des aspects structurels et fonctionnels des réseaux de neurones biologiques. Un réseau de neurones est composé d'un groupe de neurones artificiels interconnectés. L'information transite selon une approche connexionniste de calcul. Dans la plupart des cas, les réseaux de neurones sont des systèmes adaptatifs, leur

structure étant modifiée avec le flux d'information utilisé lors de la phase d'apprentissage. Ils sont composés au minimum d'une couche de neurones en entrée, dont le nombre est déterminé par celui des caractéristiques de classification à traiter, et d'une couche de neurones en sortie, avec une taille fixée par le nombre de classes à apprendre. Entre ces deux couches, il peut aussi y avoir d'autres couches dites cachées.

2.4.7.3.3 Les jeux de données de relations labélisées

Les *gold standards* ont un rôle double pour l'apprentissage de relations lexico-sémantiques. En effet, ils permettent à la fois d'évaluer la qualité de la tâche d'apprentissage, mais aussi de produire les données labellisées nécessaires à l'apprentissage. Parce que les *gold standards* dépendent du corpus, il est fréquent que l'expert les crée lui-même en fonction de ses besoins.

Des *gold standards* globaux sont disponibles de manière plus ou moins directe :

1. **WordNet** (Miller, 1995) – C'est une base de données lexicale produite et maintenue par l'université de Princeton. Elle est destinée à l'anglais et est composée de noms, verbes, adjectifs et adverbes, reliés surtout par des relations lexicales d'hyperonymie et de méronymie. Similaire à WordNet, il existe WOLF (Wordnet Libre du Français) pour le français (Sagot et Fišer, 2012), mais il est bien moins complet.
2. **FrameNet** (Baker *et al.*, 1998) – Ce projet de l'université de Berkeley a pour objectif de proposer une ressource lexicale en anglais lisible à la fois par les humains et les machines. Cette ressource est basée sur la théorie des trames sémantiques.
3. **ConceptNet** (Liu et Singh, 2004) – C'est un large graphe sémantique librement accessible. Il a été conçu pour permettre aux machines de comprendre les mots utilisés communément par les humains.
4. **DBpedia** (Auer *et al.*, 2007) – Cette base de connaissances est issue d'un effort pour structurer les informations contenues dans Wikipédia afin de les rendre librement accessibles sur le web.
5. **SemEval-2010 Task #8** (Hendrickx *et al.*, 2009) – Le SemEval (pour évaluation sémantique) est une série continue d'évaluations de systèmes d'analyse sémantique computationnelle. En 2010, lors d'un défi sur une tâche de classification de relations sémantiques entre paires de groupes nominaux pour l'anglais, ils ont produit un

gold standard de 10 717 relations de 10 classes : « cause-effet », « instrument-agent », « produit-producteur », « contenu-conteneur », « entité-origine », « entité-destination », « composant-tout », « membre-collection », « message-sujet » et « aucune ».

6. **BLESS (*Baroni and Lenci Evaluation of Semantic Spaces*)** (Baroni et Lenci, 2011) – Cet ensemble de données est aussi pour l'anglais. Il est composé de 200 noms simples en entrée auxquels sont adjoints plusieurs relations lexicales : co-hyponymie, hyperonymie, méronymie, attribution et événement.

2.4.8 Un compromis entre fiabilité et rapidité pour la construction de PCMs

Pour rappel, une ACA est basée sur les capacités de création, de formalisation et de médiation de représentations culturelles. Dans la partie précédente, j'ai présenté les PCMs comme des représentations culturelles adaptées aux exigences de granularité qui font défaut dans celles, étiques, actuellement utilisées dans les CASs. Afin de proposer une solution à la problématique temporelle liée à la construction de ces représentations émiques, j'ai expliqué comment la fouille de textes peut être appliquée à l'explicitation de modèles mentaux.

La production de PCMs par un processus ethnographique automatique est accompagnée d'un problème qui jusqu'à présent était mineur, celui de la qualité des modèles construits. Si l'explicitation dans le processus classique avait une précision humaine, cette précision va maintenant dépendre principalement de la tâche d'apprentissage des relations lexico-sémantiques. Ainsi, les PCMs ne devront plus seulement être évalués au regard de la dimension culturelle mais aussi par rapport à leur fiabilité.

Finalement, parce qu'ils sont construits par un processus informatique, les PCMs sont accessibles dans un format entre l'informel et le semi-formel. Pour qu'une ACA puisse prendre conscience des cultures représentées, il est nécessaire que ces derniers soient disponibles dans un langage formel interprétable par une machine.

2.5 Les ontologies

Si les représentations culturelles sur des bases étiques ou émiques permettent de décrire des cultures, elles ne sont pas directement exploitables pour développer une ACA. En effet, pour qu'une AI puisse interpréter ces connaissances culturelles et donc atteindre une forme de conscience, il est nécessaire que ces dernières soient formalisées dans une forme lisible par une machine.

Dans l'effort engagé pour le développement de systèmes culturellement conscients, les ontologies « sont devenues le médium de représentation de choix[18] » (Mohammed et Mohan, 2013a) puisqu'elles permettent la spécification formelle de conceptualisations partagées sur le domaine culturel (Borst, 1997). La contribution la plus remarquable est certainement celle de Blanchard et de ses collègues avec l'ontologie supérieure de la culture (*Upper Ontology of Culture* – UOC) (Blanchard *et al.*, 2010). D'autres exemples dans le domaine des CATSs sont relatifs à la gestion de la culture au niveau de l'apprenant. C'est le cas des ontologies CAE-L (Chandramouli *et al.*, 2008) et *MultiCulltural Aspects Ontology* (Motz *et al.*, 2005).

L'objectif de cette partie est de montrer comment transformer un PCM en une ontologie culturelle. Afin d'y parvenir, j'introduis la brève histoire des ontologies. Je me concentre sur la dimension formelle des ontologies, en particulier les langages formels et les syntaxes. Je présente ensuite un langage permettant de manipuler les ontologies. Enfin, je décris un processus automatique de formalisation d'ontologies applicable aux PCMs.

2.5.1 Historique

Le mot « ontologie » vient du domaine de la philosophie et qualifie « un système particulier de catégories décrivant une vision particulière du monde[19] » (Guarino, 1998). En AI, le mot « ontologie » a été défini pour la première fois par Gruber (1995) pour qui elle est « une spécification explicite d'une conceptualisation[20] ». Pour rappel, une conceptualisation est l'abstraction d'un phénomène par ses concepts principaux et ses relations importantes (Studer *et al.*, 1998). Une ontologie constitue sa représentation au travers d'un vocabulaire non-ambigu.

18. "have become the representation medium of choice".
19. "a particular system of categories accounting for a certain vision of the world".
20. "an explicit specification of a conceptualization".

Deux ans plus tard, Borst (1997) a ajouté à la définition de Gruber (1995) deux contraintes :

— la spécification doit être formelle ;

— la conceptualisation doit être partagée.

Les ontologies varient d'une formalisation fortement informelle (exprimée simplement dans un langage naturel) à une rigoureusement formelle (définie de manière méticuleuse avec une sémantique formelle). La dimension formelle des ontologies est importante puisqu'elle sous-tend leur inter-opérabilité, ré-utilisabilité et surtout interprétabilité par des machines.

Les ontologies peuvent représenter des conceptualisations allant de l'individuel au collectif. Cependant, produire des ontologies individuelles manquait d'intérêt pratique, si bien qu'on a systématiquement conçu des ontologies à partager.

2.5.2 Les différentes sortes d'ontologies

Les ontologies se distinguent selon l'objectif de leur conceptualisation. D'après Gòmez-Pérez et Benjamins (1999), il existe ainsi différentes ontologies :

1. de représentation de connaissances (Van Heijst *et al.*, 1997) ;
2. générales (Guarino, 1998) ;
3. de haut niveau (Guarino, 1997) ;
4. génériques, noyaux ou métas (Van Heijst *et al.*, 1997) ;
5. de domaine (Mizoguchi *et al.*, 1995; Van Heijst *et al.*, 1997) ;
6. linguistiques (Bateman *et al.*, 1995) ;
7. de tâche (Mizoguchi *et al.*, 1995) ;
8. de méthode (Chandrasekaran *et al.*, 1999) ;
9. d'application (Van Heijst *et al.*, 1997).

2.5.3 Les éléments constitutifs des ontologies

Une ontologie est généralement représentée sous la forme d'un graphe composé d'individus, de classes, de relations et d'axiomes. Les individus ou instances définissent des objets. Les classes sont l'équivalent des concepts. Ces classes prennent tout leur sens en établissant avec d'autres classes des relations dont la logique associée est assurée par des règles et axiomes.

2.5.4 Les langages formels

Il existe de nombreux langages qui permettent de formaliser des ontologies afin que des machines puissent interpréter des conceptualisations. Deux genres de langages formels sont applicables à la formalisation d'ontologies, ceux à destination des graphes et ceux dédiés aux ontologies.

2.5.4.1 Les langages formels pour les graphes

Bien que la formalisation d'ontologies avec des langages graphes ne permet pas le même niveau d'expressivité qu'avec des langages dédiés, elle permet dans une perspective d'ingénierie de capitaliser sur des décennies de recherche sur les graphes et réseaux ainsi que sur l'ensemble des outils qui ont été développés.

Il y a de nombreux formats pour la représentation de graphe tels que GEXF, GDF, GML, GraphML, Pajek NET, GraphViz DOT, CSV, UCINET DL, Tulip TPL, Netdraw VNA ou Spreadsheet. Ces formats diffèrent dans leur capacité à gérer des aspects plus ou moins complexes des graphes. Une comparaison de ces formats est réalisée sur le site de Gephi [21] qui est le principal logiciel de visualisation et d'exploration de graphes. Cette comparaison qui s'appuie sur ce qui est géré par Gephi est présentée sur la figure 2.4.

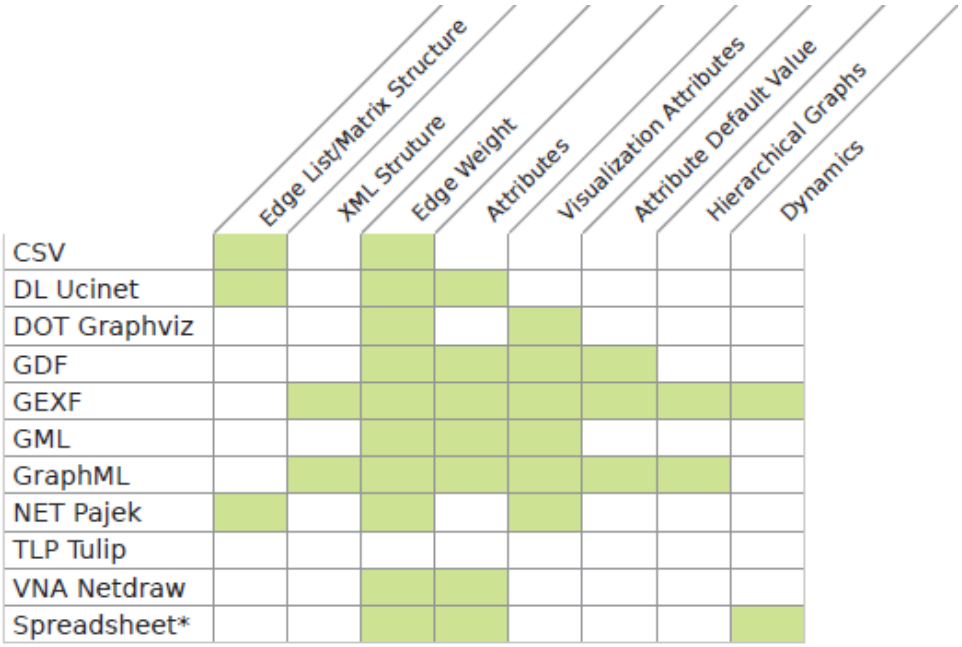

FIGURE 2.4 – Comparaison des formats utilisables pour des graphes. Pour la source, se référer à Gephi : https://gephi.org/users/supported-graph-formats/

21. https://gephi.org

D'après ce tableau comparatif, c'est le format GEXF (*Graph Exchange XML Format*) qui est le plus complet et donc le mieux adapté pour la représentation de graphes.

Un fichier GEXF est un document XML (*eXtensible Markup Language*) composé à la racine d'une balise *gexf* dans laquelle est présente une balise *meta* qui contient les métadonnées associées au graphe tel que son créateur, ainsi qu'une balise *graph* qui permet de définir l'ensemble des noeuds et arêtes du graphe. Ci-dessous est représentée une ontologie décrivant une partie du domaine de la police dans un format de fichier GEXF :

```
<?xml version="1.0" encoding="UTF-8"?>
<gexf xmlns="http://www.gexf.net/1.3" version="1.3"
    xmlns:viz="http://www.gexf.net/1.3/viz"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.gexf.net/1.3
    http://www.gexf.net/1.3/gexf.xsd">
  <meta lastmodifieddate="2017-04-18"><creator>Jean Petit</creator></meta>
  <graph defaultedgetype="directed" mode="static">
    <nodes>
      <node id="0" label="personne"/>
      <node id="1" label="policier"/>
      <node id="2" label="détective"/>
      <node id="3" label="criminel"/>
      <node id="4" label="crime"/>
      <node id="5" label="vol"/>
      <node id="6" label="meurtre"/>
      <node id="7" label="fraude"/>
      <node id="8" label="victime"/>
    </nodes>
    <edges>
      <edge id="0" source="0" target="1" label="estHyperonymeDe"/>
      <edge id="1" source="0" target="2" label="estHyperonymeDe"/>
      <edge id="2" source="0" target="3" label="estMéronymeDe"/>
      <edge id="3" source="3" target="4" label="estHyperonymeDe"/>
      <edge id="4" source="4" target="5" label="estHyperonymeDe"/>
      <edge id="5" source="4" target="6" label="estHyperonymeDe"/>
      <edge id="6" source="4" target="7" label="estHyperonymeDe"/>
```

```
    <edge id="7" source="4" target="8" label="estMéronymeDe"/>
   </edges>
  </graph>
</gexf>
```

Une illustration graphique via Gephi est offerte via la figure 2.5.

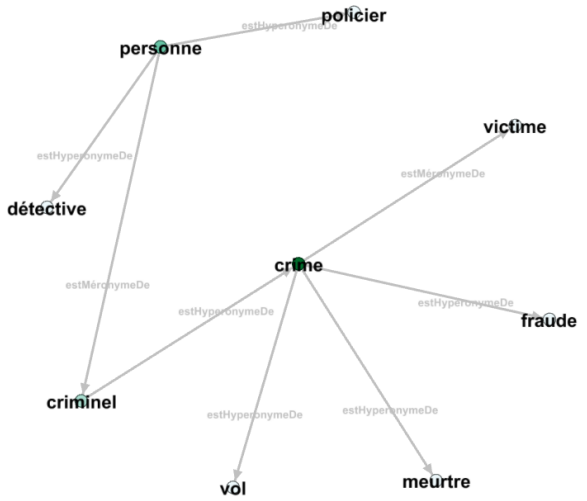

FIGURE 2.5 – Visualisation du fichier GEXF via Gephi

2.5.4.2 Les langages formels dédiés aux ontologies

Créé en 1999, le *Resource Description Framework* (RDF) est le langage formel à l'origine du développement du web sémantique dont l'idée est de qualifier, de formaliser et de structurer les savoirs présents sur le web. Il permet en effet de décrire des métadonnées au travers d'un langage basé sur des entités composées de triplets *(ressource, propriété, valeur)* sous la forme *(sujet, prédicat, objet)*. Les ressources font référence aux concepts/classes identifiées avec une URI (*Uniform Resource Identifier*). Les propriétés peuvent être des attributs ou toutes autres relations préalablement définies. Les valeurs sont des littéraux pointant soit vers un symbole (chaîne de texte, nombre, etc.) soit vers une autre ressource. Un ensemble de triplets est nommé un graphe RDF. Un exemple de graphe RDF est présenté sur la figure 2.6.

Sur la figure 2.6 trois triplets sont représentés. La ressource `http://www.exemple.com/bouclier` est liée par la propriété `http://www.exemple.com/label` à la valeur « Bouclier ». On peut aussi observer le triplet (`http://www.exemple.com/bouclier`, `http://www.exemple.com/`

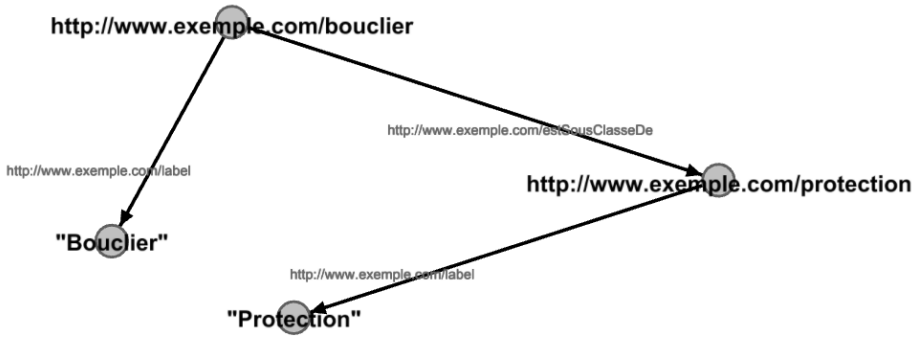

FIGURE 2.6 – Visualisation d'un graphe RDF via Gephi

estSousClasseDe, http://www.exemple.com/protection) qui indique une relation de sous-classe entre les deux ressources, le pointage d'une valeur vers une ressource étant le phénomène de réification.

RDF permet de représenter des déclarations de propriétés sur des ressources mais il est limité sur l'expression des connaissances sur les propriétés ou sur les types de ressources. Le RDF schema (RDFs) résout ce problème en permettant de définir un schéma qui autorise la qualification des relations décrites en RDF. Si les langages RDF et RDFs offrent une meilleure structuration du savoir, ils sont encore trop limités pour formaliser à eux seuls des ontologies. Ils manquent notamment d'une logique développée et de capacités de raisonnement.

En réponse à l'apparition de nombreux langages de définition et de manipulation d'ontologies (Chaudhri *et al.*, 1998; Genesereth *et al.*, 1992), le W3C (*World Wide Web Consortium*) a initié le développement du Langage Ontologique Web (*Web Ontology Language* – OWL) devenu un standard en 2004. Fondé sur les langages RDF et RDFs, OWL bénéficie d'une plus grande logique avec l'ajout de notions entre les classes telles que l'identité, la symétrie, la transitivité, etc. De plus, OWL est disponible sous trois langages d'expressivité croissante : OWL Lite, OWL DL et OWL Full. Ci-dessous une représentation OWL de ce qui était décrit dans l'exemple précédent :

```
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
   xmlns:owl="http://www.w3.org/2002/07/owl#"
   xmlns:dc="http://purl.org/dc/elements/1.1/">
```

```
   <owl:Ontology rdf:about="http://www.exemple.com/">
      <dc:title>Équipement.</dc:title>
      <dc:description>Les équipements portés par les
          policiers.</dc:description>
   </owl:Ontology>

   <owl:Class rdf:ID="protection">
      <rdfs:label>Protection</rdfs:label>
   </owl:Class>

   <owl:Class rdf:ID="bouclier">
      <rdfs:label>Bouclier</rdfs:label>
      <rdfs:subClassOf rdf:resource="#protection"/>
   </owl:Class>
</rdf:RDF>
```

2.5.4.3 La syntaxe des langages ontologiques formels

Les langages RDF, RDFs et OWL possèdent plusieurs syntaxes dont les plus connues sont RDF/XML et TURTLE. Jusqu'à présent, la syntaxe des exemples étaient en RDF/XML. L'avantage de cette syntaxe est qu'elle bénéficie du format éprouvé et familier XML, et de nombreux outils disponibles tels que les parseurs. La contrepartie est que ce formalisme est lourd. La syntaxe TURTLE offre une syntaxe beaucoup plus légère et lisible, centrée sur les triplets. L'exemple précédent devient en TURTLE :

```
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix dc11: <http://purl.org/dc/elements/1.1/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<http://www.exemple.com/>
  a owl:Ontology ;
  dc11:title "Équipement." ;
  dc11:description "Les équipements portés par les policiers." .

<http://www.exemple.com/#protection>
  a owl:Class ;
```

```
  rdfs:label "Protection" .

<http://www.exemple.com/#bouclier>
  a owl:Class ;
  rdfs:label "Bouclier" ;
  rdfs:subClassOf <http://www.exemple.com/#protection>
```

2.5.4.4 L'interprétation des ontologies

Les ontologies écrites en langages formels sont accessibles par des API et des langages de requête tels que SPARQL (*SPARQL Protocol and RDF Query Language*). SPARQL est un langage similaire au SQL (*Structured Query Language*) qui est utilisé pour communiquer avec les bases de données (BDDs) relationnelles classiques, avec une syntaxe TURTLE. Ce langage permet d'effectuer des requêtes sur des *triplestores* qui sont des bases de données constituées de triplets.

Les clauses pour les résultats peuvent être de deux types :

1. **Interrogatives** – Elles sont identifiées par l'instruction *SELECT* et permettent de consulter les connaissances présentes dans des ontologies. La structure d'une requête *SELECT* est la suivante :

```
# Déclaration des préfixes pour abbrévier les URIs et gagner en
    lisibilité
PREFIX ...
# Définition des ensembles de données à consulter
FROM ...
# Clause pour les résultats
SELECT ...
# Spécification de la requête
WHERE ...
# Modification sur les résultats
ORDER BY ...
```

   Par exemple, je vais pouvoir trouver tous les couples *(?a : ressources, ?b : labels)* avec la requête ci-dessous :

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
```

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
SELECT ?a ?b
WHERE {?a rdfs:label ?b}
```

2. **Constructives** – Elles sont caractérisées par les instructions *INSERT DATA* ou *DELETE DATA* et regroupent tout ce qui a trait à la création, modification et suppression d'ontologies. La structure des requêtes est la suivante :

```
# Déclaration des préfixes pour abbrévier les URIs et gagner en
    lisibilité
PREFIX ...
# Définition des ensembles de données àconsulter
FROM ...
# Clause pour les résultats
(INSERT | DELETE) DATA
{
#IRI du graphe
GRAPH ... {
    #les triplets
    ...
  }
}
```

Par exemple si je souhaite ajouter une classe « casque » à mon ontologie sur les équipements des forces de police, je peux faire :

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX ex: <http://exemple.com/>
INSERT { GRAPH <http://www.exemple.com>
   {<http://www.exemple.com/#casque> a owl:Class }}
```

2.5.5 L'ontologisation des PCMs en ontologies culturelles

Dans cette partie j'ai présenté en détail les ontologies qui sont définies par Studer *et al.* (1998) comme des spécifications formelles de conceptualisations partagées. Selon ses critères, au niveau de la représentation, seule la spécification formelle différencie une ontologie d'un PCM puisque ce dernier est socialement construit et consensuellement partagé. Ainsi, la transposition d'un PCM en une ontologie dite *culturelle* (en écho à l'objectif de la conceptualisation) se fait par la formalisation de ses structures de connaissances.

Le passage de connaissances explicitées en une forme interprétable par une machine peut se faire de trois manières différentes :

1. **Manuelle** – Construire et maintenir des ontologies manuellement est de moins en moins pratiqué car c'est à la fois coûteux en temps mais aussi en ressources humaines. En effet, en plus d'être laborieux, formaliser à la main des connaissances requiert une collaboration entre des experts d'une part et des ingénieurs formés pour cette tâche d'autre part. Toutefois, cette approche a pour avantage d'offrir un haut niveau de maîtrise sur les connaissances formalisées.
2. **Automatique** – Il est possible de traduire automatiquement en langage formel un ensemble de connaissances informelles par *ontologisation* (Pennacchiotti et Pantel, 2006). Pour ce faire, on doit connaître au préalable le format des connaissances en entrée et la conversion que l'on cherche à effectuer. Des règles d'alignement peuvent ainsi être établies. L'avantage de cette méthode est que le coût en temps est limité puisqu'il n'est attribué qu'à la conception des règles. L'inconvénient est son manque de flexibilité.
3. **Assistée** – De nombreux outils ont été conçus pour faciliter la création et la gestion d'ontologies tels que le populaire Protégé [22] qui permet de simplifier drastiquement ces tâches. En plus du gain de temps qu'apporte l'assistance de ces outils, ils permettent parfois aux experts de se passer complètement des ingénieurs en connaissances, bien que cela soit aussi susceptible de produire des erreurs.

Dans l'optique de contribuer au développement d'une ACA, il faut que le système soit capable de formaliser lui-même les PCMs qu'il est amené à construire. C'est

22. Disponible à `http://protege.stanford.edu/`

pourquoi, parmi les trois manières de formaliser des PCMs en ontologies, seule l'approche automatique est envisageable. Dans la partie précédente qui concernait la fouille de textes, la découverte des relations lexico-sémantiques nécessitait *a minima* une connaissance de celles recherchées. La définition de règles fixes d'alignement pour l'ontologisation des PCMs en ontologies culturelles peut alors s'effectuer sur la base des relations lexico-sémantiques précédemment définies.

Après avoir traité de la création des PCMs et de leur formalisation en ontologies culturelles, il s'agit maintenant de s'intéresser aux moyens existants permettant la production de médiations entre ces dernières.

2.6 La médiation d'ontologies

Les PCMs formalisés en ontologies culturelles permettent à une ACA de prendre conscience de diverses cultures. Contrairement à une ACA conçue autour d'un modèle étique, une ACA émique est caractérisée par des représentations culturelles hétérogènes. C'est ce contexte qui invite à repenser les modalités de l'élaboration de médiations. L'objectif majeur de cette partie est de comprendre comment les techniques d'alignement d'ontologies peuvent être mises au service de la découverte de médiations d'ordre culturel entre des ontologies qui sont elles aussi culturelles. Je commence par introduire succinctement la gestion de l'hétérogénéité conceptuelle, que ce soit par les humains ou les machines. Je décris ensuite les deux phénomènes qui peuvent causer ces différences conceptuelles dans les ontologies. Après la description des diverses équivalences possibles, je fais un rapide tour d'horizon des techniques de comparaison qui les génèrent, puis je détaille le processus général de leur production. Finalement, j'explique comment des équivalences produites par alignement d'ontologies peuvent s'apparenter dans cette thèse à des médiations culturelles pour une ACA émique.

2.6.1 Historique

La gestion au niveau métier de l'hétérogénéité des conceptions du monde est lointaine puisque l'ensemble des diplomates a dû y être confronté pour produire des médiations adéquates en fonction des points de vue des personnes qu'ils représentaient et avec lesquelles ils devaient dialoguer. Plus récemment dans les domaines relatifs à la connaissance, des personnes adoptent un rôle d'intermédiaire pour permettre l'harmonisation de différentes perspectives. D'après Lomas (1997), ces personnes sont des *knowledge mediator* (Aschoff *et al.*, 2004; Osborne, 2004) ou *knowledge broker* (Sverrisson, 2001; Meyer, 2010). Entre autres, le *knowledge mediator* « aide à la construction d'ontologies et à la découverte de consensus[23] » (Aschoff *et al.*, 2004).

La problématique de l'hétérogénéité des représentations conceptuelles existe au moins depuis les années 1990, notamment avec les modélisations des bases de données relationnelles. Les tentatives d'intégration de BDD distribuées ont marqué le commencement des réflexions d'unification de schémas structurellement et sémantiquement hétérogènes (Batini *et al.*, 1986; Ahmed *et al.*, 1991; Kim *et al.*, 1993;

23. "supports ontology construction and consensus finding".

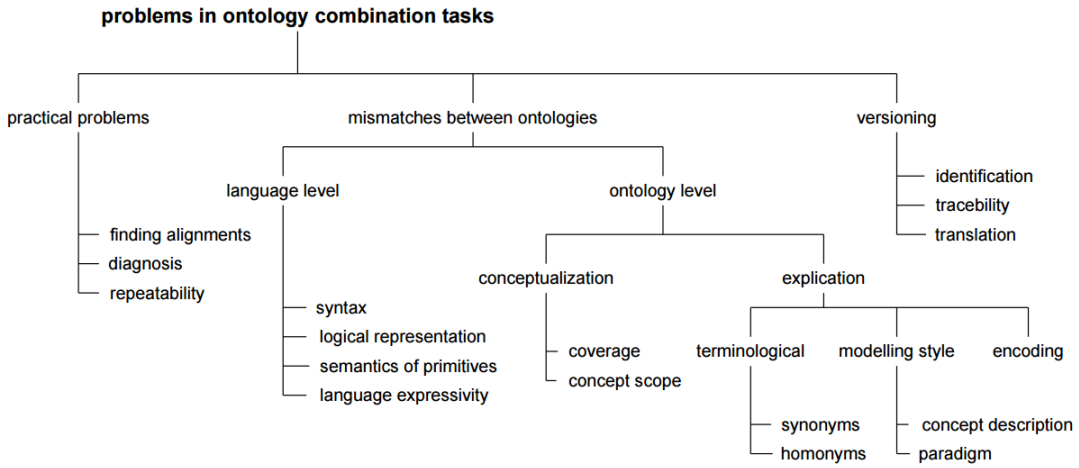

FIGURE 2.7 – Les problèmes qui apparaissent dans le processus de combinaison d'ontologies. Pour la source, se référer à Klein (2001).

Hammer et McLeod, 1993; Litwin *et al.*, 1990; Thomas *et al.*, 1990; Garcia-Molina *et al.*, 1997).

La médiation d'ontologies cherche à apporter des solutions à l'hétérogénéité causée par les disparités liées aux différences de conceptualisation dans les ontologies. Les recherches dans ce domaine héritent de celles initiées pour les schémas de BDD. Les ontologies se différencient de ces derniers par un plus haut niveau d'expressivité (Bartalos et Bielikova, 2007; Kalyanpur *et al.*, 2004). L'automatisation de la mise en correspondance d'ontologies correspond à l'alignement d'ontologies dont le but est de lier de manière cohérente plusieurs ontologies dans un accord mutuel (Noy *et al.*, 2000). Dans l'alignement d'ontologies, les ontologies elles-mêmes ne sont pas modifiées. À l'inverse, la fusion d'ontologies a pour objectif de capturer toute la connaissance de diverses ontologies locales pour former une nouvelle ontologie unique.

2.6.2 Situer les disparités conceptuelles entre ontologies

Klein (2001) fournit une classification des problèmes qui apparaissent lors de l'utilisation combinée de plusieurs ontologies. La taxonomie qu'il propose est basée principalement sur le travail de Visser *et al.* (1998). Elle est présentée sur la figure 2.7.

Klein (2001) distingue deux types de problèmes liés à la conceptualisation des ontologies. Le premier provient de la couverture des ontologies. Des ontologies peuvent représenter un même phénomène avec différentes granularités. Par exemple, différentes ontologies sur le domaine de la culture ont de fortes chances d'aborder des parties qui,

bien qu'elles se recoupent, différent largement.

Le deuxième problème identifié par Klein (2001) est la portée des concepts. Un même concept peut avoir plusieurs représentations. L'exemple donné est celui d'une classe *employé* dont la conception est amenée à varier au sein de diverses administrations (Wiederhold, 1994). Appelé « inadéquation de classes » dans des travaux antérieurs (Pepijn *et al.*, 1997), ce problème porte à la fois sur les concepts et les relations (qui peuvent constituer elles-mêmes des concepts).

2.6.3 Les équivalences entre ontologies

La mise en correspondance d'ontologies a pour objectif de relier sur la base de relations logiques des concepts ou relations similaires provenant de différentes ontologies sources (Calvanese *et al.*, 2001; Su et Gulla, 2004). En prenant deux ontologies O_1 et O_2, une correspondance est généralement représenté par un 5-tuple (i, e_1, e_2, r, s) où i est un identifiant unique, e_1, e_2 sont les éléments mis en correspondance, avec $e_1 \in O_1$ et $e_2 \in O_2$, r est la relation logique et s un score de confiance pour cette relation.

D'après les travaux sur les schémas de BDD effectués par Batini *et al.* (1986), ces équivalences peuvent être :

1. **Identiques** – Les éléments e_1 et e_2 sont exactement les mêmes. Ces équivalences sont produites quand les perceptions sont identiques et qu'aucune incohérence ne s'immisce dans la spécification.
2. **Équivalentes** – e_1 et e_2 sont similaires. Cette similarité peut être déterminée par des *comportements* proches, par exemple si une tâche requêtant l'un ou l'autre concept produit des résultats identiques. La similarité peut aussi être définie en observant que les concepts partagent les mêmes instances. Des opérations logiques peuvent aussi être appliquées pour chercher une équivalence.
3. **Compatibles** – e_1 et e_2 ne sont ni identiques, ni équivalents. Cependant, aucune contre-indication du point de vue des conceptualisations n'empêche l'équivalence.
4. **Incompatibles** – Les différences de cohérence entre les éléments e_1 et e_2 sont clairement identifiées.

2.6.4 Les techniques de comparaison d'ontologies

La comparaison d'ontologies est la tâche spécifique de recherche de correspondances entre des entités appartenant à deux ontologies. Les techniques de comparaison sont

généralement divisées entre syntaxique et sémantique. « Dans la comparaison syntaxique, les labels et parfois la structure syntaxique des ontologies sont comparées et on obtient typiquement un coefficient de similarité [0, 1], qui indique la similarité entre deux concepts. La comparaison sémantique combine un ensemble défini de relations entre les concepts, prenant en compte le sens de chaque concept [24]. » (De Bruijn *et al.*, 2006)

Une classification plus détaillée des techniques de comparaison est proposée par Shvaiko et Euzenat (2013) et représentée sur la figure 2.8.

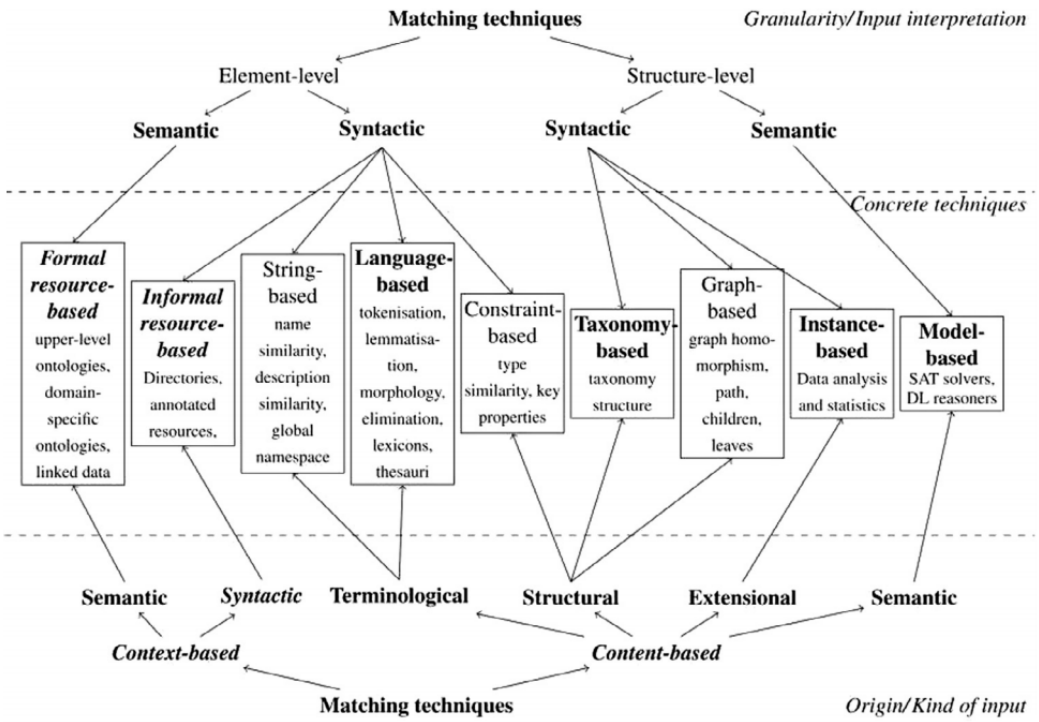

FIGURE 2.8 – Classification des techniques de comparaison récentes. Pour la source, se référer à Shvaiko et Euzenat (2013).

2.6.5 Le processus de production de correspondances entre ontologies

La production de correspondances entre ontologies est un processus divisé en deux parties que sont la découverte/évaluation et la représentation/stockage (De Bruijn *et al.*, 2006; Granitzer *et al.*, 2010). Dans la première partie, les alignements sont découverts en

24. "In syntactic matching, the labels and sometimes the syntactical structure of the [ontologies are] matched and typically some similarity coefficient [0, 1] is obtained, which indicates the similarity between two concepts. Semantic matching computes a set-based relation between the [concepts], taking into account the meaning of each [concept]."

comparant et en évaluant les éléments (concepts ou relations) appartenant à plusieurs ontologies. Les techniques de comparaison permettent de trouver automatiquement des alignements potentiels en leur associant un score de confiance. Ces alignements sont ensuite examinés manuellement par un expert pour être validés.

Dans la deuxième partie, les alignements sont formellement représentés pour être stockés. L'ensemble des alignements peut constituer une ontologie intermédiaire appelée « ontologie d'articulation » (Kalfoglou et Schorlemmer, 2003). Maedche *et al.* (2002) pour le MAFRA (*MApping FRAmework for distributed ontologies*) nomme cette dernière l'« Ontologie de Pontage Sémantique[25] ». Une fois stockées, les articulations peuvent être utilisées pour fusionner les ontologies ou les aligner (Su et Gulla, 2006).

2.6.6 La production de médiations culturelles

Dans cette partie, j'ai présenté la médiation d'ontologies qui aborde les différences conceptuelles dans une perspective d'alignement ou de fusion d'ontologies. Ces différences sont dues à un problème soit de couverture des ontologies soit de portée de leurs concepts. La construction de médiations entre ontologies se traduit par l'utilisation de diverses techniques de comparaison permettant la découverte d'équivalences. Le processus complet intègre aussi leur validation, leur représentation et leur stockage.

Les PCMs possèdent un biais culturel d'ordre conceptuel intrinsèque formé par l'organisation de leurs concepts et de leurs relations lexico-sémantiques. C'est cette organisation qui reflète un modèle culturel socialement construit et partagé. La formalisation des PCMs en ontologies n'apporte aucune altération. Deux ontologies culturelles à propos d'un même domaine sont amenées à diverger uniquement sur leur conceptualisation. Ainsi, en supposant que les domaines dans chaque ontologie soient correctement couverts, les différences de conceptualisation restantes ne peuvent avoir qu'une origine culturelle. Dans ces conditions, les indications d'équivalence produites par les techniques de comparaison d'ontologies ont une dimension culturelle. Parmi les techniques présentées, quatre en particulier sont adaptées car basées sur la structure des ontologies. Ce sont celles sur : les ressources formelles, les taxonomies, les graphes et les modèles logiques.

25. "Semantic Bridging Ontology".

2.7 Conclusion

Dans ce chapitre, j'ai présenté une solution au problème de granularité des représentations culturelles issues de modèles étiques utilisés dans la modélisation d'ACA. Cette solution est fondée sur les PCMs qui sont des représentations émiques à même de représenter la complexité des cultures. J'ai décrit l'approche ethnographique manuelle communément utilisée pour produire ces représentations et montré qu'elle ne convenait pas aux besoins techniques d'automatisme pour le développement d'une ACA. Le problème étant situé dans l'explicitation des modèles mentaux individuels, la fouille de structures conceptuelles s'impose comme l'unique solution pour répondre à cette contrainte. La compatibilité des PCMs avec les ontologies permet de les formaliser directement et automatiquement par ontologisation. Ce sont les ontologies culturelles qui vont servir de biais aux machines pour acquérir une forme de conscience des cultures représentées. Enfin, c'est la compréhension des structures conceptuelles des ontologies qui permet d'imaginer leur comparaison. Les diverses techniques d'alignement d'ontologies constituent alors une solution concrète pour faire émerger des médiations culturelles.

Ce chapitre conclut la partie *recherche* de ma thèse. La prochaine partie aborde l'aspect technique du *développement* de ma solution.

Chapitre 3

Développement

3.1 Introduction

Les recherches sur la modélisation d'ACA n'ont pas vocation à rester théoriques. La production d'une ACA émique requiert un large investissement en matière de développement. Une ACA émique est composée globalement de trois modules que je me suis efforcé de développer :

— un qui gère la construction des PCMs ;

— un autre qui s'occupe de leur formalisation en ontologies ;

— un dernier qui autorise leur médiation.

Dans le détail, la partie sur la création des PCMs est de loin la plus complexe puisqu'elle suit un processus ethnographique auquel se greffe des processus de fouille de textes. Techniquement, afin de traiter de gros volumes de données hétérogènes en un temps raisonnable, toute la fouille a été conçue dans une perspective *Big Data* avec l'ensemble des algorithmes parallélisés à l'aide d'Apache Spark[1]. Il faut rappeler que l'automatisme des tâches constitue un défi technique majeur.

Ma contribution s'est limitée à la langue anglaise. Cette décision a été prise pour simplifier au maximum le développement de l'ACA, dont la complexité va de paire avec la diversité des disciplines sollicitées et des compétences requises associées.

1. http://spark.apache.org/

3.2 Un outil de création de PCMs

La construction d'une ACA émique requiert le développement d'un processus permettant l'acquisition de PCMs détaillés en un temps raisonnable, l'idéal étant que ce soit fait de manière complètement automatique. Il faut bien comprendre que ces restrictions techniques conditionnent fondamentalement la réalisation pratique de ce nouveau genre d'ACA. L'objectif de cette partie est de présenter les différentes parties de l'outil développé pour permettre la construction quasi automatique des PCMs. Tout d'abord, je commence par introduire la gestion des échantillons et des individus qui les composent. Ensuite, je détaille le processus d'explicitation indirecte des modèles mentaux que j'ai mis en place, allant de la collection des données de l'individu à la fouille de concepts, de relations sémantiques et de relations lexico-sémantiques. Finalement, j'explique la production des PCMs à partir des éléments précédemment explicités.

3.2.1 La gestion des individus et des échantillons

L'outil que j'ai développé n'intègre pas la partie ethnographique de l'échantillonnage. Cette dernière est laissée à la charge d'une personne assumant le rôle d'ethnographe. L'identification d'une communauté est réalisée selon les critères sociaux qui lui semble pertinents. Quant à la réduction de l'échantillon, elle dépend aussi de ses choix. Néanmoins, l'outil permet de prendre en compte ses décisions en autorisant une gestion d'individus et d'échantillons avec la possibilité d'ajouter, modifier ou supprimer.

Pour des raisons à la fois éthique et pratique, les individus considérés ne sont pas des personnes en tant que telles mais des organisations, collectifs, groupes, etc. En effet, il n'était pas envisageable de demander des autorisations pour exploiter les données d'utilisateurs par rapport au temps alloué. De plus, la quantité nécessaire de données à récolter pour l'étape d'explicitation des modèles mentaux est incompatible avec ce qu'il était possible de collecter à l'échelle d'un individu.

3.2.2 La collection de données issues du web et de Facebook

La collecte des données doit permettre la récolte d'une importante quantité, et ce, de manière automatique. C'est pourquoi la collecte est conçue pour chercher les données directement sur le web. En effet, le web est devenu ces dernières années la plus grande

source de données culturelles à portée de main des ethnographes. D'ailleurs, les réseaux sociaux tels que Facebook ou Twitter permettent via des APIs de récolter facilement des données ethnographiques. Ce choix s'ensuit d'une contrainte majeure lors de l'échantillonnage puisque seules les communautés présentes sur le web deviennent analysables. Bien qu'un grand nombre de communautés soient exclues, ce n'est pas un problème puisque ces mêmes communautés ne sont certainement pas concernées par l'utilisation de systèmes culturellement conscients.

Concrètement, l'outil permet de collecter des données issues de sites web ainsi que de Facebook. HTTRACK [2] permet de copier le contenu des sites en naviguant et en enregistrant les fichiers auxquels il accède. L'API graphe de Facebook [3] a été utilisée via Facebook4J [4] pour pouvoir récupérer le contenu des posts présents sur les *murs* disponibles à tous.

3.2.3 Une gestion des données orchestrée par GATE

GATE [5] (General Architecture for Text Engineering) est un environnement complet pour gérer des données textuelles. Il permet d'intégrer intelligemment des briques de traitement allant du découpage de textes en phrases jusqu'à la reconnaissance d'entités nommées (villes, personnes, etc). GATE permet aussi d'ajouter, de retirer et de manipuler les différentes informations relatives aux données textuelles. C'est pour sa capacité à gérer diverses annotations et à capitaliser dessus pour effectuer d'autres traitements que GATE a été choisi.

3.2.4 La sélection des données textuelles

Après avoir récolté les données issues du web, la collection de documents obtenus est hétérogène. Si l'API graphe de Facebook fournit des données textuelles brutes, HTTRACK télécharge des documents variés. Bien qu'il puisse être configuré pour filtrer et ne garder que les fichiers textes, au final il y a presque toujours des images ou encore des vidéos qui sont aussi enregistrées. Apache Tika [6] a été utilisé pour assurer la création d'un corpus de documents uniquement textuels.

2. http://www.httrack.com/
3. https://developers.facebook.com/docs/graph-api
4. http://facebook4j.github.io/en/index.html
5. https://gate.ac.uk/
6. https://tika.apache.org/

Cet outil est capable de détecter et d'extraire automatiquement les métadonnées ainsi que le texte de plus d'un millier de formats de fichier différents (.doc, .odt, etc). Les métadonnées ont été utilisées pour garder seulement les documents textuels. Puis le texte brut de ces derniers a été extrait pour peupler le corpus.

3.2.5 La détection des phrases

Le choix de travailler au niveau des phrases plutôt que des documents a été guidé par plusieurs raisons, classées ci-dessous par ordre d'importance.

1. **Redondance** – Les documents provenant du web sont généralement différents les uns des autres mais possèdent néanmoins souvent du contenu similaire généré automatiquement. Sur des sites web cela peut être un menu ou encore un flux connecté à un réseau social. Sur un réseau social tel que Facebook, des posts peuvent aussi être générés systématiquement par des robots où la même information est reprise à de multiples reprises. En gérant des phrases, il est donc plus aisé de réduire considérablement le contenu dupliqué.
2. **Traitements parallèles** – Travailler sur des phrases est mieux adapté au traitement parallèle puisqu'elles sont plus simples à distribuer.
3. **Conformité** – Il est plus simple de s'assurer de la validité textuelle de phrases par rapport à celle de documents. Cela permet *in fine* de réduire considérablement le bruit.

Apache OpenNLP [7] a été utilisé pour détecter les phrases. Cette boîte à outils basée sur des algorithmes d'apprentissage permet le traitement de diverses langues naturelles dont l'anglais. Elle offre aussi une API avec une implémentation très simple pour la détection de phrases.

3.2.6 La gestion de l'aspect multilingue

Le web n'ayant pas de frontière, les langues souvent se rencontrent et se mélangent. Un processus de fouille de textes étant généralement conçu pour une langue particulière, il est important d'être à même d'identifier et de filtrer les langues *étrangères*. Après la détection de la langue de chaque phrase, seules celles anglaises ont été conservées.

La gestion de l'hétérogénéité des langues a été confiée à l'API LangDetect (Shuyo, 2010). LangDetect est basé sur un ensemble de profils de langue construits à partir de

7. https://opennlp.apache.org/

Wikipédia. Il utilise un algorithme naïf bayésien pour identifier 53 langages avec plus de 99 % de précision.

3.2.7 La correction des fautes de frappe

Les fautes de frappe sont courantes sur le web, en particulier sur les réseaux sociaux. Leur gestion est extrêmement difficile. C'est pourquoi une prise en compte minimum a été mise en place.

L'API Jazzy [8] développée pour la vérification de l'orthographe de mots anglais a été sélectionnée. J'ai essayé de maximiser la pertinence des suggestions de l'API. Pour y parvenir, j'ai calculé la distance de Levenshtein entre le mot initial et les suggestions obtenues. Cette distance permet de traduire en un score la différence entre les caractères de deux propositions. Finalement, seules les suggestions avec une distance de 1 sont retenues car la probabilité que la suggestion soit correcte est assez importante.

3.2.8 La gestion des anglais

Les langues évoluent, si bien que le français de France et celui du Québec ont de nombreuses différences notamment dans l'orthographe de certains mots. Il en est de même pour l'anglais UK (United Kingdom) et US (United States). Afin de ne pas considérer des mots tels que *offen**c**e* et *offen**s**e* comme différents, il est bénéfique en matière d'analyse de ne traiter qu'un type de langue.

Stanford CoreNLP [9] (Manning *et al.*, 2014) est une boîte à outils permettant de faire de nombreux traitements comme de la reconnaissance d'entités nommés ou de l'analyse de sentiments. Stanford CoreNLP possède notamment un *package* avec une fonction *americanize*. Elle permet de transposer les mots en anglais UK vers US.

3.2.9 L'étiquetage des informations linguistiques basiques

La majorité des tâches de fouille de données textuelles requièrent un minimum de traitements sur la langue naturelle analysée. Plusieurs outils sont disponibles pour la langue anglaise tels que OpenNLP et Stanford CoreNLP que j'ai déjà présentés, mais aussi TreeTagger [10] (Schmid, 2013). TreeTagger est uniquement un outil pour annoter

8. http://jazzy.sourceforge.net/
9. https://stanfordnlp.github.io/CoreNLP/
10. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

les mots dans un texte avec leur forme grammaticale et le lemme associé. Bien que Stanford CoreNLP soit plus *lourd* que ses concurrents, il est meilleur en matière de rapidité, de fiabilité et de performance. C'est pour ces raisons que l'annotation des éléments linguistiques basiques (tokens, classes grammaticales, lemmes) lui a été dédiée.

3.2.10 La fouille des concepts importants

L'identification des groupes nominaux dans l'outil est réalisée à partir d'une expression régulière implémentée en JAPE (Java Annotation Patterns Engine [11]). Propre à GATE, ce format de fichier permet de bénéficier des annotations produites afin de réaliser des recherches complexes de motifs dans du texte. J'ai opté pour une détection très simple des groupes nominaux car le résultat convient mieux pour la construction des patrons syntaxiques dont j'ai besoin après. Ainsi, un groupe nominal est décrit comme pouvant éventuellement commencer par un déterminant ou un pronom personnel, suivi d'une succession de potentiels adjectifs et se terminant par une suite d'au moins un nom. Ci-dessous l'expression en JAPE :

```
({Token.pos =~ "DT|PRP\$"})?
({Token.pos =~ "JJ[A-Z]*"})*
({Token.pos =~ "N[A-Z]*"})+
```

Après leur identification, les groupes nominaux peuvent être quantifiés. Cette quantification agrège les groupes nominaux sur les lemmes associés. Les groupes nominaux sont classés selon leur fréquence qui est l'indice que j'utilise pour déterminer leur importance.

Un seuil choisi par l'ethnographe permet d'exprimer le nombre de concepts importants à conserver.

3.2.11 La fouille des relations sémantiques

La fouille de relations sémantiques ne demande pas d'identification de structures particulières. Elle repose à la fois sur des mesures classiques de co-occurrence (Jaccard et PMI) et sur deux algorithmes de VSM (HAL (Lund et Burgess, 1996), Word2Vec (Mikolov *et al.*, 2013)).

L'implémentation de HAL, quant à elle, provient du kit S-Space [12]. S-Space fournit

11. https://gate.ac.uk/sale/tao/splitch8.html
12. https://github.com/fozziethebeat/S-Space/wiki

une collection d'algorithmes pour construire des espaces sémantiques et pour concevoir ces algorithmes. Word2Vec a été intégré avec la librairie MLlib. C'est la librairie d'apprentissage automatique *scalable* de Spark [13].

Pour chaque algorithme, la réduction de la dimension a été fixée à 400, valeur intermédiaire entre la taille standard de 300 et celle élevée aux alentours de 500. Les groupes nominaux avec un nombre d'occurrences inférieur à 3 ont été retirés ainsi que les phrases constituées de moins de 3 groupes nominaux parmi ceux restants.

Toutes les co-occurrences ont été construites avec une fenêtre de taille 6.

Déterminer les relations les plus sémantiques d'un concept se fait au travers d'un classement qui comprend l'ensemble des mesures. La mieux classée est donc celle qui obtient les meilleurs scores pour les mesures Jaccard, PMI, HAL et Word2Vec. Un seuil déterminé encore une fois par l'ethnographe permet de filtrer les meilleures relations pour chacun des concepts analysés.

3.2.12 La fouille des relations lexico-sémantiques

Pour la fouille de relations lexico-sémantiques, je me suis limité à un type de relation lexico-sémantique qui est l'hyperonymie.

La fouille de relations lexico-sémantiques est fondée sur la mise en place d'un processus de ML composé de trois parties : construction des caractéristiques de classification, choix de l'algorithme d'apprentissage supervisé et définition d'un *gold standard*.

L'intégration de l'identification de relations lexico-sémantiques s'est faite de manière similaire à Wang *et al.* (2006). Des patrons syntaxiques sont formalisés dans le format particulier JAPE et sont appliqués sur le corpus. La liste des patrons pour l'extraction d'hyperonymes est présentée ci-après.

```
({Token.lemma == "such"} ((GN)):e1 | ((GN)):e1 ({Token.lemma == "such"})?) as
    (((GN))+):e2
Exemple : crimes such as theft

(((GN))+):e2 ({Token.lemma == "and"} | {Token.lemma == "or"}) {Token.lemma ==
    "other"} ((GN)):e1
Exemple : murder or other crimes

(((GN))+):e1 {Token.lemma == "include"} ((GN)):e2
```

13. https://spark.apache.org/docs/1.1.0/mllib-guide.html

```
agencies include police
Exemple : agencies include police and health

((GN)):e2 {Token.lemma == "be"} ({Token.pos == "DT"})? {Token.lemma =~
    "(form|sort|type|kind)"} {Token.lemma == "of"})? (((GN))+):e1
Exemple : a knife is a weapon

((GN)):e1 {Token.lemma == "like", Token.pos == "JJ"} (((GN))+):e2
Exemple : people like victims and criminals
```

Les relations lexico-sémantiques découvertes grâce aux motifs syntaxiques sont utilisées pour produire des mesures de proximité indicatives de classes sémantiques calculées avec les formules Jaccard et PMI. Ainsi, chaque label de motif va être associée à deux mesures. Ces nouvelles mesures vont s'ajouter aux mesures de similarité produites lors de la fouille de relations sémantiques afin de constituer les caractéristiques de classification pour l'identification des relations lexico-sémantiques.

J'ai fait le choix de tester plusieurs algorithmes afin de garder celui qui offrait les meilleurs résultats. Ces tests ont été réalisés avec le logiciel RapidMiner [14]. Il permet en quelques clics de construire une chaîne de traitement de ML sous la forme d'un flux de données connectées à des briques appelées *opérateurs*. Il offre accès à de nombreux algorithmes d'apprentissage supervisé facilement configurable ainsi qu'à des opérateurs capables d'évaluer les résultats produits. Les algorithmes que j'ai sélectionnés sont nommés :

1. **Decision Tree** – Un algorithme d'arbre de décision traditionnel.
2. **Naive Bayes** – Un classificateur naïf bayésien normal.
3. **SVM** – L'implémentation du SVM fait par la librairie LibSVM [15] (Fan *et al.*, 2005).
4. **Neural Net** – Un réseau de neurones dirigé vers l'avant entraîné par un algorithme de rétropropagation.

La construction de *gold standards* équilibrés a demandé quelques efforts puisqu'ils ont dû être faits à la main. Il n'était pas envisageable d'utiliser des *gold standards* existants n'ayant aucune traçabilité quant à leur communauté d'origine pour un travail mené sur

14. https://rapidminer.com/
15. https://www.csie.ntu.edu.tw/~cjlin/libsvm/

des domaines culturels particuliers. Ainsi chaque domaine culturel a conduit à la création d'un *gold standard* sur mesure. Pour faciliter cette tâche laborieuse, je me suis aidé des relations candidates trouvées dans les corpus des membres d'un échantillon. Au final, un *gold standard* est composé de deux classes (*hypernym* et *none*) comprenant autant d'exemples.

3.2.13 La création du domaine culturel

La découverte d'un domaine culturel se fait par agrégation comme dans le processus classique de construction de PCM. L'ethnographe choisit deux seuils :

— le premier indique pour chaque individu le nombre de concepts importants à extraire ;

— le deuxième spécifie le niveau de consensus requis pour considérer qu'un concept partagé est culturel.

3.2.14 La production des relations sémantiques culturelles

Déterminer la dimension culturelle de relations sémantiques se fait là encore par agrégation. L'ethnographe décide de deux seuils :

— le premier indique pour chaque individu le nombre de relations sémantiques à conserver pour chaque concept du domaine culturel ;

— le deuxième spécifie le niveau de consensus requis pour considérer qu'une relation sémantique partagée est culturelle.

3.2.15 La découverte des relations lexico-sémantiques culturelles

La construction et la gestion des caractéristiques de classification pour les relations dont on souhaite prédire la classe lexico-sémantique sont des tâches chronophages. Hormis la phase d'apprentissage qui se fait sur un nombre de relations relativement limité, il est difficilement envisageable d'adopter une approche combinatoire lorsque l'on va chercher à identifier les relations lexico-sémantiques associées aux concepts du domaine culturel. En d'autres termes, pour un domaine culturel constitué de 2 000 concepts, il n'est pas souhaitable d'identifier la classe des 40 000 000 de relations possibles pour chaque individu. Pour limiter le nombre de relations à analyser, il est

possible d'utiliser uniquement les relations sémantiques culturelles produites, mais leur nombre reste toujours trop important. Afin de limiter encore plus les relations candidates aux prédictions, je filtre les relations culturelles avec les indices que constituent les résultats des extractions des relations lexico-sémantiques. Ainsi, pour qu'une relation soit sujette à prédiction pour l'ensemble des individus ainsi qu'à l'analyse de consensus :

1. ses concepts doivent faire partie du domaine culturel ;
2. sa nature sémantique doit être partagée ;
3. elle doit avoir été identifiée par au moins un individu dans l'échantillon.

Pour l'analyse de consensus avec le modèle formel, j'ai choisi UCINET (Borgatti *et al.*, 1999) parmi d'autres outils (Borgatti, 1992; Oravecz *et al.*, 2014). Ce choix a été guidé pour des raisons pratiques puisque les versions disponibles des autres outils ne fonctionnaient pas sur ma machine. Il peut prendre en entrée des profils qui sont normalement composés de lignes correspondant aux réponses des participants à un questionnaire. Dans mon cas, les réponses correspondent aux résultats des prédictions des relations à identifier. À partir de ces profils, l'outil génère automatiquement la matrice d'accords, les valeurs propres, les compétences et la clef de réponses.

C'est la découverte de la clef de réponses qui marque la complétion de l'outil de création de PCMs. Afin de construire une forme d'ACA, il reste encore à développer deux parties : la formalisation des PCMs et leur médiation.

3.3 Des mécanismes d'ontologisation à destination des PCMs

La formalisation des PCMs est une étape importante dans la construction d'une ACA puisqu'elle va permettre leur interprétation et par conséquent l'accès à une forme de conscience partielle des cultures représentées. Cette partie a pour but de montrer l'intégration des mécanismes d'ontologisation des PCMs. Dans cette partie, je commence par expliquer les raisons du développement d'un processus avec deux niveaux de formalisation. J'introduis le premier niveau qui formalise les PCMs en graphes. Puis je présente le deuxième niveau de formalisation qui transpose les graphes en ontologies culturelles, permettant ainsi aux machines de raisonner sur les connaissances représentées.

3.3.1 Un processus en deux niveaux de formalisation

Pour la formalisation des PCMs en ontologies culturelles, j'ai choisi de développer un processus avec deux niveaux de formalisation :

1. Un premier niveau dont l'objectif est de formaliser les PCMs en un langage qui permette de travailler facilement sur leur structure graphe.
2. Un deuxième niveau dont le but est de garantir la prise en compte de la logique des connaissances contenues dans les PCMs.

3.3.2 Un premier niveau de formalisation en graphe

La formalisation des PCMs en graphes orientés s'est faite via l'API de Gephi. Les groupes nominaux ont été convertis en labels représentant chacun un noeud unique dans le graphe. La traduction des relations lexico-sémantiques en arêtes se fait directement en liant les noeuds des groupes nominaux qui les composent. Leur classe sémantique constitue le label de l'arête et permet aussi de l'orienter. Cette première formalisation permet d'effectuer un certain nombre d'opérations capables d'améliorer la qualité des modèles prototypiques culturels.

3.3.3 Un second niveau de formalisation en OWL

La formalisation en OWL des PCMs est réalisée à partir de leur format graphe. Les noeuds sont directement convertis en classes OWL sans se préoccuper de la polysémie. J'ai pris cette décision car une gestion du sens est une tâche très complexe que je ne pouvais traiter convenablement. Le label du noeud devient naturellement le label de la classe. Par exemple pour le groupe nominal *violent_crime*, le code généré va être :

```
<owl:Class rdf:ID="violent_crime">
   <rdfs:label>violent crime</rdfs:label>
</owl:Class>
```

Avant de pouvoir formaliser les relations lexico-sémantiques, il faut s'assurer d'avoir accès aux classes sémantiques. Une intégration fine de relations lexico-sémantiques peut être réalisée grâce à d'autres ontologies telles que XKOS [16] qui permet d'inclure la méronymie et la causalité. En ce qui concerne la formalisation des hyperonymes, j'ai simplement réutilisé l'implémentation de la classe *rdfs :subClassOf*.

Après la formalisation en ontologies culturelles des PCMs, leur compréhension par la machine a été dédiée à Apache Jena [17]. Jena est un framework qui permet la manipulation de documents RDF, RDFS et OWL, et fournit en plus un moteur d'inférences permettant des raisonnements sur les ontologies.

La construction des PCMs et leur formalisation en ontologies culturelles correspondent aux deux premières parties du développement d'une ACA. Il ne reste donc qu'à développer les fonctionnalités nécessaires à la médiation de ces représentations.

3.4 Les fonctions de médiations d'ontologies culturelles

La capacité à produire des médiations sur la base de représentations culturelles émiques est nécessaire pour prendre pleinement conscience des similarités et différences entre plusieurs cultures. Dans ma démarche pour concevoir une ACA, ces médiations prennent racine dans l'alignement d'ontologies. La comparaison conceptuelle de deux concepts appartenant à des ontologies différentes mène normalement à la production

16. http://rdf-vocabulary.ddialliance.org/xkos.html
17. https://jena.apache.org/

d'un score indicatif de leur similarité. Appliqué à des ontologies culturelles, ce score doit pouvoir être représentatif d'une équivalence culturelle. L'objectif de cette partie est de décrire l'intégration de la médiation d'ontologie dans l'application. Tout d'abord, je décris la représentation du sens des concepts. Je poursuis en montrant l'utilisation de ces représentations dans la construction d'un score d'équivalence issu de la comparaison de concepts. Finalement, j'explique le déroulement du processus d'alignement d'ontologies menant à la production de l'ensemble des médiations culturelles.

3.4.1 La représentation du sens des concepts

Le sens d'un concept est défini par son contexte, c'est-à-dire l'ensemble des relations et concepts qui contribuent à lui donner du sens. En matière de développement, le sens d'un concept est obtenu en raisonnant sur une ontologie afin de récupérer l'ensemble des concepts et relations le représentant. Le sens d'un concept est alors exprimé grossièrement par la somme des labels décrivant ces concepts et relations. Concrètement, dans l'expérience n'ayant que des hyperonymes, une seule requête SPARQL permet de récupérer le sens d'un concept :

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX xkos: <http://rdf-vocabulary.ddialliance.org/xkos#>
# Retourne la représentation d'un concept labélisé avec le mot "entry"
select ?result {
    # Récupère le concept associé à "entry"
    ?tmp rdfs:label "entry".
    # Cherche tous les hyperonymes
    ?tmp (rdfs:subClassof)+ ?concept.
    # Retourne les labels de l'ensemble des concepts trouvés
    ?concept rdfs:label ?result
}
```

Par exemple si le concept *murder* est associé à *violent_crime* par *hyperonymie* qui lui même est lié à *crime*, le sens de *murder* va être :

```
Sens(murder) = {(violent_crime), (crime)}
```

3.4.2 La création d'un score d'équivalence culturelle

La comparaison conceptuelle s'effectue sur le nombre de concepts ainsi que sur leurs relations. Le score est déterminé en divisant le nombre total de caractéristiques partagées par le nombre total de caractéristiques. Pour donner un exemple, si l'on prend les concepts *murder* et *robbery*, le score d'équivalence conceptuelle va être calculé de la sorte :

$$\text{É}quivalence(murder, robbery) = \frac{Taille(sens(murder) \cap sens(robbery))}{Taille(sens(murder) \cup sens(robbery))}$$

Chaque articulation (et son score) produite par comparaison est formalisée afin de permettre à l'ethnographe de l'évaluer. Après validation, le tout est stocké dans une ontologie d'articulation qui va permettre à une machine de prendre conscience des continuités et discontinuités entre les cultures représentées.

3.4.3 La production automatique des médiations culturelles

La production des médiations est réalisée en comparant les concepts provenant de différentes ontologies qui sont symbolisées par un même label. Ces comparaisons sont ensuite stockées dans une ontologie d'articulation, le formalisme étant repris sur l'*Alignment API* proposé par Euzenat [18]. Par exemple, pour deux ontologies O_1 et O_2, si les concepts *administration* comparés possèdent un score de 0.2, la représentation formelle va être :

```
<map>
    <Cell>
        <entity1 rdf:resource="http://exemple.com/o1#administration"/>
        <entity2 rdf:resource="http://exemple.com/o2#administration"/>
        <measure
            rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.2</measure>
        <relation>=</relation>
    </Cell>
</map>
```

Cette partie marque la fin de ma contribution technique pour le développement d'une ACA émique.

18. Plus d'informations sur http://alignapi.gforge.inria.fr/tutorial/tutorial1/index.html

3.5 Conclusion

Dans ce chapitre, la première partie présentait le développement d'un outil de création de PCMs. Bien que l'échantillonnage ethnographique reste une tâche manuelle, la collection des données et leur fouille pour chaque individu se fait de façon automatique. J'ai pu concevoir l'acquisition automatique de leurs données en limitant leur collection à ce qui est accessible par le web. Quant à la fouille, hormis quelques paramètres à fixer, elle permet de trouver dans les données les concepts importants, les relations sémantiques et d'identifier si ces dernières sont des hyperonymes. C'est l'analyse du consensus sur ces découvertes qui permet finalement de construire les PCMs. Si leur composition est actuellement limitée aux hyperonymes, ajouter d'autres classes sémantiques à identifier ne pose aucune difficulté technique. Cependant, l'apprentissage automatique de multiples classes est une tâche bien moins précise que dans un contexte binaire. Le processus d'ontologisation en deux étapes permet de formaliser automatiquement les PCMs en un format dédié aux graphes et un autre aux ontologies. Ce choix a été fait afin d'assurer la bonne interprétation des structures *physique* et logique, en ajoutant une complexité minime. Finalement, les médiations sont produites en interrogeant les ontologies culturelles et en comparant les similarités de leurs concepts.

Ce chapitre conclut la partie *développement* de ma thèse. La prochaine partie vient évaluer mon travail de R&D par le biais d'une *expérience*.

Chapitre 4

Expérience

4.1 Introduction

La construction d'une ACA a une vocation pratique qui ne peut s'exprimer pleinement qu'au travers d'une expérience. Pour être dans une situation idéale, un sujet controversé a été choisi. Il est porté par deux communautés distinctes avec des visions du monde connues pour être relativement incompatibles. Ce sujet, c'est l'avortement.

Le mouvement Pro-Vie regroupe des personnes souvent proches de milieux religieux qui défendent l'idée d'un « droit à la vie » qui se traduit par un rejet, entre autres choses (comme l'euthanasie), de l'avortement qui est vu comme un acte s'opposant à la vie. Le mouvement Pro-Choix, quant à lui, défend l'idée politique et éthique que les femmes devraient avoir le contrôle de leur grossesse et de leur fertilité, pouvoir accéder facilement à une contraception et à l'avortement. Alors que les Pro-Choix considèrent qu'avoir un enfant est un choix qui affecte le corps de la femme et non du fœtus, les Pro-Vie défendent l'idée selon laquelle la liberté et le bien-être des femmes ne peuvent pas être utilisés comme argument pour contester le droit à la vie du fœtus. Les Pro-Choix ne se disent pas « pro-avortement » parce qu'ils considèrent que l'avortement est le dernier recours face à une situation difficile (ce dernier pouvant empêcher un préjudice moral et physique à la femme) et parce qu'ils combattent les avortements forcés.

Le but de cette expérience n'est pas de participer au débat ni de porter un quelconque jugement de valeur sur l'une ou l'autre position. Le but est que l'ensemble des processus développés soient capables de représenter les PCMs de ces communautés, de les formaliser et de découvrir leurs similarités et leurs différences culturelles afin de produire une forme de conscience culturelle artificielle.

4.2 La construction des PCMs des communautés Pro-Vie et Pro-Choix

Dans cette partie, je vais mener une expérience dont l'objectif est d'apporter des réponses sur la faisabilité de la construction automatique de PCMs pertinents. Je vais commencer par décrire la composition des échantillons et par expliquer le procédé ethnographique qui a mené à leur construction. Après avoir détaillé la collecte des données individuelles, je vais présenter les résultats de leur fouille. Ensuite, je vais évaluer et chercher des pistes d'amélioration pour l'apprentissage des PCMs. Puis, je conduirai une analyse ethnographique des PCMs construits automatiquement. Pour finir, je présenterai et commenterai les PCMs finaux corrigés.

4.2.1 La collecte des données des communautés Pro-Vie et Pro-Choix

J'ai cherché à construire deux échantillons différents à partir de ces deux communautés. Pour construire les échantillons, j'ai fait un ensemble de recherches de manière traditionnelle en allant sur la plateforme Facebook. Cela m'a permis de trouver des pages actives associées à un site web, de 12 groupes Pro-Vie et 16 groupes Pro-Choix. Les détails sont fournis en Annexe A. Le résultat de la récolte des données web pour chaque individu est présenté sur le tableau 4.1.

Pro-Vie	
National Right to Life	52 267
Students for Life of America	10 927
Secular Pro-Life	10 938
Americans United for Life	15 327
Abort73	10 726
Pro-Choix	
NARAL Pro-Choice America	28 855
Abortion Rights Coalition of Canada	22 878
Center for Reproductive Rights	11 629
NARAL Pro-Choice Minnesota	14 427
NARAL Pro-Choice Oregon	14 181

TABLE 4.1 – Résultats de la collecte de données en nombre de phrases valides uniques

Si de nombreux individus sont absents dans les résultats, c'est parce qu'ils ne possédaient pas un assez grand nombre de données récoltées. Un filtre a été mis en place

pour retirer les individus avec moins de 10 000 phrases trouvées. Ce nombre a été fixé pour assurer le fonctionnement minimum de la fouille.

Les échantillons sont donc composés au final de 5 individus pour les Pro-Vie et les Pro-Choix.

4.2.2 L'apprentissage des PCMs

Pour rappel, les PCMs sont produits tout d'abord en définissant un domaine culturel qui est composé de concepts partagés par les individus d'un échantillon. Puis, il faut déterminer les relations sémantiques culturelles. Finalement, c'est la découverte de la classe sémantique par une tâche d'apprentissage des relations culturelles qui permettent par analyse de consensus de former des PCMs.

4.2.2.1 L'évaluation de l'efficacité de la tâche d'apprentissage

Dans l'expérience que je mène, la tâche d'apprentissage est binaire, soit une relation est un hyperonyme, soit elle ne l'est pas. Chaque échantillon possède son propre *gold standard* qui a été construit à partir de propositions de relations lexico-sémantiques obtenues grâce aux patrons syntaxiques et partagées au minimum par deux individus de l'échantillon. Les jeux de données pour l'apprentissage diffèrent donc d'un échantillon à l'autre. Le détail est montré dans le tableau 4.2.

Échantillon	**Taille du *gold standard* équilibré**	**Exemples générés**
Pro-Vie	56 * 2 = 112	112 * 5 = 560
Pro-Choix	66 * 2 = 122	122 * 5 = 610

TABLE 4.2 – Description pour chaque échantillon de la taille de leur *gold standard* et du nombre d'exemples que cela permet de produire pour l'apprentissage

Les algorithmes d'apprentissage supervisé (*Decision Tree*, *Naive Bayes*, *SVM* et *Neural Net*) ont été testés sur chaque échantillon, c'est-à-dire en mixant des exemples provenant de tous les individus qui les composent, et leur efficacité est évaluée avec une validation croisée (avec $k = 10$). La configuration des algorithmes était celle proposée par défaut par le logiciel :

1. **Naive Bayes** – Algorithme par défaut.
2. **Decision Tree** – Algorithme classique avec une séparation basée sur le critère de *gain ratio* qui est une variante de l'*information gain*. La profondeur maximale pour l'arbre a été limitée à 6.

3. **SVM** – SVM de type *S-SVC* avec un *kernel* de type *rbf* et un *gamma* fixé à 0.1.
4. **Neural Net** – Cet algorithme a été configuré avec 4 neurones dans la couche cachée, 200 cycles pour l'apprentissage, un taux d'apprentissage à 0.2, le *momentum* à 0.1 et une normalisation.

Algorithme	Précision Pro-Vie	Précision Pro-Choix
Naive Bayes (%)	55.54 +/- 2.58	61.15 +/- 4.92
Decision Tree (%)	54.46 +/- 3.01	51.97 +/- 3.03
SVM (%)	61.61 +/- 3.93	64.25 +/- 5.02
Neural Net (%)	60.36 +/- 5.22	65.57 +/- 7.83

TABLE 4.3 – Précision des tâches d'apprentissage en fonction des algorithmes et des échantillons

Le tableau 4.3 présente les résultats bruts de l'apprentissage. Il permet de voir que les algorithmes SVM et Neural Net obtiennent les meilleurs résultats avec des précisions avoisinant les 60 % pour l'échantillon Pro-Vie et 65 % pour celui Pro-Choix. C'est l'algorithme Neural Net que je vais sélectionner pour la suite de l'expérience puisqu'il est un peu plus performant que SVM.

On observe aussi que cette précision est loin d'être parfaite : 60.36 % pour l'échantillon Pro-Vie et 65.57 % pour l'échantillon Pro-Choix. Ce score va avoir d'importantes conséquences pour la suite car il signifie concrètement qu'environ 35 % à 40 % des *réponses* apportées par les individus seront fausses. Ce n'est évidemment pas une quantité négligeable.

4.2.2.2 L'évaluation de la précision dans une optique de construction de PCMs

L'évaluation de la tâche d'apprentissage doit aussi être mise en perspective par rapport à la construction de PCMs. Pour ce faire, j'évalue la qualité d'une clef de réponses issue à la fois d'un *gold standard* utilisé pour l'apprentissage et de l'analyse de consensus culturel des prédictions faites par les individus. Bien que le *gold standard* ne constitue pas techniquement un PCM, il s'y apparente puisqu'il est normalement composé de relations labélisées et partagées par l'ensemble de l'échantillon. Pour que l'analyse ethnographique puisse se faire dans des conditions optimales, il faut que la précision de la clef de réponses soit proche de 100 %. Le tableau 4.4 permet de visualiser ces résultats.

Pro-Vie

Individu	Précision (%)	Déviation standard (%)
Abort73	63.56	8.4
Americans United for Life	58.78	11.5
National Right to Life	68.56	14.2
Secular Pro-Life	72.12	13.3
Students for Life of America	66.1	15.4
Moyenne	65.8	12.6
Clef de réponses		73.21%

Pro-Choix

Individu	Précision (%)	Déviation standard (%)
Abortion Rights Coalition of Canada	73.72	8.9
Center for Reproductive Rights	76.98	9.0
NARAL Pro-Choice America	70.44	9.1
NARAL Pro-Choice Minnesota	68.84	6.1
NARAL Pro-Choice Oregon	73.01	14.2
Moyenne	72.60	9.5
Clef de réponses		80.33%

TABLE 4.4 – Justesse des tâches d'apprentissage des individus et des clefs de réponses des échantillons

Tout d'abord, on observe sur ce tableau que la précision des tâches d'apprentissage des individus est très hétérogène, le score le plus bas étant de 58.78 % pour *Americans United for Life*, le plus haut étant de 76.98 % pour *Center for Reproductive Rights*.

Ensuite sont observables les moyennes de chaque échantillon. On constate que la précision des clefs de réponses est toujours supérieure à la précision moyenne obtenue par l'échantillon. De plus, la précision de la clef de réponses reste toujours supérieure au maximum atteint par les individus qui composent les échantillons, avec 73.21 % contre 72.12 % pour l'échantillon Pro-Vie et 80.33 % contre 76.98 % pour l'échantillon Pro-Choix.

D'après ces observations, il peut être intéressant de gérer les individus selon leur précision. En effet, on peut supposer à juste titre que la qualité des clefs de réponses peut bénéficier de l'exclusion des individus avec une très faible précision au sein des échantillons.

Les précisions obtenues pour les différentes clefs de réponses indiquent que des PCMs imparfaits pourront être générés automatiquement mais vont nécessiter des vérifications de la part de l'ethnographe avant de pouvoir être utilisés. Néanmoins, les résultats laissent présager qu'une amélioration somme toute relative de la précision de la tâche d'apprentissage des individus pourrait faire tendre celle des clefs de réponses vers des

résultats dont la qualité autoriserait la production complètement automatique des PCMs.

4.2.2.3 L'amélioration de la création des PCMs par la gestion des précisions individuelles

De nombreux facteurs liés aux individus eux-mêmes (ou plutôt aux données qui leur sont associées) peuvent être à l'origine de variations de la précision de la tâche d'apprentissage, telles que la manière dont sont écrits leurs documents ou encore la disparité des sujets abordés. La justesse des clefs de réponses est intrinsèquement dépendante de la qualité des réponses. C'est cette évidence que je cherche à étudier. Pour l'exemple, j'ai choisi l'échantillon qui possède les plus grandes variations en matière de précision, c'est-à-dire les Pro-Vie.

Nb. Individus	5	4	3
PMI (%)	65.83	67.59	68.93
PdlCdR (%)	73.21	74.11	75.00

TABLE 4.5 – Nouveaux résultats obtenus par analyse de consensus en retirant successivement l'individu le moins précis (PMI : Précision Moyenne des Individus et PdlCdR : Précision de la Clef de Réponses)

On voit sur le tableau 4.5 que la précision de la clef de réponses augmente après chaque retrait de l'individu ayant la plus mauvaise précision passant de 73.21 % à 75.00 %.

Les observations mettent en évidence l'impact négatif significatif sur les clefs de réponses d'individus avec des précisions trop faibles. Afin d'obtenir des PCMs de qualité, il semble donc nécessaire de s'assurer que la qualité de l'explicitation des relations lexico-sémantiques de chaque individu est suffisante. Parce que le nombre d'individus conditionne d'une part la qualité des clefs de réponses par effet de consensus et d'autre part la représentativité de l'échantillon, il faut être vigilant et chercher le bon compromis entre le mauvais impact de la précision de certains individus et les conséquences de la réduction de la taille de l'échantillon.

Bien que cette expérience soit réalisée dans une perspective anthropologique, il est important de noter la contribution de la démarche ethnographique pour la découverte automatique des hyperonymes sans utilisation de ressources externes. À titre de comparaison, Hearst (1992) a réussi à obtenir 50 % de précision avec l'extraction directe d'hyperonymes en utilisant des patrons syntaxiques. Cederberg et Widdows (2003) ont

amélioré ce score à l'aide de mesures de similarité provenant du modèle d'espace vectoriel LSA, atteignant 58 %. Widdows et Dorow (2002), avec une approche basée sur des graphes, réussissent à atteindre 82 % de précision. Si les précisions des individus dans mon expérience sont généralement moins bonnes que celles que l'on peut trouver dans l'état de l'art, il est à noter que l'approche dans sa globalité permet d'obtenir des scores très compétitifs avec 80.33 % de précision pour la clef de réponses de l'échantillon Pro-Choix sans remaniement. De plus, un aspect important de l'approche par analyse de consensus est qu'elle permet d'améliorer la qualité d'une tâche de fouille sans retoucher à la tâche mais en manipulant le nombre d'individus.

4.2.3 La création du domaine culturel et des relations sémantiques culturelles

Pour la construction du domaine culturel, le nombre de concepts importants extraits pour chaque individu a été fixé à 2 000. Le nombre de meilleures relations à conserver pour chaque concept appartenant à ce domaine a, lui, été défini à 200. La majorité pour le consensus requis a, lui, été fixé pour les concepts culturels à 2 individus sur 3 et 1 individu sur 3 pour les relations sémantiques culturelles, car le consensus est plus difficile à obtenir pour ces dernières. Le tableau 4.6 présente les résultats.

Échantillon	**Domaine culturel**	**Relations sémantiques culturelles**
Pro-Vie	1 265	78 382
Pro-Choix	1 432	141 978

TABLE 4.6 – Tableau récapitulatif de la taille du domaine culturel et du nombre de relations sémantiques culturelles associées pour chaque échantillon

On observe que les échantillons ont des domaines culturels de différentes tailles, respectivement 1 265 et 1 432 pour les Pro-Vie et Pro-Choix. Les différences en matière de taille de domaine culturel sont certainement dues à des niveaux de consensus hétérogènes au sein des échantillons, les individus Pro-Choix recueillis étant dans ce cas-là plus homogènes. Cela semble se confirmer avec la production des relations sémantiques culturelles qui est respectivement de 78 382 et 141 978.

4.2.4 L'analyse ethnographique des PCMs

Les PCMs des deux échantillons ont été produits en se basant sur leur domaine culturel, leurs relations sémantiques culturelles et l'identification partagée de leur classe sémantique.

4.2.4.1 La vérification d'une culture unique

Un PCM n'est valide que s'il ne représente qu'une seule et unique culture. Ce critère est déterminé par le ratio des valeurs propres produit à partir de la matrice d'accords entre les individus. Pour rappel, l'existence d'une culture unique est acceptée pour tout ratio supérieur à 3.

	Pro-Vie	**Pro-Choix**
Première valeur propre	1.397	1.010
Deuxième valeur propre	0.249	0.186
Ratio des deux	5.601	5.428

TABLE 4.7 – Résultats des valeurs propres de chaque échantillon

On constate sur le tableau 4.7 que les deux échantillons ont un ratio supérieur à 3, avec 5.601 et 5.428.

Ces scores signifient que les individus au sein des échantillons Pro-Vie ont une culture homogène, de même que ceux des échantillons Pro-Choix. Cependant, cela n'exclut pas nécessairement la présence d'individus exogènes dans les échantillons. Le ratio de valeurs propres ne donne pas de détails sur la répartition des accords entre les individus et le groupe.

4.2.4.2 L'appréciation culturelle des PCMs

Elle est possible grâce au niveau de compétences des individus. Des niveaux élevés et homogènes sont indicatifs d'une cohérence ou cohésion culturelle entre les individus. Des niveaux faibles et hétérogènes vont être révélateurs de disparités culturelles dont l'origine peut prendre racine dans la présence de sous-cultures au sein de l'échantillon. Les scores de compétences s'étendent de 0 à 1. Un score élevé signifie qu'un individu est proche du standard culturel du groupe.

Sur le tableau 4.8 on observe que la moyenne des scores de compétences n'est pas élevée puisque les Pro-Vie ont un score de 0.520 et les Pro-Choix de 0.445.

Pro-Vie

Americans United for Life	0.346
Abort73	0.529
Secular Pro-Life	0.556
Students for Life of America	0.571
National Right to Life	0.601
Moyenne : 0.520	

Pro-Choix

Center for Reproductive Rights	0.538
Abortion Rights Coalition of Canada	0.421
NARAL Pro-Choice Minnesota	0.480
NARAL Pro-Choice America	0.367
NARAL Pro-Choice Oregon	0.422
Moyenne : 0.445	

TABLE 4.8 – Score de compétences associées aux individus de chaque échantillon

Maintenant, si on regarde en détail, on se rend compte par exemple que l'individu *Americans United for Life* a un score de compétences de 0.346, ce qui est beaucoup plus faible que la moyenne de l'échantillon Pro-Vie. Il en va de même dans l'autre échantillon, avec par exemple l'individu *NARAL Pro-Choice America* qui a un score de 0.367.

En considérant que les différences de scores de compétences ont une origine culturelle, il peut être judicieux, dans le but de gagner en homogénéité, d'exclure les individus qui possèdent un niveau de compétences très en-deçà de la moyenne.

Cependant, la correspondance score de compétences et homogénéité culturelle est valide pour une analyse ethnographique classique où la qualité de l'explicitation est garantie. Dans le cas d'une explicitation automatique qui n'apporte pas cette garantie, il convient de vérifier que ces divergences ne sont pas causées par la précision somme toute relative de la tâche d'apprentissage.

4.2.4.3 L'analyse de la corrélation entre précision et score de compétences

Les scores de compétences des individus sont construits en fonction de la matrice d'accords qui résulte des réponses apportées par l'ensemble des membres d'un échantillon.

On observe sur la figure 4.1 une répartition très hétérogène des scores de compétences en fonction des précisions des tâches d'apprentissage des individus. On constate qu'il n'y a pas de corrélation entre la précision et le score de compétences de certains individus.

La construction de PCMs de qualité passe par la gestion, d'une part, de la précision de la tâche d'apprentissage de chaque individu, et, d'autre part, des compétences culturelles au sein des échantillons.

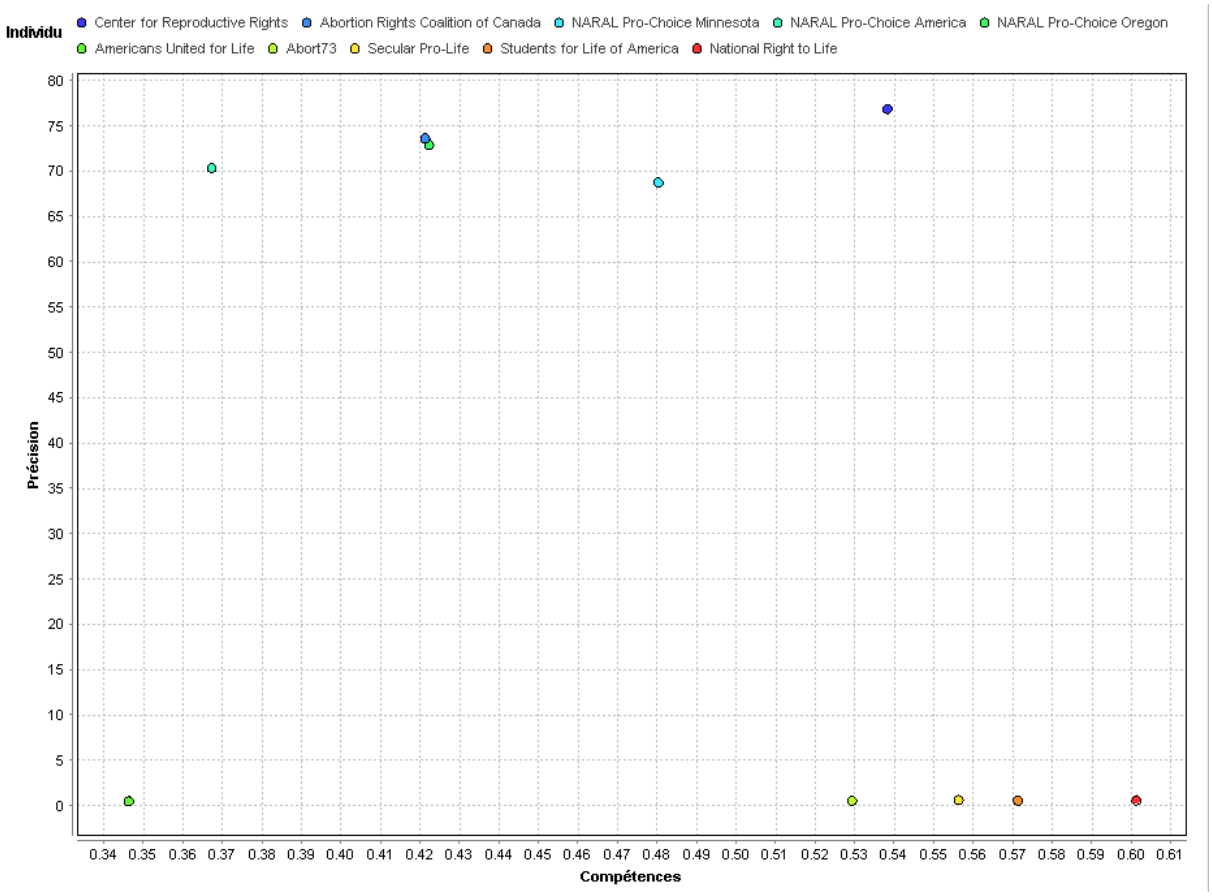

FIGURE 4.1 – Corrélation entre la précision de la tâche d'apprentissage des individus et leur score de compétences. Visualisation produite par RapidMiner.

4.2.4.4 Raffiner les PCMs par la gestion des compétences individuelles

Bien qu'un échantillon possède une culture principale, ce dernier peut être composé d'individus appartenant à ou étant influencé par d'autres cultures. Ces individus sont sources de désaccords qui ont naturellement un impact négatif dans la recherche et l'atteinte d'un consensus. C'est pourquoi détecter ce genre d'individus à partir des compétences individuelles et les exclure de l'échantillon devrait en toute logique améliorer la qualité de la clef de réponses associée. Pour tester cette hypothèse, j'ai pris l'échantillon Pro-Choix car il possède les plus mauvais scores en ce qui concerne les scores de compétences.

On observe sur le tableau 4.9 que le ratio de valeurs propres augmente en retirant l'individu avec la pire compétence allant d'un ratio de 5.42 à 10.059. En en retirant un deuxième, le ratio atteint 100, ce qui est certainement dû à un échantillon trop petit.

La gestion des compétences des individus présents dans les échantillons permet d'améliorer le ratio de valeurs propres. L'exclusion intelligente de certains individus

Nb. Individus	5	4	3
NARAL Pro-Choice Minnesota	0.479	0.458	
Center for Reproductive Rights	0.538	0.490	0.543
NARAL Pro-Choice Oregon	0.422	0.462	0.474
Abortion Rights Coalition of Canada	0.421	0.458	0.397
NARAL Pro-Choice America	0.366		
Moyenne	0.445	0.467	0.471
Ratio de valeurs propres	5.42	10.059	100

TABLE 4.9 – Nouveaux résultats obtenus par analyse de consensus en retirant successivement l'individu le moins compétent

permet d'obtenir une meilleure homogénéité culturelle et par conséquent de construire des PCMs plus pertinents.

4.2.4.5 La finalisation des modèles culturels

Pour finaliser la production des PCMs, j'ai revu l'ensemble de leurs relations afin de corriger d'éventuelles erreurs. En omettant les relations avec la classe *none*, les PCMs des trois échantillons sont composés de :

— 43 concepts et 52 hyperonymes pour les Pro-Vie ;

— 82 concepts et 96 hyperonymes pour les Pro-Choix.

Une visualisation des PCMs est offerte sous la forme de réseaux sémantiques sur les figures 4.2 et 4.3.

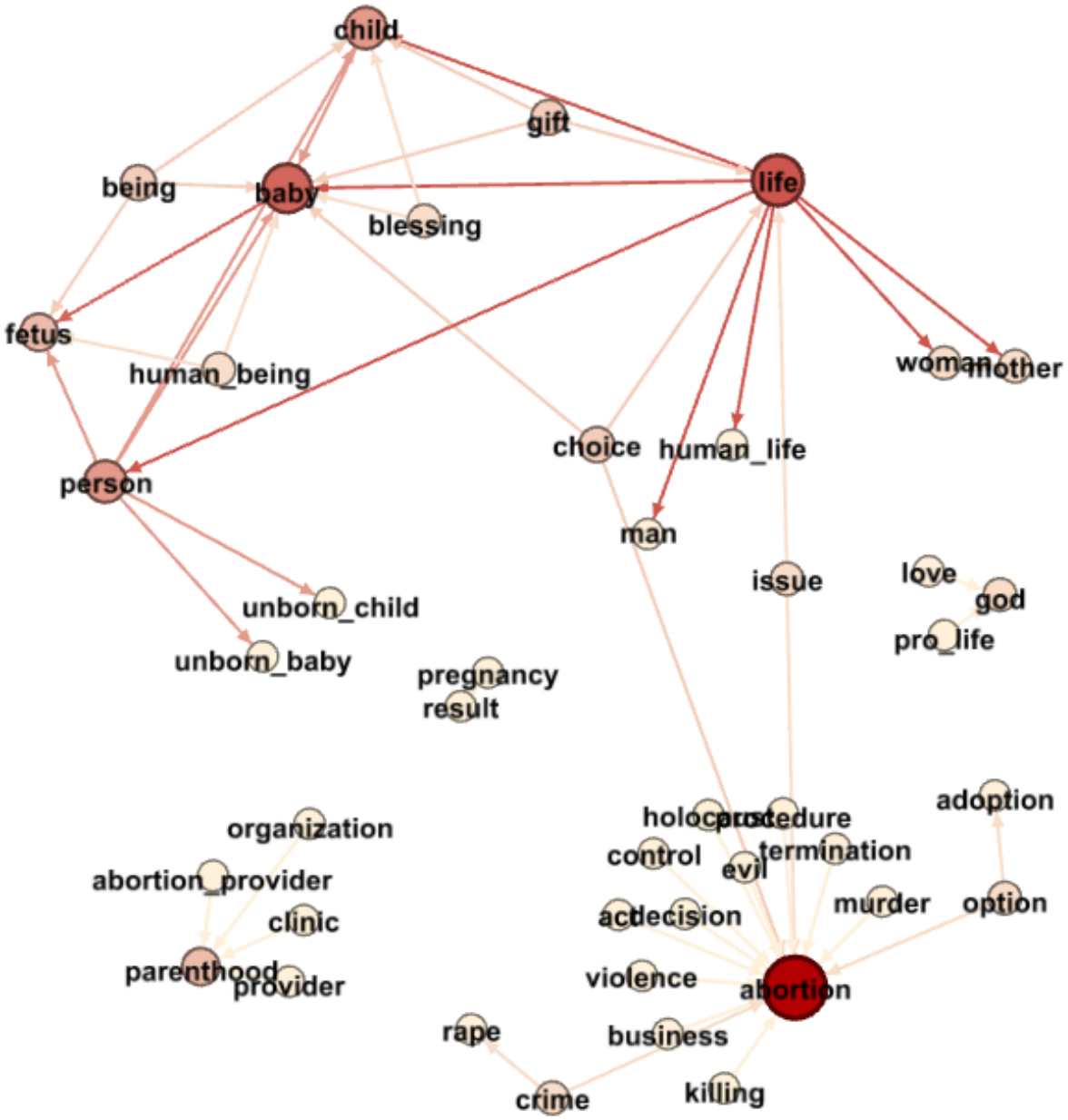

FIGURE 4.2 – Représentation du PCM pour l'échantillon Pro-Vie. Visualisation produite par Gephi.

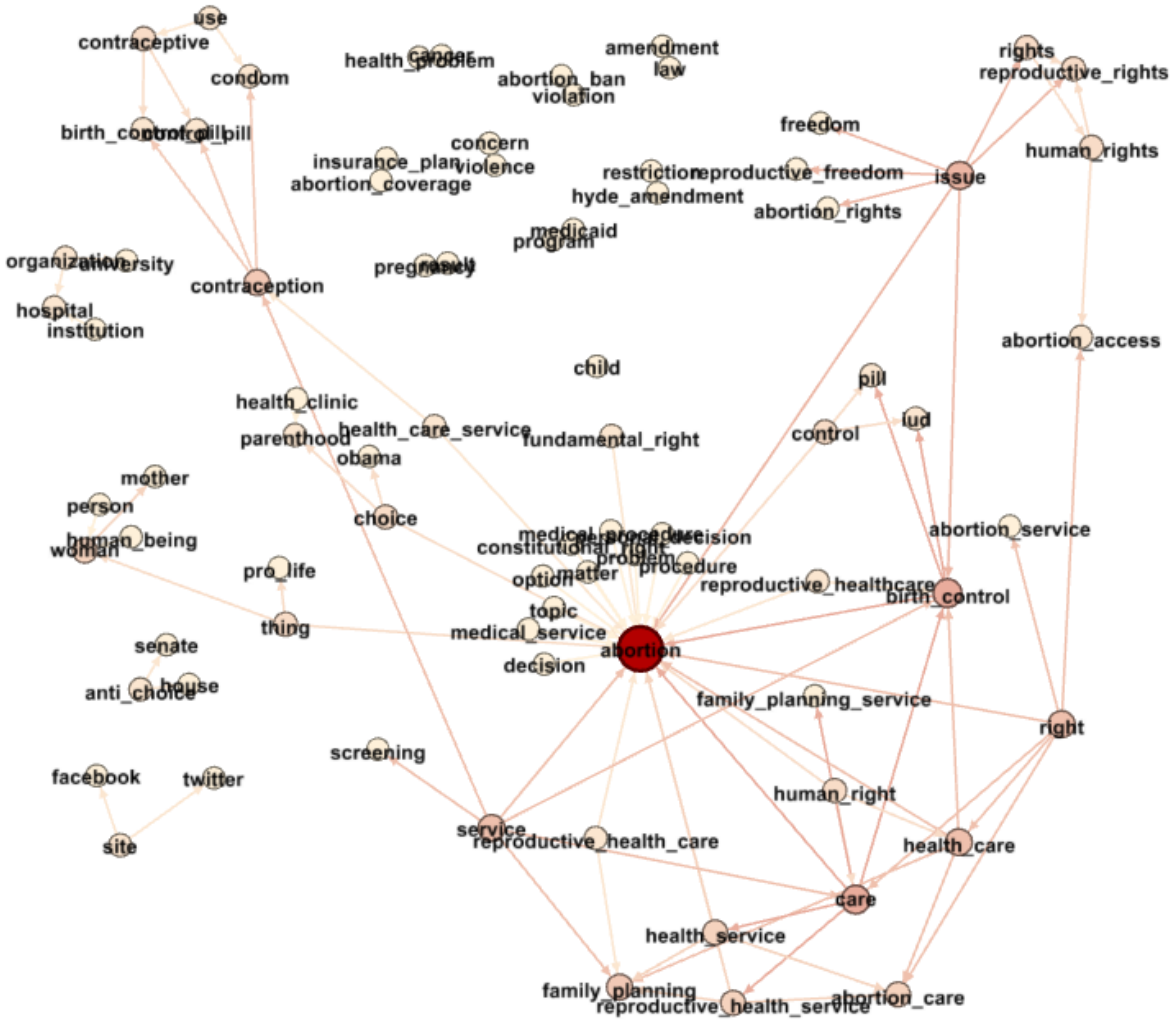

FIGURE 4.3 – Représentation du PCM pour l'échantillon Pro-Choix. Visualisation produite par Gephi.

Bien qu'une cinquantaine ou une centaine de relations sont insuffisantes pour représenter les domaines culturels dans toute leur complexité, ces représentations sont néanmoins bien plus riches qu'un système de valeurs tel que celui d'Hofstede (Hofstede et Hofstede, 2001; Hofstede et Bond, 1984) avec ses 6 dimensions. Il faut aussi prendre en considération l'unique prise en compte des hyperonymes alors que d'autres relations lexico-sémantiques courantes et pertinentes capables d'enrichir les PCMs existent.

Une analyse ethnographique légère permet de mettre en évidence sur les figures 4.2 et 4.3 une thématique commune relative au domaine de l'avortement (*abortion*). Cependant, de nombreux autres sujets sont spécifiques aux deux communautés. Par exemple, il y a la vie (*life*) et les enfants (*child*) chez les Pro-Vie, alors que les Pro-Choix sont plus centrés sur la contraception (*contraception*) et les droits (*rights*). Ces différences sont tout à fait en accord avec la présentation que j'avais faite de ces communautés.

L'objectif de cette partie consistait à évaluer la faisabilité d'une production quasi automatique de PCMs. À l'issue de l'expérience conduite sur le domaine de l'avortement, les résultats indiquent que ce but est atteignable. En effet, cette approche autorise la construction de PCMs d'une qualité raisonnable par une gestion double des précisions et compétences individuelles. Il faut maintenant regarder dans quelle mesure la formalisation automatique des PCMs peut se faire.

4.3 Les ontologies culturelles des communautés Pro-Vie et Pro-Choix

Je vais poursuivre mon expérience en appliquant le processus d'ontologisation afin de convertir les PCMs en graphes puis en ontologies. Dans cette partie, l'accent va être mis sur les améliorations que permet la formalisation. Je vais tout d'abord analyser d'un point de vue structurel les graphes créés à partir des PCMs. Je proposerai une solution simple pour remédier aux absences des relations lexico-sémantiques évidentes. Je vais ensuite montrer comment supprimer les relations logiques redondantes. Pour continuer avec le nettoyage, je présenterai une possible manière de gérer les groupes isolés de concepts et relations sémantiques.

4.3.1 Les résultats de la formalisation des PCMs

La formalisation des PCMs doit permettre leur interprétation par une machine. Si l'on observe les graphes produits à partir des PCMs sur les figures 4.2 et 4.3, on s'aperçoit qu'ils sont difficilement lisibles en l'état :

1. L'acquisition de structures conceptuelles à partir de textes basés sur la méthode d'Hearst (1992) ne prend généralement pas en considération les relations composites. Une relation composite est celle existant par exemple dans l'échantillon Pro-Choix pour le concept *reproduction rights* qui contient intrinsèquement les relations :

   ```
   (rights)-[hypernym]->(reproduction_rights)
   ```

   Ces relations sont tellement évidentes qu'elles ne sont presque jamais explicitées au sein d'un texte au travers de motifs syntaxiques tels que : « ... les droits reproductifs et autres droits ... ».

2. On constate qu'il existe de la redondance logique dans la structure des PCMs. Par exemple, pour l'échantillon Pro-Choix, le concept « problème » (*issue*) est à la fois l'hyperonyme de « droits » (*rights*) et de « droits reproductifs » (*reproductive rights*) qui sont eux-mêmes des droits. Il n'y a donc pas besoin de spécifier directement que les droits reproductifs sont des problèmes.

3. Finalement, on remarque des agrégations de concepts et de relations de tailles

variées, formant visuellement des îlots plus ou moins gros. Bien que les petits îlots ne soient pas des erreurs (ce sont effectivement des relations valides), ils ne contribuent pas au modèle global et s'apparentent à du bruit dans les représentations des PCMs.

La formalisation des PCMs en graphes donne la possibilité à une machine d'interpréter ces représentations, mais offre surtout la capacité de travailler sur leur structure afin de les améliorer.

4.3.2 L'ajout des relations trop évidentes

Un graphe sémantique permet d'inférer l'existence de certaines relations évidentes avec un très haut niveau de certitude en se basant sur le contexte des concepts. C'est le cas des groupes nominaux qui ont dans leur voisinage des concepts avec lesquels ils ont une relation de composition :

```
(human_rights)-[hypernym]->(reproductive_rights)
(rights)-[hypernym]->(reproductive_rights)

// Il faut ajouter
++ (rights)-[hypernym]->(human_rights)
```

Échantillon	Propositions	Valides	Nouvelles relations	Total
Pro-Vie	6	5	5	57
Pro-Choix	27	26	25	121

TABLE 4.10 – Résultats des ajouts possibles de taxons évidents pour les graphes des échantillons Pro-Vie et Pro-Choix

Sur le tableau 4.10 on observe qu'il n'y a en effet presque aucune erreur dans la production des hyperonymes à partir de l'analyse des relations composites et du graphe sémantique. Les seules erreurs générées sont liées aux mouvements Pro-Vie et Pro-Choix qui ont produit les concepts vie (*life*) et choix (*choice*).

L'ajout des hyperonymes évidents pour l'être humain dans des graphes constitués uniquement de ces mêmes relations est une tâche qui peut être réalisée sans risque majeur d'erreur. L'enrichissement apporté constitue un apport conséquent de nouvelles relations. Cependant, l'ajout de ces relations contribue au problème de redondance logique précédemment évoqué.

4.3.3 La suppression des duplicata logiques

La redondance logique est basée sur la transitivité des hyperonymes. Ces duplicata forment un bruit inutile au niveau de la structure des graphes. C'est pourquoi les relations qui peuvent s'exprimer par transitivité sont supprimées. Pour illustrer :

```
(rights)-[hypernym]->(human_rights)
(rights)-[hypernym]->(freedom)
(human_rights)-[hypernym]->(freedom)

// Il faut supprimer
-- (rights)-[hypernym]->(freedom)
```

En d'autres termes, en parcourant le graphe, si à partir d'un même noeud il existe plusieurs chemins pour atteindre une même destination, le plus long chemin est conservé alors que le plus court est supprimé.

Échantillon	**Nb. de suppressions**	**Nb. total de relations**
Pro-Vie	15	42
Pro-Choix	37	84

TABLE 4.11 – Résultats des suppressions de relations logiques redondantes pour les graphes des échantillons

On constate sur le tableau 4.11 que le nombre de suppressions est un peu supérieur au nombre d'ajouts précédemment effectués. Comme pressenti, cela s'explique par le fait que le traitement des relations évidentes a généré de la redondance logique en plus de celle qui existait déjà.

La redondance logique n'entrave pas l'interprétation par la machine des PCMs, mais altère la structure *physique* de ces derniers. La gestion de cette redondance permet de nettoyer les connaissances formalisées et d'améliorer leur lisibilité.

4.3.4 La suppression des îlots

La représentation formelle en graphe permet de raisonner sur le voisinage de chaque nœud afin de s'assurer qu'ils ne soient pas isolés. Dans cette expérience, je considère qu'un ensemble de nœuds connectés par des relations formant un îlot est isolé si la distance maximale trouvée entre deux nœuds est inférieure à 2. Très concrètement, je récupère tous les nœuds du graphe, pour chaque nœud je vais chercher son plus lointain voisin

par itérations successives. La distance est incrémentée de 1 pour chaque nouveau nœud rencontré. À la fin, la valeur maximale obtenue est attribuée à l'ensemble des nœuds parcourus. Les nœuds identifiés comme appartenant à des îlots isolés sont finalement supprimés. Une visualisation des graphes nettoyés des différents PCMs est offerte avec les figures 4.4 et 4.5.

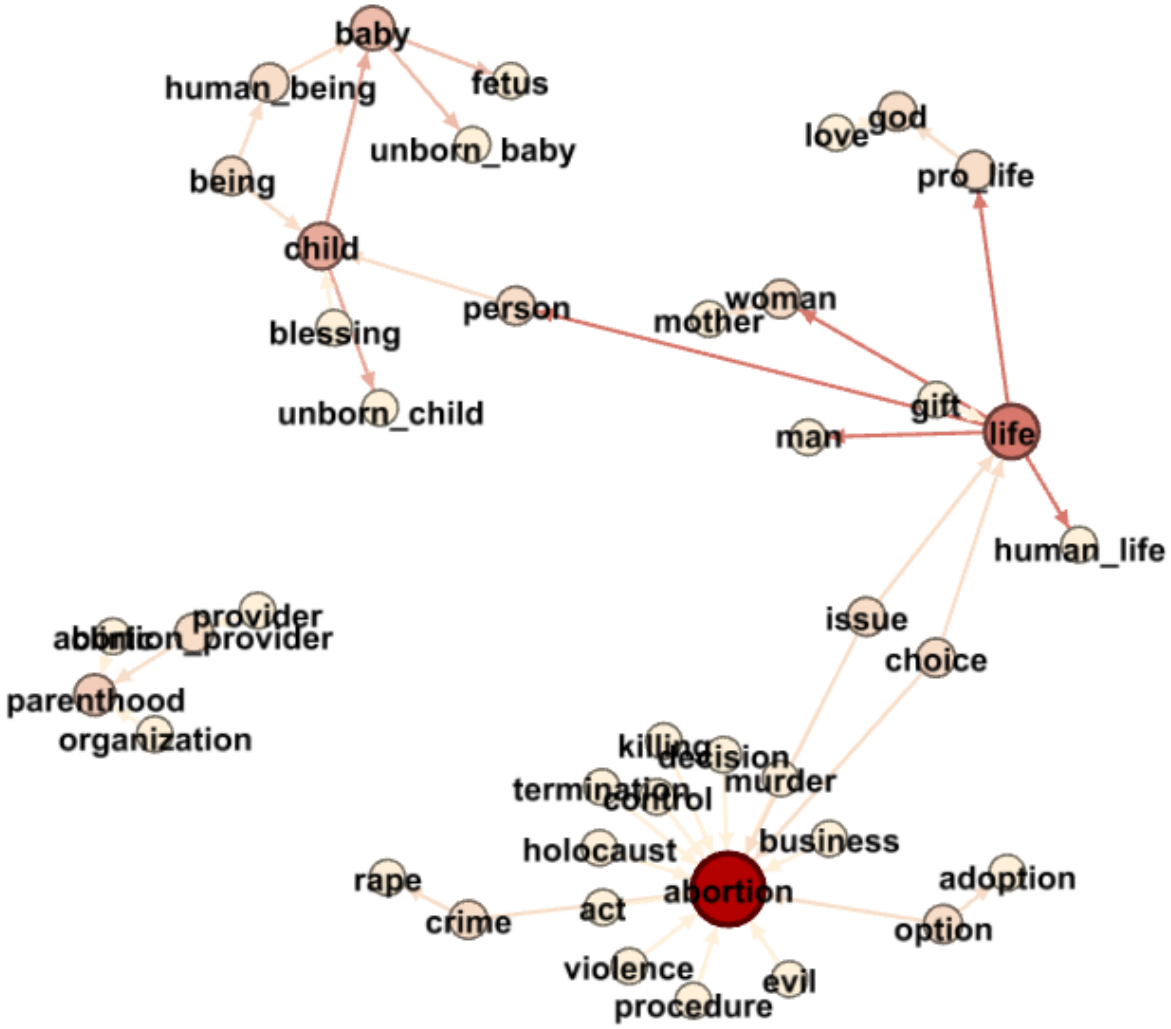

FIGURE 4.4 – Représentation du graphe nettoyé du PCM pour l'échantillon Pro-Vie. Visualisation produite par Gephi.

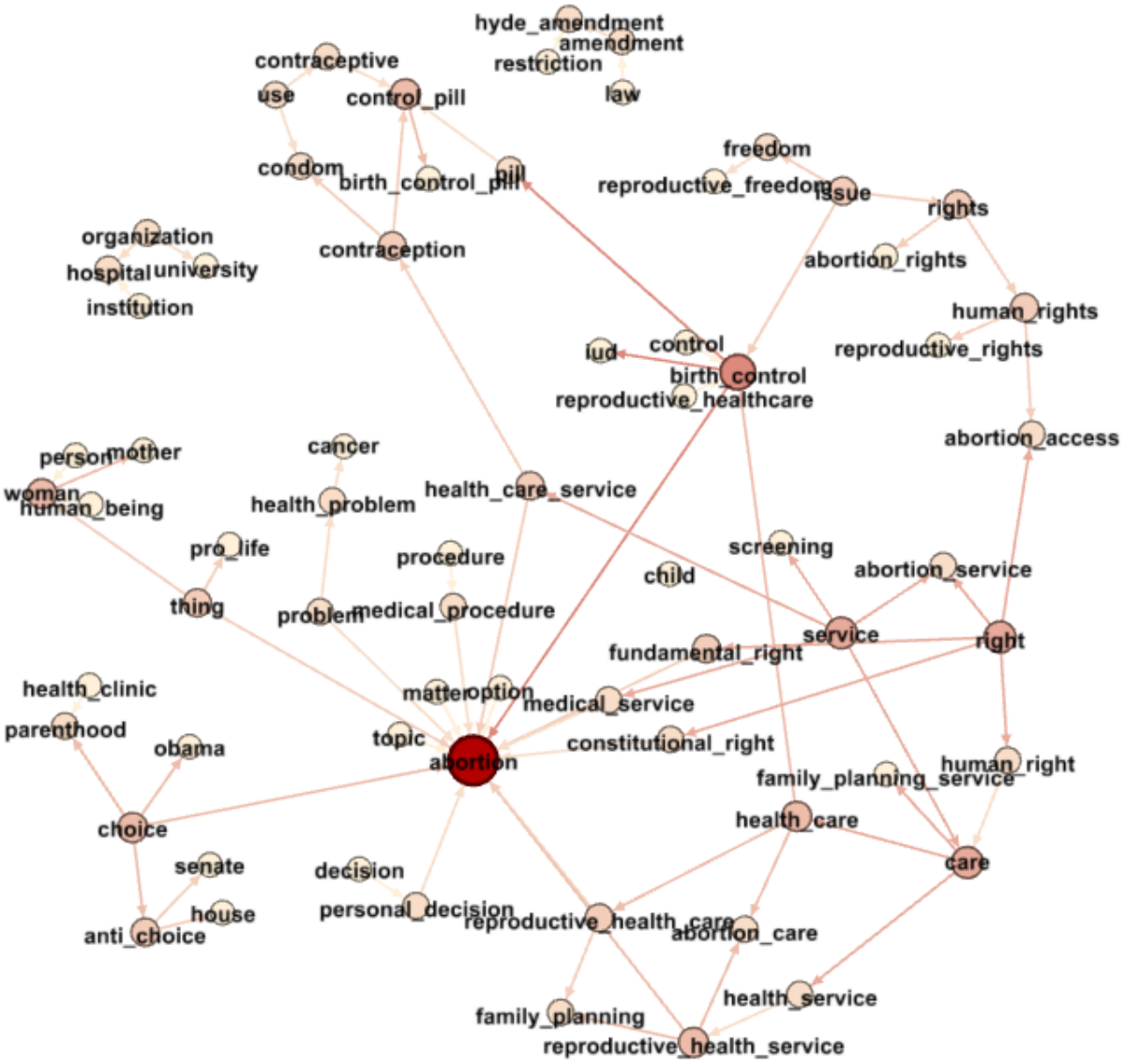

FIGURE 4.5 – Représentation du graphe nettoyé du PCM pour l'échantillon Pro-Choix. Visualisation produite par Gephi.

L'ajout des hyperonymes évidents, la suppression des relations logiquement redondantes et la destruction des îlots isolés permet d'améliorer drastiquement la lisibilité des PCMs produits par chaque échantillon.

Après avoir formalisé en graphes les PCMs, les avoir enrichis et nettoyés, ils vont être maintenant exprimés en OWL, qui est le langage formel pour les ontologies. C'est ce deuxième niveau de formalisation qui va réellement permettre à une machine d'interpréter les représentations culturelles émiques des échantillons Pro-Vie et Pro-Choix. Transposées en ontologies culturelles, les cultures représentées vont maintenant pouvoir être comparées dans le but de découvrir d'éventuelles similarités et différences culturelles ou, plus exactement, dans le but de produire des médiations culturelles.

4.4 Les médiations culturelles entre Pro-Vie et Pro-Choix

Dans cette partie, je vais continuer mes expériences afin de confirmer ou d'infirmer la capacité des scores de comparaison conceptuelle à indiquer des similarités ou des différences culturelles entre ontologies. Dans un premier temps, je vais analyser les résultats produits automatiquement à partir des ontologies culturelles issues de l'expérience précédente. Dans un deuxième temps, je vais étudier l'impact des problèmes de couverture d'ontologies sur les résultats que j'obtiens.

4.4.1 L'évaluation des médiations culturelles suggérées automatiquement

Les scores d'équivalence sont produits en comparant des représentations de deux concepts possédant un label similaire et provenant d'ontologies différentes. Dans le cas d'ontologies culturelles, ce score est censé traduire des équivalences culturelles. Normalement un score faible, entre 0 et 0.5, doit indiquer une forte différence culturelle. À l'inverse, un score élevé, entre 0.5 et 1, doit exprimer une similarité culturelle forte.

L'étude de la pertinence des médiations culturelles suggérées par l'application s'est faite en attribuant une note à chacune des médiations :

— 0 – Le score n'est pas représentatif de la réalité culturelle. Cette situation arrive par exemple quand deux concepts culturellement similaires obtiennent un score inférieur à 0.5.

— 1 – Le score indique partiellement la bonne réponse. C'est ce qui peut se passer quand deux concepts culturellement très différents obtiennent un score de 0.4.

— 2 – Le score traduit correctement l'équivalence culturelle.

Parmi les concepts des ontologies Pro-Choix et Pro-Vie, 14 concepts seulement sont comparables. D'après le tableau 4.12, les médiations proposées sont valides à environ 61 %. On remarque que les médiations fausses sont principalement celles qui obtiennent des scores d'équivalence de 0, ce qui signifie que les concepts sont définis par des hyperonymes mais qu'ils n'en ont aucun en commun. Quant à eux, les scores de 1 indiquent qu'aucun hyperonyme n'a été trouvé et que les concepts se partagent eux-mêmes. Seuls les concepts de l'« avortement » (*abortion*) et de la « mère » (*mother*) ont des scores entre 0 et 1.

Concept	Score d'équivalence	Note
pro_life	0.0	0
woman	0.0	0
person	0.0	0
human_being	0.0	0
parenthood	0.0	0
mother	0.125	1
abortion	0.17	2
choice	1.0	2
issue	1.0	2
option	1.0	2
decision	1.0	2
procedure	1.0	2
organization	1.0	2
control	1.0	2

TABLE 4.12 – Résultats de l'évaluation des médiations culturelles produites (classés par ordre croissant)

Pour le premier concept, le score semble juste puisqu'on s'attend à une forte différence culturelle. Pour le second, en revanche, le score indique une disparité culturelle forte qui, a priori, ne reflète pas la réalité.

Si les médiations avec un score de 1 sont généralement bonnes, celles avec un score de 0 ne le sont pas souvent. Des scores de 0 sont symptomatiques d'un problème de couverture qu'il convient d'analyser.

4.4.2 La gestion des problèmes de couverture

Le score produit en comparant les concepts ne peut pas faire la distinction entre les différentes origines de conceptualisation, qu'elle soit liée à un problème de couverture des ontologies ou qu'elle soit réellement associée à des visions du monde divergentes. L'étude plus fine de la pertinence de médiations culturelles produites sur la base de comparaisons conceptuelles requiert de s'assurer d'une couverture adéquate du domaine. Une nouvelle évaluation des médiations a été réalisée après s'être assuré que les domaines culturels de chaque échantillon était mieux couvert. Concrètement, les relations suivantes ont été ajoutées :

```
# Pour les Pro-Vie
(human_being)-[hypernym]->(woman)


# Pour les Pro-Choix
```

```
(being)-[hypernym]->(human_being)
(clinic)-[hypernym]->(health_clinic)
```

Concept	Score d'équivalence	Note
pro_life	0.0	0
person	0.0	0
parenthood	0.167	2
abortion	0.171	2
woman	0.25	2
mother	0.333	2
procedure	1.0	2
clinic	1.0	2
being	1.0	2
option	1.0	2
decision	1.0	2
human_being	1.0	1
issue	1.0	2
organization	1.0	2
control	1.0	2
choice	1.0	2

TABLE 4.13 – Résultats de l'évaluation des médiations culturelles produites après avoir retouché manuellement la couverture des domaines (classés par ordre croissant)

Sur le tableau 4.13 on observe que les notes valident à 82 % les médiations produites. Les scores obtenus pour les concepts de « femme » (*woman*) et de « mère » (*mother*) sont significatifs des différences culturelles sous-jacentes entre ces deux communautés. Le score associé au « planning familial » (*parenthood*) est lui aussi révélateur d'une différence importante entre, d'une part, des Pro-Vie qui perçoivent ces cliniques comme des fournisseurs de services, et d'autre part, des Pro-Choix qui voient seulement l'aspect santé. Le concept de « Pro-Vie » (*pro_life*) est resté à 0 puisqu'il a pour origine une erreur. Contrairement à ce dernier, si le concept de « personne » (*person*) est toujours à 0, c'est parce qu'il n'a pas bénéficié des changements effectués.

Globalement, les nouveaux résultats obtenus prouvent qu'assurer la couverture des domaines permet de générer des scores d'équivalence culturelle qui ont plus de sens.

La comparaison d'ontologies culturelles permet de faire émerger des médiations qui elles-mêmes le sont, à condition que la couverture des domaines soit correctement réalisée. La réussite de cette opération prend donc racine dans la construction des PCMs. Une solution pour améliorer les résultats consiste alors à traiter un plus grand nombre de relations lexico-sémantiques lors de la tâche d'apprentissage.

4.5 Conclusion

Le but de cette expérience était de prouver que le développement d'une ACA émique était possible. La construction des PCMs Pro-Vie et Pro-Choix a démontré que le processus ethnographique automatique autorisant leur acquisition permettait effectivement la création de représentations spécifiques et détaillées pour chacune de ces communautés. Limité aux relations d'hyperonymie et avec une qualité d'explicitation encore imparfaite, le processus peut, d'une part, s'étendre à l'identification de plusieurs classes sémantiques sans aucune problème, et, d'autre part, gagner en précision en améliorant la précision de la tâche de fouille.

L'expérience a permis de montrer la facilité avec laquelle l'ontologisation des PCMs en ontologies culturelles se fait. De plus, cette formalisation autorise un certain nombre d'opérations d'enrichissement et de nettoyage qui augmente significativement la qualité des représentations.

L'alignement des ontologies des Pro-Vie et Pro-Choix a prouvé qu'il était possible de découvrir leurs similarités et différences culturelles. Bien qu'il soit possible de produire des médiations culturelles, la bonne couverture des domaines conditionne le succès de cette opération. Le développement d'une ACA émique dépend donc principalement de la qualité des PCMs initialement construits.

Ce chapitre conclut la partie *expérience* de ma thèse, le prochain, lui, la conclut.

Chapitre 5

Conclusion

L'objectif de cette thèse était de contribuer à la modélisation d'une conscience culturelle artificielle émique par les ontologies. Une telle conscience est fondée sur la capacité à produire des représentations et médiations culturelles formelles. Cet objectif a été atteint en développant un ensemble de processus permettant de produire des modèles culturels prototypiques, de les formaliser en ontologies culturelles et de les comparer. Les PCMs sont des représentations culturelles émiques. L'explicitation de ces modèles, basée sur l'échantillonnage, la collecte de données web, la fouille de textes et l'analyse de consensus culturel ont prouvé qu'une production quasi automatique était possible. Si cette méthode est à moindre coût, elle est associée à des contraintes qui limitent son application à certaines communautés dont les membres doivent posséder un volume important de données accessibles sur la toile. Les modèles culturels prototypiques sont adaptés à la formalisation en ontologies dites culturelles. Ces ontologies permettent aux machines d'accéder à la conscience de différentes cultures par l'interprétation de leur représentation. La médiation d'ontologies, en particulier leur alignement sur la base de comparaisons conceptuelles, est liée à deux problèmes : la couverture des ontologies et la portée de leurs concepts. La médiation d'ontologies complètes sur un même domaine permet alors la production d'équivalences dont la nature culturelle est garantie par celle des ontologies. Ces médiations culturelles permettent finalement aux machines d'identifier les similarités et différences entre diverses cultures.

Les contributions

Le processus développé et l'expérience réalisée dans cette thèse ont permis de mettre en évidence la faisabilité du développement d'une conscience culturelle artificielle émique.

Le processus ethnographique de découverte de modèles culturels mis en place permet la production de représentations culturelles de manière à la fois assez rapide et précise (précisions supérieures entre 73.21 % et 80.33 %). Ces résultats révèlent notamment la pertinence de l'approche ethnographique pour la fouille de textes.

La formalisation en deux temps des PCMs en graphes, puis en ontologies culturelles permet, sans ajouter de complexité, d'augmenter les possibilités de traitement pour améliorer la qualité des modèles. Par exemple, la représentation en graphe orienté permet une gestion assez aisée de la transitivité et permet, par exemple via l'outil Gephi, qu'un ethnographe puisse appliquer facilement des modifications/corrections.

Finalement, lors de la recherche de production de médiations culturelles, l'expérience a permis de montrer la pertinence de la comparaison conceptuelle pour produire des médiations culturelles de qualité. Cependant, les résultats obtenus mettent aussi en évidence la nécessité d'une bonne couverture des domaines par chacune des ontologies afin que les mesures produites expriment pleinement des continuités ou discontinuités culturelles.

Les perspectives

Dans cette thèse de nombreux aspects ont été simplifiés pour des raisons pratiques. Par exemple, la culture doit être pensée avec le multilinguisme tout comme doit l'être le développement d'une conscience culturelle artificielle.

En l'état, le processus ethnographique de construction des PCMs n'accomplit qu'une tâche de découverte d'hyperonymes. Cependant, l'intégration d'un apprentissage automatique de qualité pour des classes sémantiques multiples pourrait améliorer amplement la richesse des représentations.

Dans le travail réalisé, presque aucune distinction n'était faite entre le concept et le signe (sous la forme d'un groupe nominal), notamment parce que travailler au niveau d'un domaine permet de restreindre fortement le sens qu'un signe peut prendre. Il convient alors de considérer les concepts plus sérieusement en intégrant la polysémie.

Les médiations culturelles produites s'expriment uniquement en matière d'équivalence ou non équivalence. Cependant, des médiations plus complexes basées sur une logique plus forte peuvent certainement être produites.

Le travail réalisé sur la conception d'un processus ethnographique semi-automatique ouvre des perspectives méthodologiques en anthropologie cognitive pour étudier de

manière plus fine l'évolution des modèles culturels d'une communauté.

Finalement, cette contribution participe aux efforts effectués pour le développement de systèmes culturellement conscients, qui s'inscrivent dans ceux pour la conception des systèmes culturellement intelligents. Un aspect qu'il est alors important d'explorer est l'utilisation possible d'une conscience culturelle artificielle émique dans l'enculturation des systèmes.

Bibliographie

Rafi Ahmed, Philippe DeSmedt, Weimin Du, William Kent, Mohammad A. Ketabchi, Witold A Litwin, Abbas Rafii et M-C Shan : The pegasus heterogeneous multidatabase system. *Computer*, 24(12):19–27, 1991.

Patricia A Alexander, Diane L Schallert et Victoria C Hare : Coming to terms : How researchers in learning and literacy talk about knowledge. *Review of educational research*, 61(3):315–343, 1991.

Felix-Robinson Aschoff, Franz Schmalhofer et Ludger Van Elst : Knowledge mediation : A procedure for the cooperative construction of domain ontologies. *In International Conference on Knowledge Engineering and Knowledge Management*, pages 506–508. Springer, 2004.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak et Zachary Ives : Dbpedia : A nucleus for a web of open data. *The semantic web*, pages 722–735, 2007.

Collin F Baker, Charles J Fillmore et John B Lowe : The berkeley framenet project. *In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998.

Will Baker : From cultural awareness to intercultural awareness : Culture in elt. *ELT journal*, page ccr017, 2011.

The World Bank : Thinking with mental models, 2015.

Marco Baroni et Alessandro Lenci : How we blessed distributional semantic evaluation. *In Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10. Association for Computational Linguistics, 2011.

Peter Bartalos et Maria Bielikova : An approach to objectontology mapping. *In IIT, SRCâASStudent Research Conference, Bratislava, Slovakia*, pages 9–16, 2007.

Frederic C Bartlett : Remembering : An experimental and social study. *Cambridge : Cambridge University*, 1932.

John A Bateman, Bernardo Magnini et Giovanni Fabris : The generalized upper model knowledge base : Organization and use. *Towards very large knowledge bases*, pages 60–72, 1995.

Gregory Bateson : A theory of play and fantasy. *Psychiatric research reports*, 1955.

Carlo Batini, Maurizio Lenzerini et Shamkant B. Navathe : A comparative analysis of methodologies for database schema integration. *ACM computing surveys (CSUR)*, 18(4):323–364, 1986.

Giovanni Bennardo et Victor C De Munck : *Cultural models : Genesis, methods, and experiences*. Oxford University Press, 2014.

Kati Berninger, Daniel Kneeshaw et Christian Messier : The role of cultural models in local perceptions of sfm–differences and similarities of interest groups from three boreal regions. *Journal of environmental management*, 90(2):740–751, 2009.

EMMANUEL G Blanchard et RIICHIRO Mizoguchi : Designing culturally-aware tutoring systems with mauoc, the more advanced upper ontology of culture. *Research and Practice in Technology Enhanced Learning*, 9(1):41–69, 2014.

Emmanuel G Blanchard, Riichiro Mizoguchi et Susanne P Lajoie : Structuring the cultural domain with an upper ontology of culture. *The Handbook of Research on Culturally-Aware Information Technology : Perspectives and Models*, pages 179–212, 2010.

Emmanuel G Blanchard, Marguerite Roy, Susanne P Lajoie et Claude Frasson : An evaluation of sociocultural data for predicting attitudinal tendencies. *In AIED*, pages 399–406, 2009.

Benjamin Blount : Situating cultural models in history and cognition. *In approaches to language, culture, and cognition*, pages 271–298. Springer, 2013.

Stephen P Borgatti : Anthropac 4.983/x. *Columbia, South Carolina : Analytic Technologies*, 1992.

Stephen P Borgatti, Martin G Everett et Linton C Freeman : Ucinet 6.0 version 1.00. *Natick : Analytic Technologies*, 1999.

Willem Nico Borst : *Construction of engineering ontologies for knowledge sharing and reuse*. Universiteit Twente, 1997.

Janet E Burge : Knowledge elicitation tool classification. *Artificial Intelligence Research Group, Worcester Polytechnic Institute*, 2001.

Michael Byram : *Teaching and assessing intercultural communicative competence*. Multilingual Matters, 1997.

Ruth MJ Byrne : *The rational imagination*. MIT Press, 2005.

Diego Calvanese, Giuseppe De Giacomo et Maurizio Lenzerini : Ontology of integration and integration of ontologies. *Description Logics*, 49(10-19):30, 2001.

SA Caraballo : *Automatic Acquisition of a Hypernym-Labeled Noun Hierarchy from Text. Brown University Ph. D.* Thèse de doctorat, Thesis, 2001.

D Douglas Caulkins : Identifying culture as a threshold of shared knowledge a consensus analysis method. *International Journal of Cross Cultural Management*, 4(3):317–333, 2004.

Scott Cederberg et Dominic Widdows : Using lsa and noun coordination information to improve the precision and recall of automatic hyponymy extraction. *In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 111–118. Association for Computational Linguistics, 2003.

Krishna Chandramouli, Craig Stewart, Tim Brailsford et Ebroul Izquierdo : Cae-l : An ontology modelling cultural behaviour in adaptive education. *In Semantic Media Adaptation and Personalization, 2008. SMAP'08. Third International Workshop on*, pages 183–188. IEEE, 2008.

Balakrishnan Chandrasekaran, John R Josephson et V Richard Benjamins : What are ontologies, and why do we need them ? *IEEE Intelligent Systems and their applications*, 14(1):20–26, 1999.

Vinay K Chaudhri, Adam Farquhar, Richard Fikes, Peter D Karp et James P Rice : Open knowledge base connectivity 2.0. *Knowledge Systems Laboratory*, 1998.

Michelene TH Chi : Three types of conceptual change : Belief revision, mental model transformation, and categorical shift. *International handbook of research on conceptual change*, pages 61–82, 2008.

Krzysztof J Cios, Witold Pedrycz et Roman W Swiniarski : Data mining and knowledge discovery. *In Data Mining Methods for Knowledge Discovery*, pages 1–26. Springer, 1998.

Allan Collins et Dedre Gentner : How people construct mental models. *Cultural models in language and thought*, 243, 1987.

William A Corsaro et David R Heise : Event structure models from ethnographic data. *Sociological methodology*, pages 1–57, 1990.

Hernani Pereira Costa, Hugo Gonçalo Oliveira et Paulo Gomes : Using the web to validate lexico-semantic relations. *In Portuguese Conference on Artificial Intelligence*, pages 597–609. Springer, 2011.

K.J.W Craik : *The nature of explanation.* Cambridge University Press, Cambridge, UK, 1943.

Alan Cruse : Meaning in language : An introduction to semantics and pragmatics. 2011.

Roy D'Andrade : *A folk model of the mind.* Cambridge University Press, 1987.

Roy G D'Andrade : Character terms and cultural models. *Directions in cognitive anthropology*, pages 321–343, 1985.

Roy G D'Andrade : *The development of cognitive anthropology.* Cambridge University Press, 1995.

Dmitry Davidov et Ari Rappoport : Classification of semantic relationships between nominals using pattern clusters. *In ACL*, pages 227–235, 2008.

Jos De Bruijn, Marc Ehrig, Cristina Feier, Francisco Martìn-Recuerda, François Scharffe et Moritz Weiten : Ontology mediation, merging and aligning. *Semantic web technologies*, pages 95–113, 2006.

Victor C De Munck : Cognitive approaches to the study of romantic love : Semantic, cross-cultural, and as a process. *A companion to cognitive anthropology*, pages 513–530, 2011.

Victor C De Munck : A theory explaining the functional linkage between the self, identity and cultural models. *Journal of cognition and culture*, 13(1-2):179–200, 2013.

Leslie A DeChurch et Jessica R Mesmer-Magnus : Measuring shared team mental models : A meta-analysis., 2010.

James K Doyle : *Measuring change in mental models of dynamic systems : An exploratory study*. Thèse de doctorat, Department of Economics, Siena College, 1998.

Roy D'andrade : A cognitivist's view of the units debate in cultural anthropology. *Cross-cultural research*, 35(2):242–257, 2001.

Michael Eraut : Non-formal learning and tacit knowledge in professional work. *British journal of educational psychology*, 70(1):113–136, 2000.

Rong-En Fan, Pai-Hsuen Chen et Chih-Jen Lin : Working set selection using second order information for training support vector machines. *Journal of machine learning research*, 6(Dec):1889–1918, 2005.

Usama M Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth et Ramasamy Uthurusamy : *Advances in knowledge discovery and data mining*, volume 21. AAAI press Menlo Park, 1996.

G Ferraro : The cultural dimension of international business. (3rd Edition), 1998.

William R Ferrell : Discrete subjective probabilities and decision analysis : Elicitation, calibration and combination. 1994.

Howard Freeman, A Romney, Joao Ferreira-Pinto, Robert Klein et Tom Smith : Guatemalan and us concepts of success and failure. *Human Organization*, 40(2):140–145, 1981.

Stephanie A Fryberg et Hazel Rose Markus : Cultural models of education in american indian, asian american and european american contexts. *Social Psychology of Education*, 10(2):213–246, 2007.

Hector Garcia-Molina, Yannis Papakonstantinou, Dallan Quass, Anand Rajaraman, Yehoshua Sagiv, Jeffrey Ullman, Vasilis Vassalos et Jennifer Widom : The tsimmis approach to mediation : Data models and languages. *Journal of intelligent information systems*, 8(2):117–132, 1997.

Linda C Garro : Remembering what one knows and the construction of the past : A comparison of cultural consensus theory and cultural schema theory. *Ethos*, 28(3):275–319, 2000.

John B Gatewood et John W Lowe : Employee perceptions of credit unions : Implications for member profitability. *Madison, WI : Filene Research Institute*, 2008.

Tatiana Gavrilova et Tatiana Andreeva : Knowledge elicitation techniques in a knowledge management context. *Journal of Knowledge Management*, 16(4):523–537, 2012.

James Paul Gee : Video games and embodiment. *Games and Culture*, 3(3-4):253–263, 2008.

Michael R Genesereth, Richard E Fikes *et al.* : Knowledge interchange format-version 3.0 : reference manual. 1992.

Dedre Gentner et Donald R Gentner : Flowing waters or teeming crowds : Mental models of electricity. Rapport technique, DTIC Document, 1982.

Roxana Girju, Adriana Badulescu et Dan Moldovan : Learning semantic constraints for the automatic discovery of part-whole relations. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 1–8. Association for Computational Linguistics, 2003.

Roxana Girju, Adriana Badulescu et Dan Moldovan : Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1):83–135, 2006.

Roxana Girju, Dan I Moldovan *et al.* : Text mining for causal relations. *In FLAIRS Conference*, pages 360–364, 2002.

Ward Hunt Goodenough : *Culture, language, and society.* Benjamin-Cummings Pub Co, 1981.

Michael Granitzer, Vedran Sabol, Kow Weng Onn, Dickson Lukose et Klaus Tochtermann : Ontology alignment—a survey with focus on visually supported semi-automatic techniques. *Future Internet*, 2(3):238–258, 2010.

Clarence C Gravlee : Research design and methods in medical anthropology. *A Companion to Medical Anthropology. Merrill Singer and Pamela I. Erickson, eds*, pages 69–91, 2011.

Thomas R Gruber : Toward principles for the design of ontologies used for knowledge sharing ? *International journal of human-computer studies*, 43(5-6):907–928, 1995.

Nicola Guarino : Some organizing principles for a unified top-level ontology. *In AAAI Spring Symposium on Ontological Engineering*, pages 57–63, 1997.

Nicola Guarino : Some ontological principles for designing upper level lexical resources. *arXiv preprint cmp-lg/9809002*, 1998.

Asunciòn Gòmez-Pérez et Richard Benjamins : Overview of knowledge sharing and reuse components : Ontologies and problem-solving methods. IJCAI and the Scandinavian AI Societies. CEUR Workshop Proceedings, 1999.

Joachim Hammer et Dennis McLeod : An approach to resolving semantic heterogeneity in a federation of autonomous, heterogeneous database systems. *International Journal of Intelligent and Cooperative Information Systems*, 2(01):51–83, 1993.

Zellig S Harris : Distributional structure. *Word*, 10(2-3):146–162, 1954.

M Hearst : Wordnet : An electronic lexical database and some of its applications, 1998.

Marti A Hearst : Automatic acquisition of hyponyms from large text corpora. *In Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics, 1992.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid ò Séaghdha, Sebastian Padò, Marco Pennacchiotti, Lorenza Romano et Stan Szpakowicz : Semeval-2010 task 8 : Multi-way classification of semantic relations between pairs of nominals. *In Proceedings of the Workshop on Semantic Evaluations : Recent Achievements and Future Directions*, pages 94–99. Association for Computational Linguistics, 2009.

Iris Hendrickx, Roser Morante, Caroline Sporleder et Antal Van Den Bosch : Ilk : Machine learning of semantic relations with shallow features and almost no data. *In Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 187–190. Association for Computational Linguistics, 2007.

Geert Hofstede et Michael H Bond : Hofstede's culture dimensions an independent validation using rokeach's value survey. *Journal of cross-cultural psychology*, 15(4):417–433, 1984.

Geert H Hofstede et Geert Hofstede : *Culture's consequences : Comparing values, behaviors, institutions and organizations across nations*. Sage, 2001.

Dorothy Holland et Naomi Quinn : *Cultural models in language and thought*. Cambridge University Press, 1987.

Robert J House, Paul J Hanges, Mansour Javidan, Peter W Dorfman et Vipin Gupta : *Culture, leadership, and organizations : The GLOBE study of 62 societies*. Sage publications, 2004.

W Lewis Johnson : A simulation-based approach to training operational cultural competence. 2010a.

W Lewis Johnson : Using immersive simulations to develop intercultural competence. *In Culture and computing*, pages 1–15. Springer, 2010b.

Philip Nicholas Johnson-Laird : *Mental models : Towards a cognitive science of language, inference, and consciousness*. Numéro 6. Harvard University Press, 1983.

Natalie Jones, Helen Ross, Timothy Lynam, Pascal Perez et Anne Leitch : Mental models : an interdisciplinary synthesis of theory and methods. *Ecology and Society*, 16 (1), 2011.

Natalie A Jones, Helen Ross, Timothy Lynam et Pascal Perez : Eliciting mental models : a comparison of interview procedures in the context of natural resource management. 2014.

David A Jurgens, Peter D Turney, Saif M Mohammad et Keith J Holyoak : Semeval-2012 task 2 : Measuring degrees of relational similarity. *In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 :*

Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation, pages 356–364. Association for Computational Linguistics, 2012.

Yannis Kalfoglou et Marco Schorlemmer : Ontology mapping : the state of the art. *The knowledge engineering review*, 18(01):1–31, 2003.

Aditya Kalyanpur, Daniel Jiménez Pastor, Steve Battle et Julian A Padget : Automatic mapping of owl ontologies into java. *In SEKE*, volume 4, pages 98–103. Citeseer, 2004.

Anne R Kearney et Stephen Kaplan : Toward a methodology for the measurement of knowledge structures of ordinary people the conceptual content cognitive map (3cm). *Environment and Behavior*, 29(5):579–617, 1997.

Willett Kempton, James S Boster et Jennifer A Hartley : *Environmental values in American culture*. MIT Press, 1996.

Nouf Khashman et Andrew Large : Investigating the design of arabic web interfaces using hofstede's cultural dimensions : A case study of government web portals. *In Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI*, 2013.

Christopher SG Khoo et Jin-Cheon Na : Semantic relations in information science. 2006.

Won Kim, Injun Choi, Sunit Gala et Mark Scheevel : On resolving schematic heterogeneity in multidatabase systems. *Distributed and Parallel Databases*, 1(3):251–279, 1993.

Michel Klein : Combining and relating ontologies : an analysis of problems and solutions. *In IJCAI-2001 Workshop on ontologies and information sharing*, pages 53–62. USA., 2001.

Alfred L Kroeber et Talcott Parsons : The concepts of culture and of social system. *American Sociological Review*, 23(5):582–583, 1958.

Alfred Louis Kroeber et Clyde Kluckhohn : Culture : A critical review of concepts and definitions. *Papers. Peabody Museum of Archaeology & Ethnology, Harvard University*, 1952.

Thomas K Landauer et Susan T Dumais : A solution to plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

Robert B Lees et Noam Chomsky : Syntactic structures. *Language*, 33(3 Part 1):375–408, 1957.

Witold Litwin, Leo Mark et Nick Roussopoulos : Interoperability of multiple autonomous databases. *ACM Computing Surveys (CSUR)*, 22(3):267–293, 1990.

Hugo Liu et Push Singh : Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004.

Jonathan Lomas : Research and evidence–based decision making. *Australian and New Zealand Journal of Public Health*, 21(5):439–441, 1997.

Kevin Lund et Curt Burgess : Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208, 1996.

Alexander Maedche, Boris Motik, Nuno Silva et Raphael Volz : Mafra—a mapping framework for distributed ontologies. *In International Conference on Knowledge Engineering and Knowledge Management*, pages 235–250. Springer, 2002.

Kateryna Maltseva et Roy D'Andrade : Multi-item scales and cognitive ethnography. *A companion to cognitive anthropology*, pages 153–170, 2011.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard et David McClosky : The stanford corenlp natural language processing toolkit. *In ACL (System Demonstrations)*, pages 55–60, 2014.

Aaron Marcus et Emilie West Gould : Crosscurrents : cultural dimensions and global web user-interface design. *interactions*, 7(4):32–46, 2000.

Samuel Mascarenhas et Ana Paiva : Creating virtual synthetic cultures for intercultural training. *In Third International Workshop on Culturally-Aware Tutoring Systems (CATS2010)*, page 25, 2010.

Raphael Mathevet, Michel Etienne, Tim Lynam et Coralie Calvet : Water management in the camargue biosphere reserve : insights from comparative mental models analysis. *Ecology and Society*, 16(1), 2011.

Holly F Mathews : Uncovering cultural models of gender from accounts of folktales. *In Finding Culture in Talk*, pages 105–155. Springer, 2005.

John E Mathieu, Tonia S Heffner, Gerald F Goodwin, Janis A Cannon-Bowers et Eduardo Salas : Scaling the quality of teammates' mental models : Equifinality and normative comparisons. *Journal of Organizational Behavior*, 26(1):37–56, 2005.

John E Mathieu, Tammy L Rapp, M Travis Maynard et Phillip M Mangos : Interactive effects of team and task shared mental models as related to air traffic controllers' collective efficacy and effectiveness. *Human Performance*, 23(1):22–40, 2009.

David Ricky Matsumoto : *Cultural influences on research methods and statistics*. Brooks/Cole Publishing Company, 1994.

Diana Maynard, Adam Funk et Wim Peters : Sprat : a tool for automatic semantic pattern-based ontology population. *In International conference for digital libraries and the semantic web, Trento, Italy*, 2009.

Marissa F McBride et Mark A Burgman : What is expert knowledge, how is such knowledge gathered, and how do we use it to address questions in landscape ecology ? *In Expert knowledge and its application in landscape ecology*, pages 11–38. Springer, 2012.

Mary A Meyer et Jane M Booker : *Eliciting and analyzing expert judgment : a practical guide*. SIAM, 2001.

Morgan Meyer : The rise of the knowledge broker. *Science communication*, 32(1):118–127, 2010.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado et Jeff Dean : Distributed representations of words and phrases and their compositionality. *In Advances in neural information processing systems*, pages 3111–3119, 2013.

George A Miller : The magical number seven, plus or minus two : some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

George A Miller : Wordnet : a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Nick R Milton : *Knowledge technologies*, volume 3. Polimetrica sas, 2008.

Riichiro Mizoguchi, Johan Vanwelkenhuysen et Mitsuru Ikeda : Task ontology for reuse of problem solving knowledge. *Towards Very Large Knowledge Bases : Knowledge Building & Knowledge Sharing*, 46:59, 1995.

Phaedra Mohammed et Emmanuel G Blanchard : Leveraging comparisons between cultural frameworks : Preliminary investigations of the mauoc ontological ecology. *In Sixth International Workshop on Culturally-Aware Tutoring Systems (CATS2015)*, page 1, 2015.

Phaedra Mohammed et Permanand Mohan : The design and implementation of an enculturated web-based intelligent tutoring system for computer science education. *In Advanced Learning Technologies (ICALT), 2011 11th IEEE International Conference on*, pages 501–505. IEEE, 2011.

Phaedra Mohammed et Permanand Mohan : Breakthroughs and challenges in culturally-aware technology enhanced learning. *In Proc. Workshop on Culturally-aware Technology Enhanced Learning in conjuction with EC-TEL 2013, Paphos, Cyprus, September 17*, 2013a.

Phaedra Mohammed et Permanand Mohan : Contextualised student modelling for enculturated systems. *In AIED 2013 Workshops Proceedings Volume 5*, page 20. Citeseer, 2013b.

Neville Moray : Models of models of... mental models. *Perspectives on the human controller*, pages 271–285, 1997.

Regina Motz, Jacqueline Guzmàn, Claudia Deco et Cristina Bender : Applying ontologies to educational resources retrieval driven by cultural aspects. *Journal of Computer Science & Technology*, 5, 2005.

M Lynne Murphy : *Semantic relations and the lexicon : antonymy, synonymy and other paradigms*. Cambridge University Press, 2003.

NCCC : National center for cultural competence, 2004. URL https://nccc.georgetown.edu/documents/Cultural_Broker_Guide_English.pdf.

Nancy J Nersessian : The cognitive basis of model-based reasoning in science. *The cognitive basis of science*, pages 133–153, 2002.

Natalya Fridman Noy, Mark A Musen *et al.* : Algorithm and tool for automated ontology merging and alignment. *In Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00). Available as SMI technical report SMI-2000-0831*, 2000.

Charles Kay Ogden, Ivor Armstrong Richards, Sv Ranulf et E Cassirer : The meaning of meaning. a study of the influence of language upon thought and of the science of symbolism, 1923.

Zita Oravecz, Joachim Vandekerckhove et William H Batchelder : Bayesian cultural consensus theory. *Field Methods*, 26(3):207–222, 2014.

Thomas Osborne : On mediators : Intellectuals and the ideas trade in the knowledge society. *Economy and Society*, 33(4):430–447, 2004.

Patrick Pantel et Dekang Lin : Discovering word senses from text. *In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–619. ACM, 2002.

Patrick Pantel et Deepak Ravichandran : Automatically labeling semantic classes. *In HLT-NAACL*, volume 4, pages 321–328, 2004.

Pertti J Pelto et Gretel H Pelto : intra-cultural diversity : some theoretical issues. *American ethnologist*, 2(1):1–18, 1975.

Marco Pennacchiotti et Patrick Pantel : Ontologizing semantic relations. *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 793–800. Association for Computational Linguistics, 2006.

RSV Pepijn, MJ Dean, TJM Bench-capon et M Shave : An analysis of ontological mismatches : Heterogeneity versus interoperability. *In Proceedings of AAAI Spring Symposium on Ontological Engineering, Stanford, USA*, 1997.

Jean Petit, Jean-Charles Boisson et Francis Rousseaux : Building time-affordable cultural ontologies using an emic approach. volume 497 de *IFIP Advances in Information and Communication Technology series*. Springer, 2017.

Jean Piaget et Margaret Cook : *The origins of intelligence in children*, volume 8. International Universities Press New York, 1952.

Kenneth L Pike : Language in relation to a unified theory of the structure of human behavior. 1954.

Michael Polanyi : *Personal knowledge : Towards a post-critical philosophy*. University of Chicago Press, 1958.

Reinhard Rapp : Word sense discovery based on sense descriptor dissimilarity. *In Proceedings of the ninth machine translation summit*, pages 315–322, 2003.

Katharina Reinecke et Abraham Bernstein : Tell me where you've lived, and i'll tell you what you like : Adapting interfaces to cultural preferences. *In International Conference on User Modeling, Adaptation, and Personalization*, pages 185–196. Springer, 2009.

Silja Renooij : Probability elicitation for belief networks : issues to consider. *The Knowledge Engineering Review*, 16(03):255–269, 2001.

A Kimball Romney, William H Batchelder et Susan C Weller : Recent applications of cultural consensus theory. *American Behavioral Scientist*, 31(2):163–177, 1987.

A Kimball Romney, Susan C Weller et William H Batchelder : Culture as consensus : A theory of culture and informant accuracy. *American anthropologist*, 88 (2):313–338, 1986.

Gun Roos : Pile sorting :"kids like candy.". *Using methods in the field : a practical introduction and casebook. Walnut Creek (CA) : AltaMira*, pages 97–110, 1998.

Eleanor Rosch et Barbara Bloom Lloyd : *Cognition and categorization*, volume 1. Citeseer, 1978.

David E Rumelhart : Notes on a schema for stories. *Representation and understanding : Studies in cognitive science*, 211(236):45, 1975.

Andrew Rutherford et John R Wilson : Models of mental models : an ergonomist-psychologist dialogue. *In Selected papers of the 8th Interdisciplinary Workshop on Informatics and Psychology : Mental Models and Human-Computer Interaction 2*, pages 39–58. North-Holland Publishing Co., 1989.

Benoît Sagot et Darja Fišer : Automatic extension of wolf. *In GWC2012-6th International Global Wordnet Conference*, 2012.

Roger C Schank et Robert P Abelson : Scripts, plans, goals, and understanding : An inquiry into human knowledge structures (artificial intelligence series). 1977.

E Schein : Organizational culture. american psychologist, 1990.

Edgar H Schein : Coming to a new awareness of organizational culture. *Sloan management review*, 25(2):3–16, 1984.

Helmut Schmid : Probabilistic part-of speech tagging using decision trees. *In New methods in language processing*, page 154. Routledge, 2013.

Shalom H Schwartz : *Beyond individualism/collectivism : New cultural dimensions of values*. Sage Publications, Inc, 1994.

N Shadbolt et P Smart : Knowledge elicitation : Methods, tools and techniques. *Evaluation of Human Work (4th ed.). CRC Press, Boca Raton, Florida, USA*, pages 163–200, 2015.

Farzad Sharifian : On cultural conceptualisations. *Journal of Cognition and Culture*, 3 (3):187–207, 2003.

Nakatani Shuyo : Language detection library for java. *Retrieved Jul*, 7:2016, 2010.

Pavel Shvaiko et Jérôme Euzenat : Ontology matching : state of the art and future challenges. *IEEE Transactions on knowledge and data engineering*, 25(1):158–176, 2013.

John F Sowa : Conceptual structures : information processing in mind and machine. 1983.

Helen Spencer-Oatey et P Franklin : What is culture. *A compilation of quotations. GlobalPAD Core Concepts*, 2012.

Rudi Studer, V Richard Benjamins et Dieter Fensel : Knowledge engineering : principles and methods. *Data & knowledge engineering*, 25(1-2):161–197, 1998.

William C Sturtevant : Studies in ethnoscience. *American Anthropologist*, 66(3):99–131, 1964.

Xiaomeng Su et Jon Atle Gulla : Semantic enrichment for ontology mapping. *In International Conference on Application of Natural Language to Information Systems*, pages 217–228. Springer, 2004.

Xiaomeng Su et Jon Atle Gulla : An information retrieval approach to ontology mapping. *Data & Knowledge Engineering*, 58(1):47–69, 2006.

Árni Sverrisson : Translation networks, knowledge brokers and novelty construction : Pragmatic environmentalism in sweden. *Acta Sociologica*, 44(4):313–327, 2001.

Leonard Talmy : How language structures space. *In Spatial orientation*, pages 225–282. Springer, 1983.

Gomer Thomas, Glenn R Thompson, Chin-Wan Chung, Edward Barkmeyer, Fred Carter, Marjorie Templeton, Stephen Fox et Berl Hartman : Heterogeneous distributed database systems for production use. *ACM Computing Surveys (CSUR)*, 22(3):237–266, 1990.

Barry Tomalin et Susan Stempleski : *Cultural awareness*. Oxford University Press, 2013.

Yannick Toussaint : *Fouille de textes : des méthodes symboliques pour la construction d'ontologies et l'annotation sémantique guidée par les connaissances*. Thèse de doctorat, Université Henri Poincaré-Nancy I, 2011.

Shigehisa Tsuchiya : Improving knowledge creation ability through organizational learning. *In ISMICK'93 Proceedings, International Symposium on the Management of Industrial and Corporate Knowledge*, pages 87–95, 1993.

Victor Witter Turner : *The forest of symbols : Aspects of Ndembu ritual*, volume 101. Cornell University Press, 1967.

Edward Burnett Tylor : *Primitive culture : researches into the development of mythology, philosophy, religion, art, and custom*, volume 2. Murray, 1871.

Gertjan Van Heijst, A Th Schreiber et Bob J Wielinga : Using explicit ontologies in kbs development. *International journal of human-computer studies*, 46(2-3):183–292, 1997.

Pepijn RS Visser, Dean M Jones, Trevor JM Bench-Capon et Michael JR Shave : Assessing heterogeneity by classifying ontology mismatches. *In Proceedings of the FOIS*, volume 98, 1998.

Carole Vuillot, Nadège Coron, François Calatayud, Clélia Sirami, Raphael Mathevet et Annick Gibon : Ways of farming and ways of thinking : do farmers' mental models of the landscape relate to their land management practices ? *Ecology and Society*, 21(1):1–23, 2016.

Wolfgang Wagner et Nicky Hayes : *Everyday discourse and common sense : The theory of social representations*. Palgrave Macmillan, 2005.

Ting Wang, Yaoyong Li, Kalina Bontcheva, Hamish Cunningham et Ji Wang : Automatic extraction of hierarchical relations from text. *In European Semantic Web Conference*, pages 215–229. Springer, 2006.

Stanley Wasserman et Katherine Faust : *Social network analysis : Methods and applications*, volume 8. Cambridge university press, 1994.

Susan C Weller : Shared knowledge, intracultural variation, and knowledge aggregation. *American Behavioral Scientist*, 31(2):178–193, 1987.

Susan C Weller : Cultural consensus theory : Applications and frequently asked questions. *Field methods*, 19(4):339–368, 2007.

Dominic Widdows et Beate Dorow : A graph model for unsupervised lexical acquisition. *In Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics, 2002.

Gio Wiederhold : An algebra for ontology composition. *In Proceedings of 1994 Monterey Workshop on Formal Methods*, pages 56–61, 1994.

Anna Wierzbicka : *English : Meaning and culture*. Oxford University Press, 2006.

K Wiig : People-focused knowledge management. how effective decision making leads to corporate success. 2004.

James C Young : A model of illness treatment decisions in a tarascan town. *American Ethnologist*, 7(1):106–131, 1980.

Annexe

Individu	Status	URL du site web	Facebook
Priests for Life	C	www.priestsforlife.org/	102906873084255
Pro-Life Action League	C	prolifeaction.org/	10150096009920008
National Right to Life	F	www.nrlc.org/	286781440146
SPUC	C	www.spuc.org.uk/	107582715996851
Pro Life Campaign	C	www.prolifecampaign.ie/	224615650378
Students for Life of America	F	www.studentsforlife.org/	79125112926
Secular Pro-Life	F	secularprolife.org/	156362631095762
Prolife NZ	C	www.prolife.org.nz/	135807799796790
PLAGAL	C	www.plagal.org/	99136188374
Americans United for Life	F	www.aul.org/	101095043361
Abort73	F	www.abort73.com/	7644864415
I am Pro-Life	C	www.focusonthefamily.com/pro-life	202600839762879
National Abortion Federation	C	www.prochoice.org/	149445681749910
Abortion Rights	C	www.abortionrights.org.uk/	251165918287722
NARAL Pro-Choice America	F	www.prochoiceamerica.org/	80562389320
Abortion Rights Coalition of Canada	F	www.arcc-cdac.ca/	127514140600625
Center for Reproductive Rights	F	www.reproductiverights.org/	73528878871
NARAL Pro-Choice Texas	C	www.prochoicetexas.org/	5998839406
NARAL Pro-Choice California	C	www.prochoicecalifornia.org/	103841584893
NARAL Pro-Choice Colorado	C	www.prochoicecolorado.org/	25115496086
NARAL Pro-Choice Connecticut	C	www.prochoicect.org/	115683471815619
NARAL Pro-Choice Maryland	C	www.prochoicemd.org/	19903890507
NARAL Pro-Choice Massachusetts	C	www.prochoicemass.org/	27550760394
NARAL Pro-Choice Minnesota	F	www.prochoiceminnesota.org/	145429001769
NARAL Pro-Choice Missouri	C	www.prochoicemissouri.org/	16643179251
NARAL Pro-Choice Montana	C	www.prochoicemontana.org/	92320384393
NARAL Pro-Choice North Carolina	C	www.prochoicenc.org/	96538134831
NARAL Pro-Choice Ohio	C	www.prochoiceohio.org/	205621732795876
NARAL Pro-Choice Oregon	F	www.prochoiceoregon.org/	73174731788

TABLE 5.1 – Détails sur les individus des échantillons

Titre – Contribution à la Modélisation d'une Conscience Culturelle Artificielle Émique par les Ontologies

Résumé – Depuis l'expansion du web, de nombreuses applications cherchent à répondre aux besoins d'utilisateurs ou de machines aux origines culturelles variées. De ce contexte de diversité culturelle émergent de nombreux conflits liés à des conceptions du monde différentes. Proposer des services adaptés requiert l'intégration au sein du système d'une forme de conscience culturelle. Une conscience culturelle artificielle est composée de représentations et de médiations culturelles formelles offrant au système les moyens pour interpréter les cultures représentées et déterminer leurs différences. Jusqu'à présent les représentations utilisées dans le développement des systèmes culturellement conscients sont issues de modèles universels ou « étiques ». Ces modèles grossiers, bien qu'ils soient adaptés, limitent la compréhension possible des cultures représentées. Par conséquent ils constituent un goulot d'étranglement dans le développement des systèmes culturellement conscients.

Cette thèse explore le développement d'une conscience culturelle artificielle plus fine sur la base de modèles culturels dits « émiques », c'est-à-dire spécifiques à chaque culture. J'étudie la construction, la formalisation et la médiation des représentations culturelles émiques. Mes contributions principales sont la conception et la validation, d'une part, d'un nouveau processus ethnographique semi-automatique de construction de modèles émiques via la fouille de textes et, d'autre part, d'une conscience culturelle artificielle émique fondée sur l'alignement d'ontologies culturelles issues de ces modèles.

Mots clefs – système culturellement conscient, conscience culturelle artificielle émique, modèle culturel prototypique, relation sémantique, ontologie culturelle, médiation culturelle.

Title – Contribution to the Modeling of Emic Artificial Cultural Awareness through Ontologies

Abstract – With the growing web, a number of applications seek to meet the needs of users or machines having diverse cultural backgrounds. From this context of cultural diversity arises conflicts linked to different world conceptions. Offering adaptated services requires the integration of a form of cultural awareness in the system. An artificial cultural awareness is composed of formal cultural representations and mediations providing the system with the means to interpret the represented cultures and to determine their differences. So far the representations used for the development of culturally-aware systems come from universal or "etic" models. Those coarse-grained models, even though they are adapted, limit the possible understanding of the represented cultures. As a consequence they constitute a bottleneck for the development of culturally-aware systems.

This thesis investigates the development of a finer-grained artificial cultural awareness based on cultural models called "emic", in other words specific to each culture. I study the construction, the formalisation and the mediation of these emic cultural representations. My main contributions are the design and validation of, on the one hand, a new semi-automatic ethnographic process for building emic models through text-mining, on the other hand, an emic artificial cultural awareness based on the mapping of cultural ontologies coming from those models.

Keywords – culturally-aware system, emic artificial cultural awareness, prototypical cultural model, semantic relation, cultural ontology, cultural mediation.

Discipline – Informatique, ingénierie des connaissances, fouille de données, fouille de textes, apprentissage de relations lexico-sémantiques, ontologies, alignement d'ontologies.
Pluridisciplinarité – Anthropologie, sciences cognitives, anthropologie cognitive, traitement de la langue.

CReSTIC EA 3804,
Moulin de la Housse,
51687 Reims, France