Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Sommaire

L'évolution des technologies et de la capacité de nos systèmes a permis de produire et stocker de plus en plus de données, sous des formats très variés, avec des moyens d'une grande efficacité. Aujourd'hui, nous pouvons déclarer sans conteste que la première et plus grande source d'information disponible est le web. Le web a rapidement connu un essor sans précédent tant dans le nombre de sites présents que dans la diversité et la richesse de leur contenu. Les articles de journaux, les billets de blogs, les messages et conversations sur les réseaux sociaux, les bibliothèques spécialisées sont autant de ressources disponibles, comportant des données intéressantes à analyser pour en extraire de la connaissance.
Dans le cadre de la thèse, nous nous intéressons plus particulièrement aux publications scientifiques disponibles à partir de bibliothèques spécialisées comme Wiley, Elsevier ou Springer qui mettent en ligne de nombreux articles, donnant accès à de nombreuses informations représentant une source intarissable pour les scientifiques de nombreux domaines. La communauté scientifique a désormais la possibilité de partager des informations et d'accéder à de nouvelles informations à travers les documents publiés et stockés dans les bases en ligne du web. Toutefois, malgré ces avancées prodigieuses, de nouveaux verrous doivent être levés pour pouvoir analyser les données. La valorisation de l'information reste un défi majeur car, ne l'oublions pas, une information non utilisée, non analysée n'a finalement pas grande valeur. Cette problématique rassemble plusieurs disciplines depuis plusieurs années dont la discipline du Traitement Automatique des Langues (TAL) et

celles de l'Ingénierie des Connaissances (IC).
Le TAL s'intéresse à la découverte de méthodes pour reconnaitre l'information pertinente en fonction de son expression dans les textes et à leur extraction dans les documents. En effet, identifier et extraire l'information pertinente se révèlent être des tâches complexes car la grande majorité des documents collectés sur le web est, en général, partagée au format textuel non structuré et en langage naturel. Le langage naturel, du fait de sa richesse et de sa variété est souvent difficile à appréhender. Un même mot revêt plusieurs significations, une même information peut s'exprimer de multiples manières, souvent implicitement, générant des ambiguïtés difficiles à cerner automatiquement par les machines. L'IC s'intéresse à la capitalisation de la connaissance, par exemple, à partir de l'approche ontologique et du web sémantique pour définir un vocabulaire commun partagé. Le rôle de l'ontologie est double car, d'une part, elle définit les concepts d'un domaine et, d'autre part, elle fournit une sémantique formelle aux concepts, indispensable à l'automatisation de l'intégration des données. Les données extraites et ainsi structurées au sein d'ontologies sont valorisées, peuvent être partagées sur le web de données et sont disponibles à l'interrogation et à l'analyse automatique par les agents logiciels sillonnant le web.

1.1 Cadre général

Les publications scientifiques, disponibles à partir de bibliothèques spécialisées en ligne, sont une source d'information précieuse à exploiter et analyser par les experts du domaine pour, par exemple, paramétrer des modèles d'aide à la décision. Le nombre d'articles publiés et disponibles en ligne est toujours grandissant. Aujourd'hui, le défi n'est pas de trouver l'information mais d'être en mesure de l'identifier et l'extraire automatiquement. En effet, la problématique liée aux documents textuels réside essentiellement dans le fait que les données se trouvent dans un format libre et sont, par conséquent, complexes à identifier et extraire automatiquement. Les travaux de recherche, menés dans le domaine de l'extraction d'information, portent sur le développement de méthodes innovantes pour permettre l'extraction automatique de l'information pertinente des documents. Leur représentation sous des formats structurés permet de capitaliser et partager la connaissance du domaine.
Depuis de nombreuses années, les travaux en recherche d'information sont menés dans le cadre des conférences MUC (Message Understanding Conference), qui organisent et favorisent les travaux en extraction d'information, en proposant un cadre d'évaluation, à partir de corpus textuels annotés et analysés, mis à disposition des équipes de recherche confrontant leurs méthodes. Les travaux ont concerné, par exemple pour la conférence MUC-6 [1], l'extraction d'entités nommées (lieux, noms de personnes, organisations...) et

1. http ://www.cs.nyu.edu/cs/faculty/grishman/muc6.html

de coréférences. La conférence MUC-7 [2], dans la continuité des travaux de recherche en cours du MUC-6, s'intéresse également à l'extraction de relations entre entités.
Dans le cadre des travaux menés sur l'extraction automatique de relations dans les documents, nous trouvons deux grands axes de recherche, l'extraction des relations binaires et l'extraction des relations n-aires, dont les travaux à l'état de l'art sont cités dans (Bach and Badaskar, 2007; Zhou et al., 2014).
Les travaux de recherche concernant l'extraction de relations binaires ont connu un vif essor dans le domaine de la biologie médicale. Ce domaine s'y intéresse particulièrement, notamment pour découvrir les interactions entre gènes et entre protéines, à la base des découvertes de nouvelles thérapies. L'enjeu n'est pas tant l'identification des entités nommées, mais plutôt l'identification des relations que vont entretenir certaines entités. En effet, dans le domaine biomédical, la terminologie de spécialité constitue un ensemble que l'on peut considérer comme étant un ensemble fini et les méthodes développées pour structurer et enrichir les ressources sont extrêmement riches. La communauté peut se reposer sur ces systèmes robustes de connaissance, e.g UMLS (Zweigenbaum, 2004).
Nous nous sommes également intéressés aux travaux menés en chimie où les problématiques d'extraction de relations, représentant une expérience impliquant deux molécules chimiques est également un enjeu majeur. Dans ce domaine, les travaux menés (Hawizy et al., 2011; Akhondi et al., 2015) utilisent l'identification du verbe principal de la phrase comme déclencheur de la relation binaire entre molécules chimiques.

Les méthodes utilisées pour découvrir des relations binaires sont essentiellement de deux types et sont illustrées à partir du schéma représentatif 1.1. Le schéma montre, à partir d'une phrase présentant une relation binaire entre deux protéines, que les méthodes reposent sur la construction de patrons linguistiques à partir de règles syntaxiques, et, sur des méthodes d'apprentissage, notamment des méthodes de classification.

- Les méthodes à base de règles consistent à produire un ensemble de règles écrites sous forme d'expressions régulières, intégrant certains mots ou étiquettes grammaticales obtenues en préparant les textes à partir de l'analyse des partis du discours ou en anglais *Part Of Speech (POS)*. Ces règles peuvent être écrites smanuellement ou apprises à partir de données d'entraînement pour construire les patrons linguistiques d'extraction des relations binaires recherchées. Ces méthodes nécessitent une forte implication humaine et ont un faible taux de recouvrement sur toutes les expressions possibles décrivant les interactions recherchées ;
- Les méthodes fondées sur l'apprentissage utilisent la classification binaire pour identifier les relations. Le modèle appris permet de déterminer si une nouvelle phrase

2. http ://www.aclweb.org/anthology/M/M98/

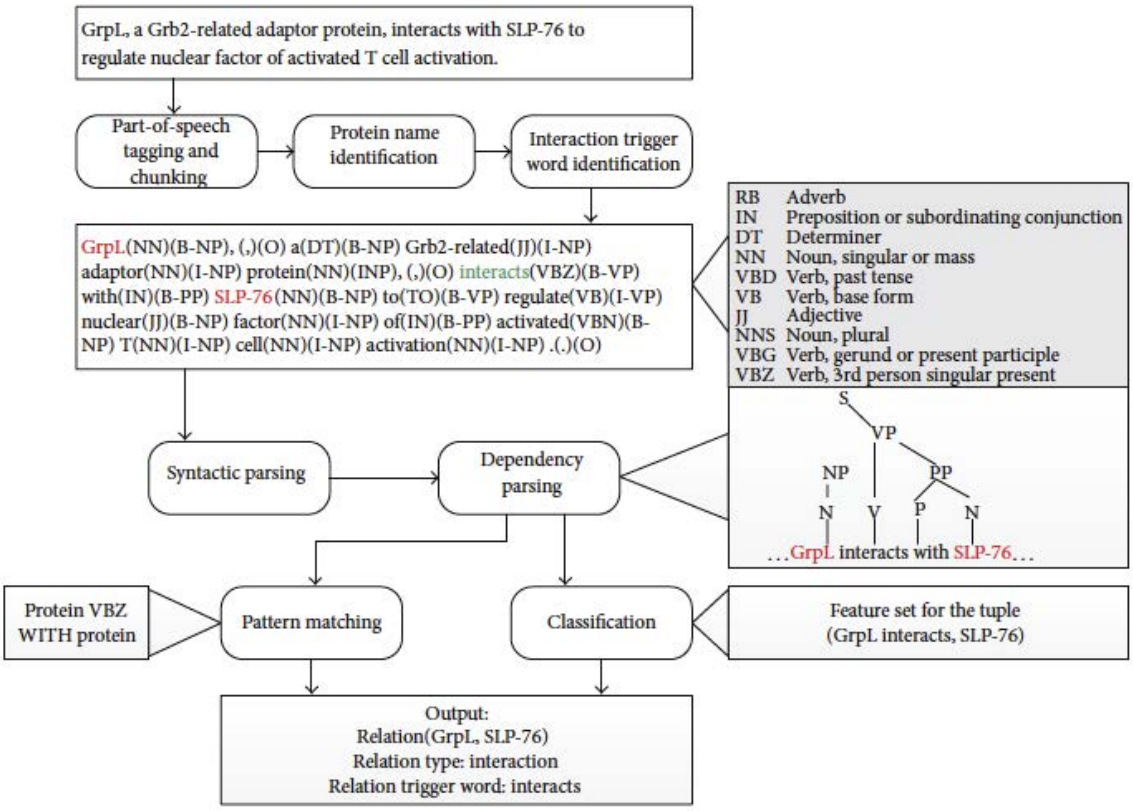


FIGURE 1.1: Méthodologies employées pour l'extraction de relations binaires dans le domaine biomédical (Zhou et al., 2014).

contient l'interaction recherchée ou pas. Le processus d'apprentissage repose sur un ensemble d'attributs variés : des attributs sémantiques (i.e. des entités de domaine), des attributs issus de l'analyse syntaxique ou de l'étiquetage grammatical du POS. Les méthodes par apprentissage supervisé obtiennent de très bons résultats mais requièrent néanmoins un grand nombre de données annotées pour l'apprentissage.

Les travaux de (Charnois et al., 2009) proposent une méthode basée sur la fouille de données séquentielles afin d'extraire des informations concernant les interactions d'un gène avec d'autres gènes, sous forme de relations binaires, et tentent ainsi de répondre à la question : *"Avec quel(s) gène(s), le gène X interagit-il ? et sous quelle forme ?"*. Les travaux menés proposent donc d'extraire, à partir d'un ensemble de phrases reconnues comme contenant des interactions entre gènes et dont les gènes ont été étiquetés par un expert, les motifs séquentiels fréquents. Cette approche nous intéresse particulièrement pour l'adapter à l'extraction des relations n-aires dispersées dans plusieurs phrases d'un corpus brut (non annoté et non analysé).

Les travaux de recherche concernant les relations complexes (ou relations n-aires), c'est-à-dire des relations faisant intervenir plus de deux arguments, est un problème beaucoup plus compliqué à résoudre. Tout d'abord, la relation n-aire fait intervenir plusieurs arguments de différents types. Ces arguments à regrouper dans la relation à identifier, peuvent se trouver dans une phrase ou dans l'ensemble du document ou également dans des tableaux. Chaque cas nécessite une méthode spécifique, d'autant que l'expression de la relation ou des arguments formant la relation peut être explicite ou implicite dans le texte.

Dans le domaine de la Biologie médicale, (Mcdonald et al., 2005) sont les premiers à s'intéresser à la problématique de l'extraction des relations complexes. Leurs travaux proposent une méthode hybride combinant apprentissage et règles appliquées sur une représentation en graphe. Le système propose d'identifier une relation n-aire phrastique comportant 3 arguments. Chaque relation binaire entre arguments de la relation est identifiée par classification binaire, puis, un graphe d'entités est construit où les noeuds représentent les relations binaires identifiées, comme cela est représenté dans la figure 1.2, empruntée à (Ludovic, 2011). La relation complexe est construite en sélectionnant la clique maximale dans le graphe des relations. La clique du graphe des relations correspond à un sous-graphe complet du graphe des relations qui permet de reconstituer les relations complexes avec les 3 arguments.
De manière générale, comme le montre l'illustration 1.3, les méthodes à l'état de l'art (Zhou et al., 2014) d'extraction de relations n-aires dans les documents suivent l'architecture globale suivante composée de trois étapes fondamentales :

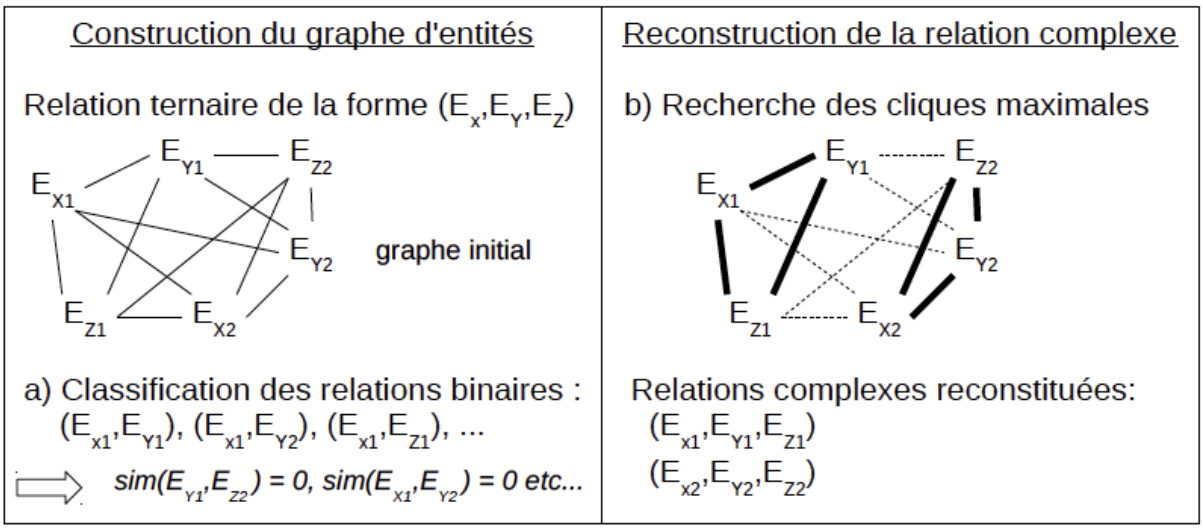


FIGURE 1.2: Principe de reconstitution des relations complexes avec la clique maximale

- La première étape consiste à identifier les entités nommées. Cette phase ne pose, *a priori*, pas de problèmes majeurs sachant qu'il existe des dictionnaires ou des bases de connaissances contenant la terminologie du domaine d'intérêt. Nous disons "*a priori*" car cette tâche peut également soulever des difficultés lorsqu'il s'agit d'identifier les variations typographiques des entités que nous rencontrons dans tous les domaines de spécialité. Par exemple, en Biologie, les protéines peuvent être rédigées selon une nomenclature différente, composée de sigles variés. L'identification devient alors plus complexe nécessitant des étapes d'enrichissement des sources de connaissances (dictionnaires, ontologies...),
- La seconde étape consiste à trouver l'élément déclencheur de la relation. Cette phase est une étape clé dans le processus et soulève de nombreuses ambiguïtés car le choix de l'élément déclencheur n'est pas trivial. Des méthodes à base de règles ou d'apprentissage supervisé sont également utilisées afin d'identifier l'élément déclencheur dans la relation,
- La troisième étape consiste à relier les arguments de la relation autour de l'élément déclencheur. Les méthodes sont également à base de règles ou d'apprentissage. Les meilleurs résultats obtenus sont ceux à base d'apprentissage utilisant des attributs proches de l'élément déclencheur et des entités identifiées dans le processus global.

Dans le domaine de la géographie, une méthode a été proposée par (Nguyen et al.,

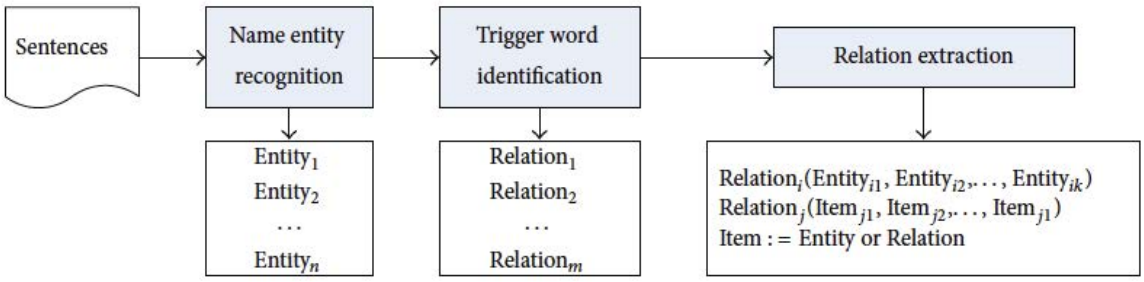


FIGURE 1.3: Architecture globale pour l'extraction des relations n-aires.

2010) fondée uniquement sur l'analyse syntaxique des phrases. Plusieurs patrons linguistiques sont construits, pour les 4 cas de relations n-aires définies par le W3C (Fridman Noy and Rector, 2006) et susceptibles d'être rencontrées dans le contexte de descriptions textuelles d'itinéraires.

- une relation à laquelle il faut ajouter un attribut décrivant l'instance de relation et qui aura des liens avec tous les participants, correspondant à des informations complémentaires de cette instance, comme illustré dans la figure 1.4. Cette relation est vue comme une relation binaire.

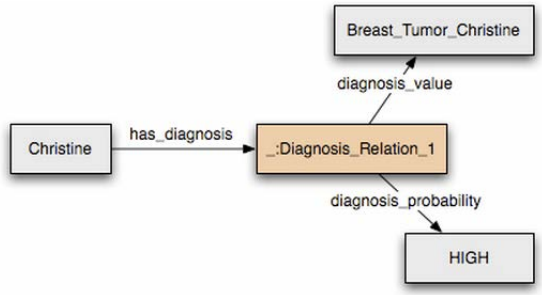


FIGURE 1.4: Cas 1 de représentation de relation n-aire à partir de la phrase *Christine has breast tumor with high probability*

- une relation à laquelle on relie un individu et les faits représentant un attribut de la relation. Dans l'exemple de la figure 1.5, l'instance de relation relie l'individu Steve et les différents attributs de sa température

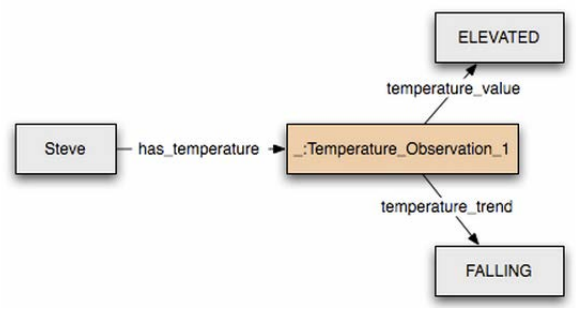


FIGURE 1.5: Cas 2 de représentation de relation n-aire à partir de la phrase *Steve has temperature, which is high, but falling*

- une relation qui relie des individus ayant des rôles différents sans qu'aucun d'entre eux ne joue le rôle central de "sujet" de la relation, comme cela est illustré dans la figure 2.1

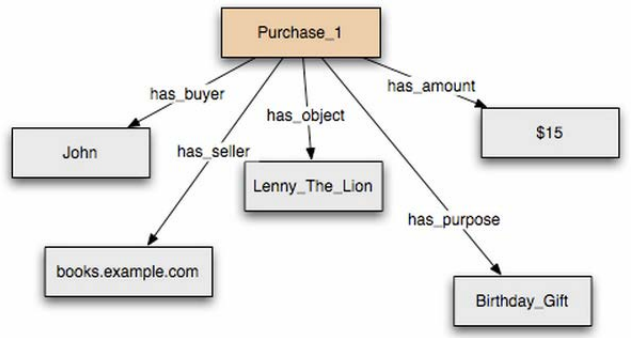


FIGURE 1.6: Cas 3 de représentation de relation n-aire à partir de la phrase *John buys a Lenny the Lion book from books.example.com for $ 15 as a birthday gift*

- une relation entre une liste d'arguments ordonnés et reliés selon la séquence en question, figure 1.7

La méthode suppose que les phrases à traiter comportent un verbe de placement. La méthode permet ensuite de générer 8 patrons pour le cas 1 de représentation des relations n-aires, 2 patrons pour le cas 2, 4 patrons pour le cas 3 et 2 patrons pour le cas 4. L'auteur précise que la base de patrons est construite de manière empirique en observant des échantillons de phrases qui comportent des relations n-aires annotées manuellement. Les résultats sont communiqués en termes d'extraction terminologique (le nombre de termes nouveaux extraits pour enrichir une ontologie) avec l'apport des relations n-aires, définies

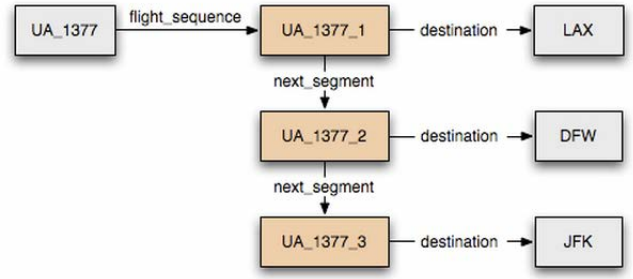


FIGURE 1.7: Cas 4 de représentation de relation n-aire à partir de la phrase *United Airlines flight 3177 visits the following airports : LAX, DFW, and JFK*

en patrons selon les cas, avec une précision entre 0.54 et 0.59 selon les patrons utilisés. La méthode proposée par les auteurs s'applique à l'extraction terminologique dans des relations n-aires qui se situent dans la phrase, or, comme nous l'avons évoqué plus tôt, il est assez fréquent dans les documents scientifiques de rencontrer des relations n-aires qui s'étendent sur plusieurs phrases.

L'identification des relations n-aires dans le domaine médical est abordée dans les travaux de (Grau et al., 2009; Minard et al., 2013). La méthode proposée est évaluée sur un corpus de 20 documents, convertis au format XML et annotés en projetant les termes et relations présentes dans une base de données QKDB (*Quantitative Kidney DataBase*). Le corpus est ensuite validé et complété manuellement si nécessaire. La méthode d'extraction s'appuie sur trois étapes : détection de la valeur numérique du résultat, reconnaissance des descripteurs d'expérience (reconnaissance terminologique à partir de la base de données) et, mise en relation des attributs ou descripteurs avec la valeur numérique, en prenant en compte la distance entre la valeur numérique du résultat et les descripteurs ou attributs, ainsi que des critères de fréquence dans l'article complet. Le système général obtient une précision de 0.51, avec une précision de 0.63 pour l'extraction des résultats quantitatifs. À partir de l'extraction des résultats quantitatifs correctement extraits de 63%, l'extraction des résultats expérimentaux obtient une précision de 0.81. Les résultats obtenus servent de référence à notre travail qui s'inscrit dans la même démarche d'extraction des arguments de relations n-aires, représentant également des résultats expérimentaux engageant des arguments symboliques et des arguments quantitatifs. Notre démarche consiste à proposer une approche qui permette de découvrir les liens implicites qui existent entre les arguments de la relation n-aire afin de désambiguiser leur mise en relation. Dans cette nouvelle approche, les arguments ne sont donc plus considérés comme indépendants au sein de la relation n-aire.

Les méthodes pour l'identification et l'extraction des relations binaires proposées à l'état de l'art montrent une certaine maturité et révèlent des résultats plutôt satisfaisants. En revanche, la problématique d'identification et d'extraction automatiques des relations n-aires dans les textes non structurés reste un défi de recherche majeur. La problématique d'extraction des relations n-aires ne doit pas être considérée comme plusieurs extractions de relations binaires, i.e. *n-1* extractions de relations binaires. En effet, en supposant qu'une relation n-aire n'est finalement qu'un ensemble de relations binaires, la relation n-aire est alors factorisée en *n-1* relations binaires à extraire. Contrairement à ce que nous pouvons supposer, (Zhou et al., 2014) montre que dans les faits, si la précision d'extraction de la relation binaire est de p, alors la précision de l'extraction de la relation n-aire sera, en utilisant les mêmes méthodes d'extraction de la relation binaire, affaiblie à p^{n-1}. Supposons que la précision de l'extraction de la relation binaire soit de $p=0.8$ alors la précision de la relation n-aire, avec $n=5$, sera de $p=0.512$.

Nos travaux s'inscrivent dans cette démarche d'extraction et de capitalisation des connaissances dans le domaine des sciences du vivant et de l'environnement. La communauté fait face à un nombre croissant de données, en particulier de données publiées sur le web, sous forme d'articles scientifiques. Ces articles comportent de l'information, comme les résultats expérimentaux des travaux de recherche, indispensable pour confronter, comparer et découvrir de nouvelles connaissances. Leur pérennisation est incontestablement un enjeu majeur.
Dans ce contexte, de nombreuses initiatives ont déjà été prises afin de trouver des solutions durables pour la gestion et l'analyse de données, en commençant par la définition de thésaurus, étape fondamentale pour constituer un vocabulaire partagé.
L'un des plus importants thésaurus constitués du domaine est AGROVOC (Caracciolo et al., 2012) par la FAO (Food and Agriculture Organization of the United Nations) [3]. Disponible en 19 langues, il regroupe près de 40000 termes dans chacune des langues et couvre les domaines de l'agriculture, de la sylviculture, de la pêche, de l'alimentation et de domaines apparentés comme l'environnement.
NALT (National Agricultural Library Thesaurus) est un autre thésaurus faisant référence dans le milieu international. Il est constitué de 91000 termes en anglais et espagnol et couvre également de nombreux domaines. Environ 14000 termes dans AGROVOC sont reliés aux termes de NALT. Ces références sont publiées sur le web de données et actuellement reliées au vocabulaire de 11 ressources internationales (comme GeoNames ou DBpedia).

3. La liste complète des publications en rapport avec AGROVOC se trouve à l'adresse : http ://aims.fao.org/fr/publications

D'autres initiatives dans le domaine des sciences du vivant et de l'environnement ont émergé pour créer des ontologies permettant de décrire des connaissances :

- Le domaine de la culture des plantes avec l'ontologie Ref-TO (Arnaud et al., 2012),
- L'ontologie pour représenter les connaissances sur les capteurs en agriculture et environnement (Bendadouche et al., 2012; Compton et al., 2012),
- L'ontologie décrivant les relations microorganismes-habitat (Bossy et al., 2012).

1.2 Contexte et problématiques identifiées

La thèse s'inscrit dans une nouvelle initiative d'extraction et de capitalisation des connaissances. Cette initiative a été appliquée aux domaines du risque alimentaire microbiologique étendu aux emballages et à la bioraffinerie. De nombreux résultats expérimentaux sont publiés dans les articles scientifiques et diffusés sur le web dans un format textuel non structuré. Ces résultats expérimentaux sont associés à des paramètres de contrôle d'importance pour le domaine. Ces paramètres sont mesurés au cours des différentes expérimentations menées par les chercheurs et restitués au sein des articles scientifiques. Les travaux récents ont permis de définir et standardiser le vocabulaire du domaine dans une ontologie (Touhami et al., 2011), ou plus précisément une Ressource Termino-Ontologique (RTO). La RTO du domaine permet de représenter en langage formel structuré les paramètres de contrôle sous forme d'arguments engagés dans une relation n-aire, représentant le résultat expérimental associé à ses paramètres de contrôle. La modélisation et la capitalisation des données d'intérêt dans une RTO représentent des étapes fondamentales pour le partage des connaissances sur le web de données, l'analyse des données afin de produire de la connaissance et enfin, proposer un support d'aide à la décision (Guillard et al., 2015). En effet, outre la capitalisation des connaissances, la possibilité de raisonner à partir de l'ontologie permet de proposer un langage de requêtes pour des prises de décision stratégiques. Par exemple, le projet EcoBiocap, dans lequel sont intégrés les travaux de recherche relatifs aux emballages alimentaires, utilise la RTO comme support dans un outil d'aide à la décision sous forme de requêtes : on définit des paramètres comme l'aliment à emballer, les conditions de température auxquelles on souhaite conserver l'aliment, en précisant des valeurs de perméabilité optimales, l'outil aide au choix de l'emballage à utiliser, en proposant une liste d'emballages répondant aux critères définis. Cet exemple montre l'intérêt de capitaliser la connaissance mais également le besoin de peupler les ontologies ou RTO de domaine, avec de nouvelles informations. La RTO du domaine joue un double rôle fondamental puisqu'elle sert, à partir des instances référencées, de support aux outils d'aide à la décision et, elle guide les étapes qui permettent son

propre peuplement avec de nouvelles instances. Dans ce contexte, les travaux menés dans le domaine des emballages alimentaires ont abouti à l'élaboration du système ONDINE (ONtology-based Data INtEgration), illustré dans la figure 1.8, qui propose un processus complet d'intégration de données (Buche et al., 2013c) où la RTO joue un rôle central. Le système repose sur deux sous-systèmes :

1. Le sous-système d'acquisition et d'annotation @web permet d'annoter avec des concepts de la RTO des données, représentant des résultats expérimentaux d'intérêt, issues des tableaux trouvés dans les articles scientifiques publiés sur le web.
2. Le sous-système d'interrogation MIEL++(Méthode d'Interrogation Elargie) propose un système d'interrogation unifiée et flexible que nous mentionnons pour information mais que nous ne détaillons pas dans cette thèse.

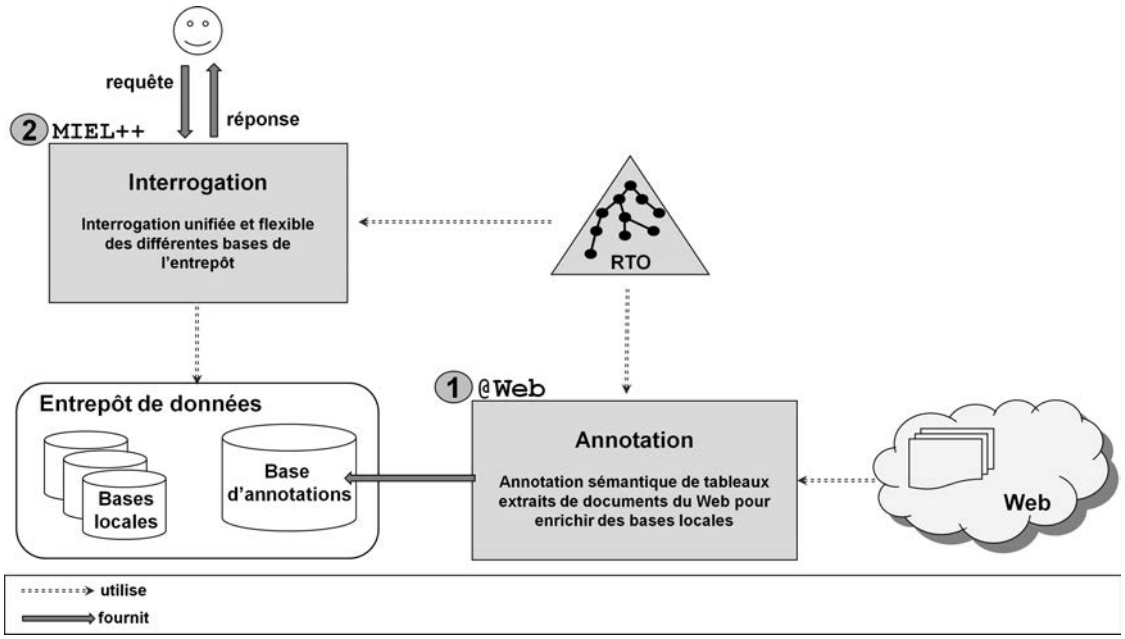
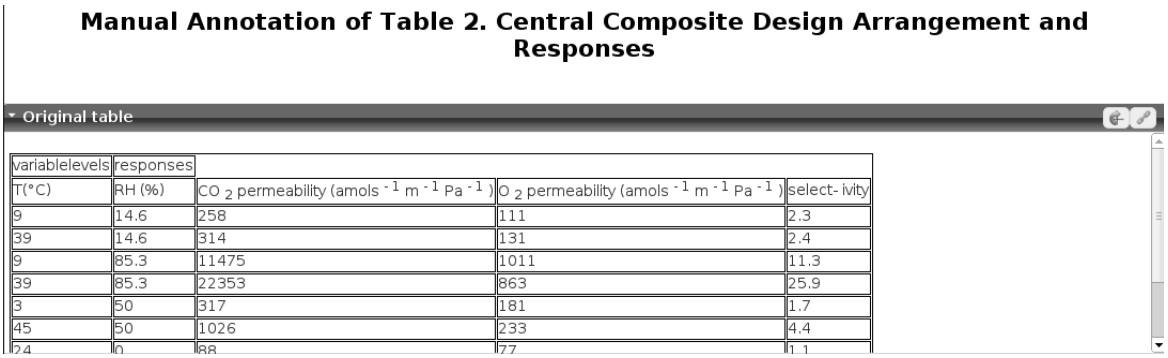


FIGURE 1.8: Architecture du système ONDINE

Le sous-système @web (Buche et al., 2013b) a été conçu selon deux modules : un module qui permet la gestion de la RTO du domaine et un module pour guider l'annotation des instances de relations n-aires que l'on retrouve dans certains tableaux de documents scientifiques préalablement sélectionnés pour leur pertinence quant au domaine étudié. Lors de l'annotation des données d'intérêt, plusieurs problématiques ont été identifiées et nécessitent de nouvelles contributions pour optimiser les processus d'annotation et d'extraction des nouvelles instances de relations n-aires de données expérimentales.

Tout d'abord dans les tableaux, une des problématiques relevées dans ce processus d'annotation des tableaux dans @web est l'incomplétude dans le renseignement des arguments d'une instance de concept relation. En effet, dans le tableau d'origine à annoter, les instances peuvent être incomplètes avec certains arguments manquants. La figure 1.9 montre un tableau pertinent à annoter à partir du sous-système @web.
Le tableau présente deux types de relation n-aire : *CO2 Permeability* et *O2 Permeability*. Nous pouvons constater que deux arguments qui sont des paramètres de contrôle de l'expérimentation sont manquants dans ce tableau d'origine, l'instance d'arguments *Thickness* et *Partial Pressure*. Les informations manquantes se trouvent dans le corps du texte du document dans lequel a été sélectionné le tableau.



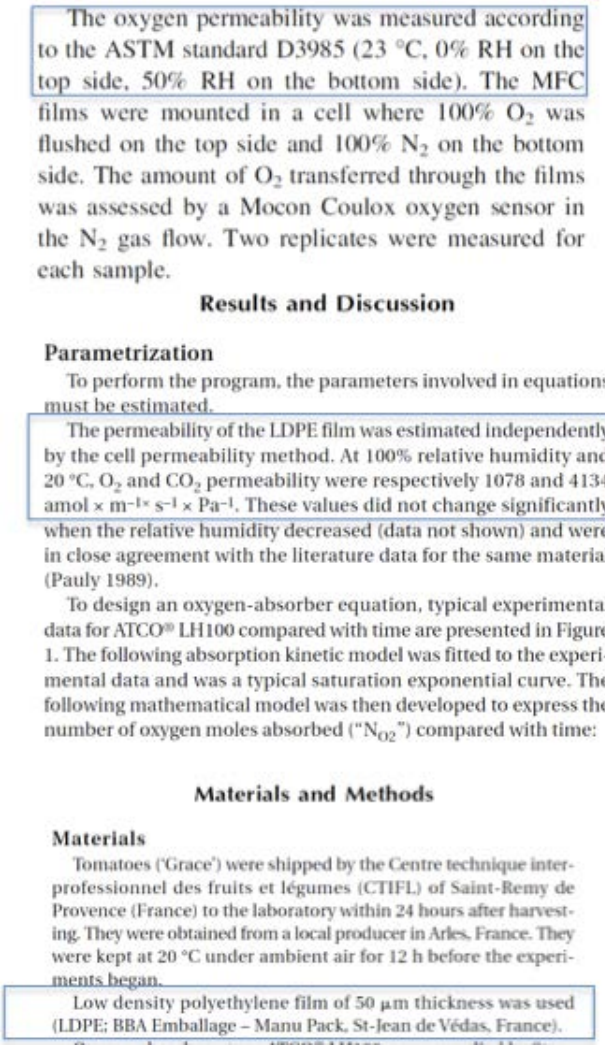
variablelevels	responses			
T(°C)	RH (%)	CO_2 permeability (amols $^{-1}$ m $^{-1}$ Pa $^{-1}$)	O_2 permeability (amols $^{-1}$ m $^{-1}$ Pa $^{-1}$)	select- ivity
9	14.6	258	111	2.3
39	14.6	314	131	2.4
9	85.3	11475	1011	11.3
39	85.3	22353	863	25.9
3	50	317	181	1.7
45	50	1026	233	4.4
24	0	88	77	1.1

FIGURE 1.9: Instance de concept de relation incomplète.

Plus généralement, toutes les données ne se retrouvent pas uniquement dans les tableaux. L'exemple de la figure 1.10, montre que l'instance recherchée se trouve dans le texte. Les arguments recherchés (*packaging, thickness, temperature, relative humidity, permeability*) de l'instance de relation n-aire à extraire se retrouvent dans 3 paragraphes différents, *Experiments, Materials and Methods* et *Results and discussion*.

Considérant ces observations, la thèse s'intéresse à l'extraction des instances de relations n-aires complètes ou partielles, décrivant des résultats expérimentaux, que l'on retrouve dans le corps du texte de l'article scientifique. Elle a pour objectif de proposer une nouvelle méthode d'extraction de ces nouvelles instances de relations n-aires afin de peupler une RTO de domaine. La complexité de la tâche d'extraction des instances de relations n-aires de données expérimentales est liée à plusieurs verrous scientifiques :

- Un premier verrou concerne la forte variation typographique des unités de mesure associées aux résultats quantitatifs dans les documents. En effet, elle impacte fortement le processus d'annotation automatique des textes guidé par la RTO car les



The oxygen permeability was measured according to the ASTM standard D3985 (23 °C, 0% RH on the top side, 50% RH on the bottom side). The MFC films were mounted in a cell where 100% O_2 was flushed on the top side and 100% N_2 on the bottom side. The amount of O_2 transferred through the films was assessed by a Mocon Coulox oxygen sensor in the N_2 gas flow. Two replicates were measured for each sample.

Results and Discussion

Parametrization

To perform the program, the parameters involved in equations must be estimated.

The permeability of the LDPE film was estimated independently by the cell permeability method. At 100% relative humidity and 20 °C, O_2 and CO_2 permeability were respectively 1078 and 4134 amol × m^{-1}× s^{-1} × Pa^{-1}. These values did not change significantly when the relative humidity decreased (data not shown) and were in close agreement with the literature data for the same material (Pauly 1989).

To design an oxygen-absorber equation, typical experimental data for ATCO® LH100 compared with time are presented in Figure 1. The following absorption kinetic model was fitted to the experimental data and was a typical saturation exponential curve. The following mathematical model was then developed to express the number of oxygen moles absorbed ("N_{O2}") compared with time:

Materials and Methods

Materials

Tomatoes ('Grace') were shipped by the Centre technique interprofessionnel des fruits et légumes (CTIFL) of Saint-Remy de Provence (France) to the laboratory within 24 hours after harvesting. They were obtained from a local producer in Arles, France. They were kept at 20 °C under ambient air for 12 h before the experiments began.

Low density polyethylene film of 50 µm thickness was used (LDPE; BBA Emballage – Manu Pack, St-Jean de Védas, France).

FIGURE 1.10: Un extrait d'instance de relation n-aire avec des arguments dispersés dans les sections du document

variants qui ne sont pas référencés, ne sont pas reconnus. De ce fait, l'annotation de certains arguments quantitatifs de la relation échoue ;

- Un autre verrou concerne l'expression de la relation n-aire avec des arguments dispersés dans plusieurs sections du document et selon un discours implicite, comme cela est illustré également dans la figure 1.10 à partir d'un document du corpus des emballages alimentaires ;

- La présence dans les documents de nombreux résultats expérimentaux incluant la mesure de paramètres de contrôle autres que ceux représentés dans les relations n-aires que nous cherchons à extraire. En effet, par exemple, dans le corpus des emballages, outre les résultats de perméabilité, les documents restituent d'autres types de résultats comme les tests de résistance mécanique ou des propriétés antimicrobiennes des emballages. L'exemple 1 illustre ce cas, à partir de deux phrases extraites du corpus des emballages exprimant d'autres conditions expérimentales. L'exemple montre également que les unités de mesure associées aux arguments des instances de la relation n-aire peuvent être les mêmes (dans l'exemple, l'unité de température $°C$) pour définir les résultats quantitatifs de ces autres paramètres de contrôle.

Exemple 1
La phrase suivante restitue les résultats des tests de résistance mécanique. Les tests sont effectués à une température d'environ 23 °C, la longueur initiale du film de 100 mm est étirée de 7,5 mm/min :
All films were conditioned for 48 h at $\mathbf{23 \pm 2°C}$ *and* $50\% \pm 2\%$ *RH before testing using a saturated salt solution of magnesium nitrate (Fisher Scientific, Fair Lawn, NJ). The ends of the equilibrated strips were mounted and clamped with pneumatic grips on an Instron Model 55R4502 Universal Testing Machine (Instron, Canton, MA) with a 100 N load cell. The initial gauge length was set to 100 mm and films were stretched using a crosshead speed of 7.5 mm/min.*

Cette autre phrase restitue les résultats des tests concernant les propriétés antimicrobiennes de l'emballage évalué (AAPEF) à une température de $37°C$ *:*
Disc inhibition zone assays were performed as a qualitative test for antimicrobial activity of the films. AAPEF with and without EOs and OCs were aseptically cut into 12 mm diameter discs and then placed on MacConkey-Sorbitol agar (Biokar Diagnostics, Beauvais, France) plates for E. coli O157 :H7, which had been previously

spread with 0.1 mL of inoculum containing 10^5 *CFU/mL of tested bacterium. Plates were incubated at* $\mathbf{37°C}$ *for 48 h.*

Nos travaux tout au long de ce mémoire cherchent à répondre à ces problématiques et contribuer au travail d'extraction des instances de relations n-aires de données expérimentales dans les textes, comme nous le détaillons dans la section suivante.

1.3 Contributions

Dans le but de répondre aux problématiques soulevées dans la section précédente, nous proposons deux contributions principales, s'appuyant sur une RTO de domaine. La première consiste à localiser et à identifier les termes dénotant les variants d'unités de mesure réputés difficiles à extraire, afin d'enrichir la RTO. La seconde est une nouvelle approche hybride pour l'extraction d'instances d'arguments des relations n-aires des données expérimentales. Plus précisément, nous proposons une nouvelle méthode guidée par la RTO permettant l'extraction et l'identification des variants d'unités de mesure en deux étapes :

- La première étape propose de prédire la localisation des variants d'unités de mesure en nous appuyant sur l'apprentissage supervisé afin de réduire l'espace de recherche des variants dans les textes ;
- La deuxième étape propose, une fois l'espace de recherche réduit, une nouvelle mesure de similarité adaptée à la syntaxe des unités de mesure afin d'identifier les variants extraits des documents. Ces variants d'unités de mesure permettent d'enrichir la partie terminologique de la RTO.

Notre nouvelle méthode hybride contribuant à l'extraction des instances d'arguments de la relation n-aire est également constituée de deux parties :

- Dans une première partie, nous tirons profit de la capacité des méthodes de fouille de données à faire émerger des régularités et des motifs afin de prendre en compte la diversité d'expressions des instances d'arguments de la relation n-aire dans les documents,
- Dans une deuxième partie, nous proposons d'extraire les relations de dépendances syntaxiques proches de la définition de la relation n-aire afin d'enrichir les motifs découverts au cours de la première partie. Les motifs sont étendus par combinaison avec les relations syntaxiques extraites pour l'extraction d'instances d'arguments de la relation n-aire.

1.4 Organisation du mémoire

Le mémoire est organisé de la manière suivante :

- Le chapitre 2, dans la section 2.1 justifie le choix d'une RTO plutôt qu'une ontologie, la section 2.2 présente la RTO naRyQ *n-ary Relations between Quantitative experimental data* et sa modélisation dans les deux domaines d'application s'intégrant dans le cadre de la thèse, le domaine du risque alimentaire microbiologique étendu aux emballages et le domaine de la bioraffinerie. La section 2.3 présente les principaux éléments de travail sur lesquels reposent nos propositions ;

- Le chapitre 3 présente en détails notre proposition pour enrichir la RTO de nouveaux variants d'unités de mesure. La section 3.2 dresse un état de l'art concernant les travaux sur les unités de mesure et plus généralement sur les données quantitatives. La section 3.3 détaille la première étape de notre contribution concernant la réduction de l'espace de recherche des variants dans les documents textuels. Nous exposons notre méthode basée sur l'apprentissage supervisé pour prédire la localisation des variants. La section 3.4 détaille la deuxième étape de notre contribution concernant l'identification des variants d'unités de mesure. Nous présentons notre mesure adaptée à l'identification des variants extraits des documents et nous comparons les résultats obtenus avec des mesures classiques de la littérature ;

- Le chapitre 4 présente notre nouvelle approche hybride fondée sur les méthodes de fouille de données combinées à de l'analyse syntaxique. La section 4.2 détaille notre approche s'appuyant sur les méthodes de fouille de données, adaptée à notre contexte de données de type expérimental, afin d'extraire de la connaissance implicite concernant l'expressivité des arguments de la relation n-aire. La section 4.3 présente l'extraction des relations de dépendances syntaxiques d'intérêt pour le domaine en étant guidés par la RTO. Notre méthode permet d'extraire les relations syntaxiques pertinentes à l'extension des motifs émergeant de l'étape de fouille de données. Nous montrons également comment, à partir de notre méthode hybride, nous obtenons des motifs étendus pour l'extraction des instances comportant plus de 2 arguments.

Chapitre 2

Contexte d'étude

Sommaire

Nous avons présenté en introduction le contexte général de la thèse avec les problématiques soulevées et les contributions apportées. Tout au long du manuscrit, le travail effectué et les contributions proposées s'appuient, d'une part, sur la connaissance de domaine, capitalisée au sein d'une RTO, et d'autre part, nous souhaitons contribuer à la problématique d'extraction de données d'intérêt, représentées sous forme de relations n-aires, en évaluant les méthodes d'apprentissage supervisé et de fouille de données. Dans ce chapitre, il nous semble important de revenir et motiver les choix des éléments méthodologiques sur lesquels s'appuient nos propositions :

- le choix d'une modélisation en Ressource Termino-Ontologique (RTO) plutôt qu'une ontologie,
- la description des RTO des domaines d'application intégrés dans le cadre de la thèse,
- les définitions et hypothèses de travail sur lesquelles reposent les contributions proposées dans le cadre de la thèse.

2.1 Ontologie ou Ressource Termino-Ontologique ?

La notion d'ontologie est utilisée dans le domaine de l'Intelligence Artificielle, et plus précisément dans la branche de l'Ingénierie des Connaissances, pour la conception des systèmes à base de connaissances.

Une définition consensuelle utilisée dans la littérature en Intelligence Artificielle est celle de (Gruber, 1993) : «Une ontologie est une spécification explicite d'une conceptualisation.» Le terme "conceptualisation" situe les ontologies au niveau sémantique. Elle pose ainsi le sens des termes utilisés et fortement corrélés au domaine considéré. La caractérisation du sens des termes dépend du contexte dans lequel ils apparaissent. En effet, la linguistique est concernée par la question des ontologies dans la mesure où les données dont on dispose pour élaborer les ontologies consistent en des expressions linguistiques de connaissances. Nous parlons alors d'ontologie élaborée pour une tâche donnée et dans un contexte de référence (Bachimont, 2000). L'expression "spécification explicite" fait des ontologies un objet syntaxique. La conceptualisation est faite dans un langage formel qui définit les concepts et les contraintes d'utilisation. On obtient un réseau sémantique et un ensemble de formules logiques sous-jacentes.
La construction du vocabulaire conceptuel standardisé est établie en définissant un ensemble de primitives de représentation pour modéliser le domaine. En OWL, les primitives d'une ontologie sont (Guarino et al., 2009) :

- les concepts de l'ontologie sous forme de `owl :Class`,
- les attributs de concepts sous forme de `owl :DatatypeProperty`,
- les relations binaires entre concepts sous forme de `owl :ObjectProperty`.

OWL a été élaboré dans l'optique de l'indexation de ressources sur le Web. Il permet donc de représenter le lexique sous la forme duquel un concept pourra apparaître dans un document. La modélisation des termes désignant le concept dans le langage OWL se fait par association du terme à la classe correspondante au moyen d'une propriété d'annotation, `rdfs :label`. Cette modélisation limitée pose plusieurs problématiques, en particulier pour le domaine de recherche s'intéressant à l'extraction et à l'annotation des données dans les documents textuels. En effet, un terme ainsi représenté n'a pas d'existence propre, on ne peut pas lui associer directement de propriétés, e.g. une étiquette grammaticale. Le terme est relié au concept qu'il désigne et de ce fait, la dissociation des informations conceptuelles et lexicales est impossible.

La nécessité de matérialiser la notion de `terme`, de manière à pouvoir la manier aussi aisé-

ment qu'un concept, est apparue à partir des travaux de (Reymonet et al., 2007; Aussenac-Gilles et al., 2006) avec la naissance de la notion de Ressource Termino-Ontologique (RTO). Dans cette nouvelle modélisation, la manifestation linguistique (le terme) est dissociée de la notion qu'elle dénote (le concept), en la représentant et en lui octroyant une existence propre et indépendante.

Ceci peut être fait en utilisant SKOS comme proposé dans les travaux de (Touhami et al., 2011). SKOS, Simple Knowledge Organisation System, est une recommandation du W3C [1] permettant de représenter les thésaurus, classifications et autres vocabulaires contrôlés. SKOS s'appuie sur le langage RDF afin de permettre la publication facile de vocabulaires structurés pour leur utilisation dans le cadre du web sémantique. La représentation s'établit sous forme de triplets RDF tels que la ressource de base est un `skos :Concept`. Plusieurs propriétés sont octroyées à un objet de type `skos :Concept` :

- des termes préférés (un maximum par langue) en utilisant le `skos :prefLabel`,
- des termes alternatifs qui vont représenter les synonymes ou les abréviations (plusieurs par langue) en utilisant le `skos :altLabel`,
- des termes cachés pour gérer les variantes correspondant à des fautes d'orthographe courantes en utilisant le `skos :hiddenLabel`,
- d'autres propriétés permettent de poser des définitions, des notes et des exemples.

Les instances de `skos :Concept` peuvent être reliées par des relations de spécialisation ou par des relations associatives, i.e. `skos :related`.
SKOS-XL, SKOS eXtension for Labels, est une extension de SKOS qui permet de représenter plus finement les termes en proposant la représentation des relations entre les différentes formes lexicales, e.g. relation d'acronymie. Cette représentation est particulièrement intéressante, dans le cadre de la RTO, car elle permet de considérer les termes comme une ressource à part entière, avec la propriété `skosxl :Label`.

Dans le cadre de la thèse, le travail repose sur une modélisation en RTO dédiée à la représentation des relations n-aires, où composante conceptuelle et terminologique sont clairement dissociées. Cette RTO est présentée dans la section 2.2. Dans la suite du manuscrit, nous utilisons indifféremment les termes `Ontologie` et `RTO` en sous-entendant la notion de RTO.

1. http ://www.w3.org/TR/skos-primer/

2.2 Présentation des Ressources Termino-Ontologiques des domaines d'application

La RTO sur laquelle s'appuie nos travaux, appelée généralement RTO naRyQ pour *n-ary Relations between Quantitative experimental data*, a été modélisée pour représenter des relations n-aires entre des données quantitatives expérimentales. Le choix de la représentation des relations n-aires de la RTO naRyQ (Touhami et al., 2011; Buche et al., 2013a) s'est porté sur la représentation sans arguments différenciés, c'est-à-dire le cas d'utilisation 3 définit par le W3C et illustré dans la figure 2.1. Ce cas d'utilisation définit une relation reliant des individus ayant des rôles différents sans qu'aucun d'entre eux ne joue le rôle central de "sujet" de la relation. Dans le cas de la RTO naRyQ, ce choix est intéressant car il correspond à un bon compromis entre généricité dans la représentation des relations n-aires de données de type expérimental et simplicité d'utilisation et de manipulation pour un utilisateur non informaticien.

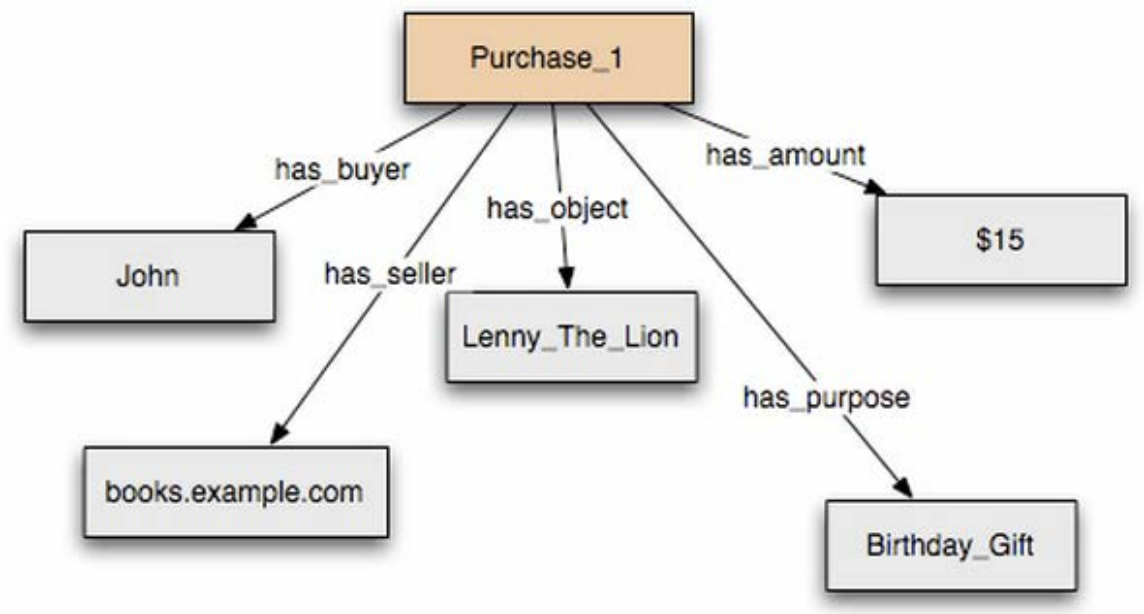


FIGURE 2.1: Cas 3 de représentation de relation n-aire à partir de la phrase *John buys a Lenny the Lion book from books.example.com for $ 15 as a birthday gift*

Dans les sections suivantes, nous présentons la composante conceptuelle et terminologique de la RTO naRyQ.

2.2.1 La composante conceptuelle

La composante conceptuelle de la RTO naRyQ est composée de deux parties, *une ontologie noyau* qui permet de représenter des relations n-aires et *une ontologie de domaine* qui permet de représenter les concepts spécifiques à un domaine donné. La représentation des relations n-aires prend en compte des arguments symboliques, notamment les objets sur lesquels portent les expérimentations, et des arguments quantitatifs définis par des valeurs numériques et des unités de mesure. Le modèle de RTO naRyQ propose, d'une part, de représenter la modélisation des données quantitatives de manière générique, dans l'ontologie noyau *core ontology*, comme l'illustre la figure 2.2, et d'autre part, la modélisation de la relations n-aire spécifique au domaine étudié, *domain ontology*.

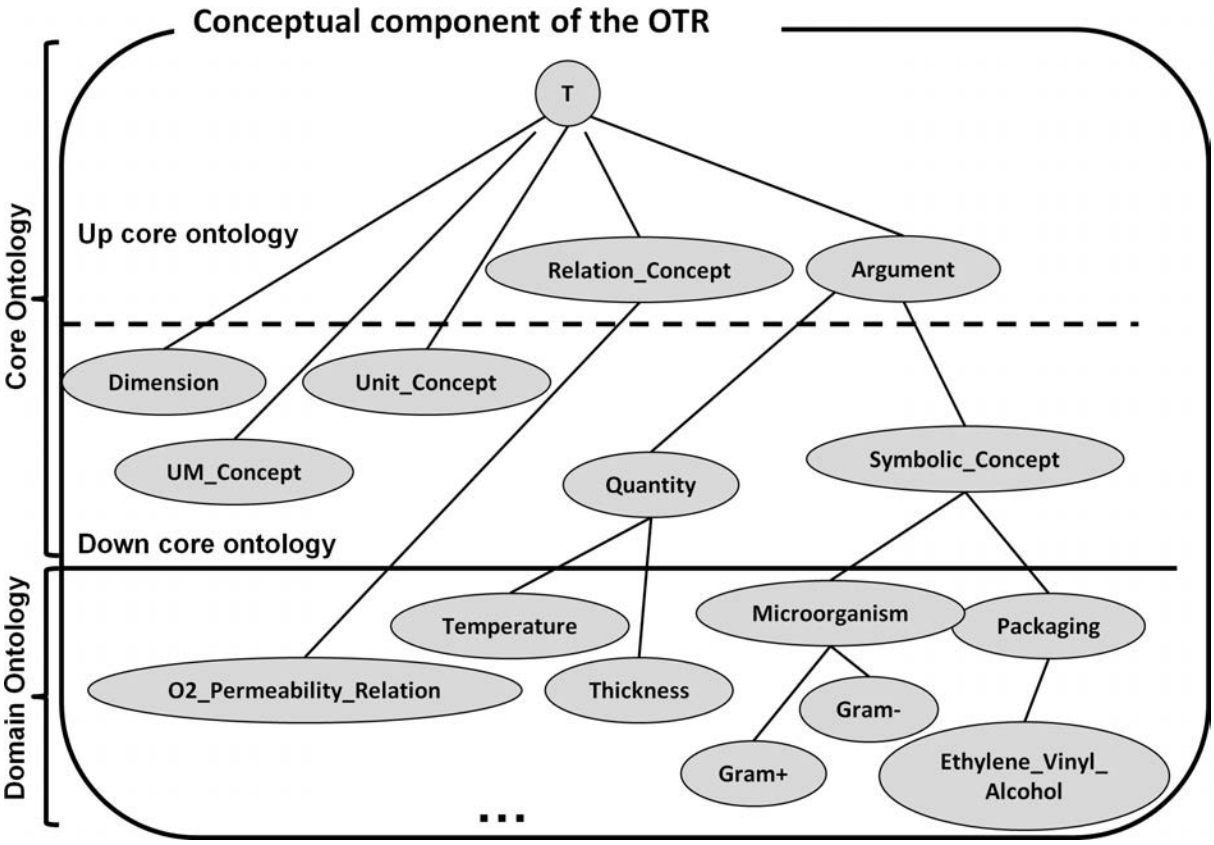


FIGURE 2.2: Un extrait de la RTO naRyQ_emb dans le domaine du risque alimentaire microbiologique étendu aux emballages

L'ontologie noyau supérieure *up core ontology* modélise les concepts génériques *Relation_Concept* et *Argument* permettant de représenter les relations n-aires et leurs argu-

ments.

Définition 1 *La relation n-aire :*
Un concept relation est caractérisé par son label (i.e. un terme), défini dans la composante terminologique de la RTO, et par sa signature qui permet de définir l'ensemble des concepts, sous-concepts du concept générique Argument, qui peuvent être reliés par la relation.

Exemple 2 *La figure 2.3 présente la relation générique Permeability_ Relation. Cette relation peut avoir trois types de signature, O2Permeability_ Relation, H2OPermeability_ Relation, CO2Permeability_ Relation si les paramètres de sortie mesurés sont respectivement O2_ permeability, H20_ permeability et CO2_ permeability. Ces concepts relation permettent de représenter la perméabilité à l'oxygène, à l'eau et au dioxyde de carbone d'un emballage dans des conditions expérimentales données par des paramètres de contrôle : l'épaisseur de l'emballage, l'humidité relative, la pression partielle et la température ambiante.*

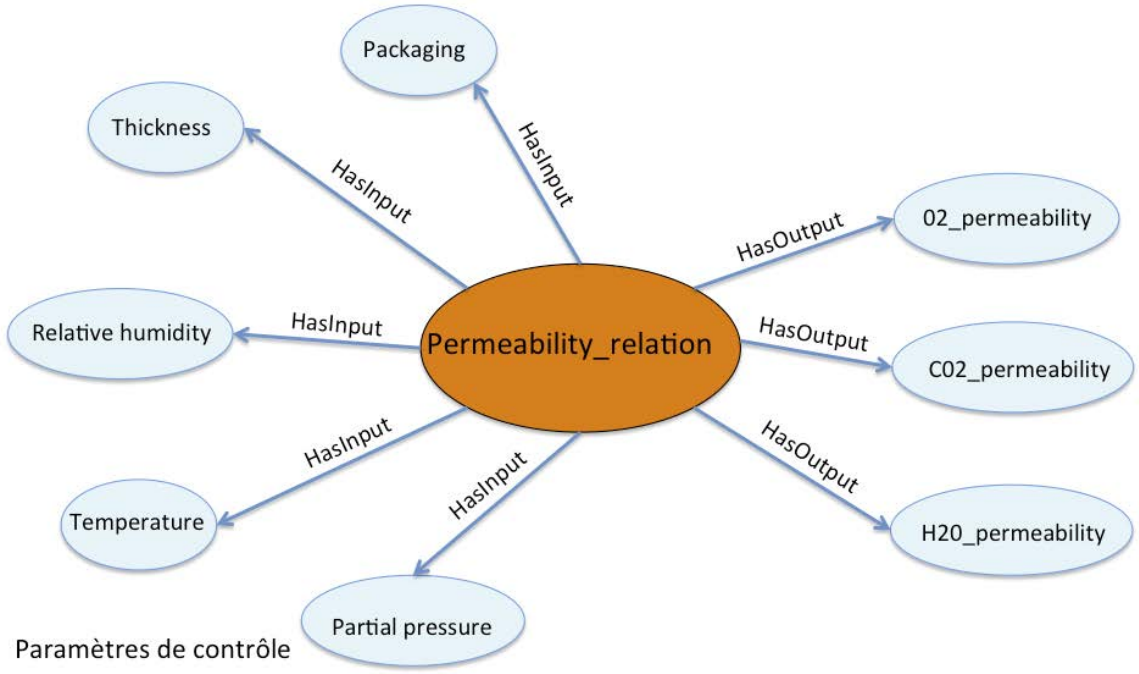


FIGURE 2.3: Représentation des relations n-aires *Permeability_ Relation*

L'ontologie noyau inférieure *down core ontology* modélise les concepts génériques suivants :

- *Dimension* : contient les dimensions qui permettent aux quantités et à leurs unités de mesures associées d'être classifiées. Par exemple, le concept quantité *thickness* et son unité de mesure sont rattachés à la dimension *length*,
- *UM_ Concept* permet de gérer les conversions entre unités de mesure,
- *Unit_ Concept*, représenté par *Quantity* dans la figure 2.4, permet de gérer les concepts unités de mesure comme illustré dans la figure 2.6. Le concept Unit_Concept possède 4 sous-concepts *Singular_ Unit, Unit_ Division_ Or_ Multiplication, Unit_ - Multiple_ Or_ Submultiple,*
  Unit_ Exponentiation. Les unités sont des instances de ces sous-concepts. Chaque instance est associée à une partir terminologique, e.g $°C$ dénote l'instance d'unité *Degree_ Celsius*,
- *Quantity_ Concept* permet de gérer les arguments de type quantitatif. Le concept se subdivise en sous-concepts, dépendant du domaine étudié, représentant les arguments quantitatifs des relations n-aires représentées dans la RTO,
- *Symbolic_ Concept*, représenté dans la figure 2.5, permet de gérer les arguments non numériques de la relation n-aire.

L'ontologie de domaine correspond à la modélisation spécifique du domaine étudié. Elle contient tous les concepts spécifiques qui sont modélisés comme des sous-concepts des concepts génériques de l'ontologie noyau. Ils sont organisés en hiérarchie par la relation de subsomption OWL `subClassOf`.

2.2.2 La composante terminologique

La composante terminologique de la RTO naRyQ contient l'ensemble des termes du domaine étudié. Chaque concept de la hiérarchie modélisée dans la composante conceptuelle est considéré à la fois comme `owl :Class` et comme une instance de la classe `skos :concept`. Comme nous l'avons présenté dans la section 2.1, cette représentation est spécifiquement dédiée à la terminologie en utilisant un ensemble de propriétés telles que preflabel et altlabel.

La figure 2.7 illustre cette représentation à l'aide des propriétés SKOS, à partir de l'exemple donné sur le concept *Ethylene Vinyl Alcohol*. La figure montre qu'il est possible d'ajouter un `skos :Concept` de type emballage. Cette représentation autorise l'utilisation d'une seule et même URI pour *Ethylene Vinyl Alcohol*. *Ethylene Vinyl Alcohol*

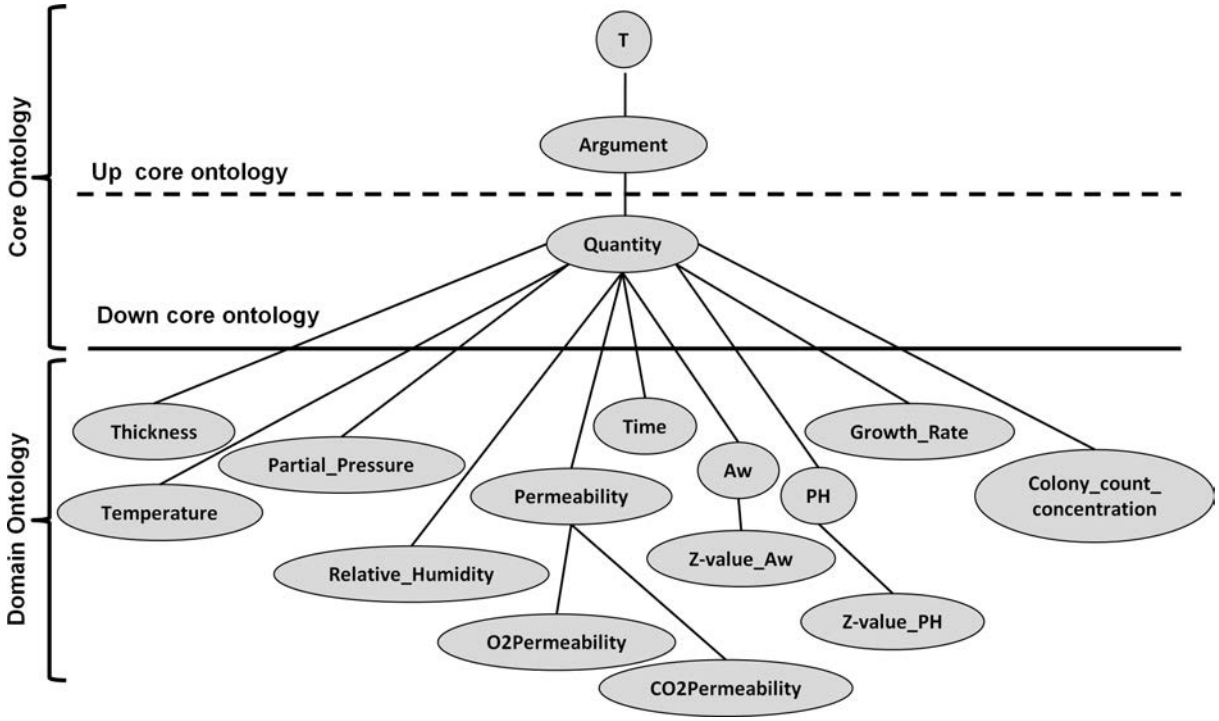


FIGURE 2.4: Un extrait de la hiérarchie des quantités de la RTO naRyQ_emb

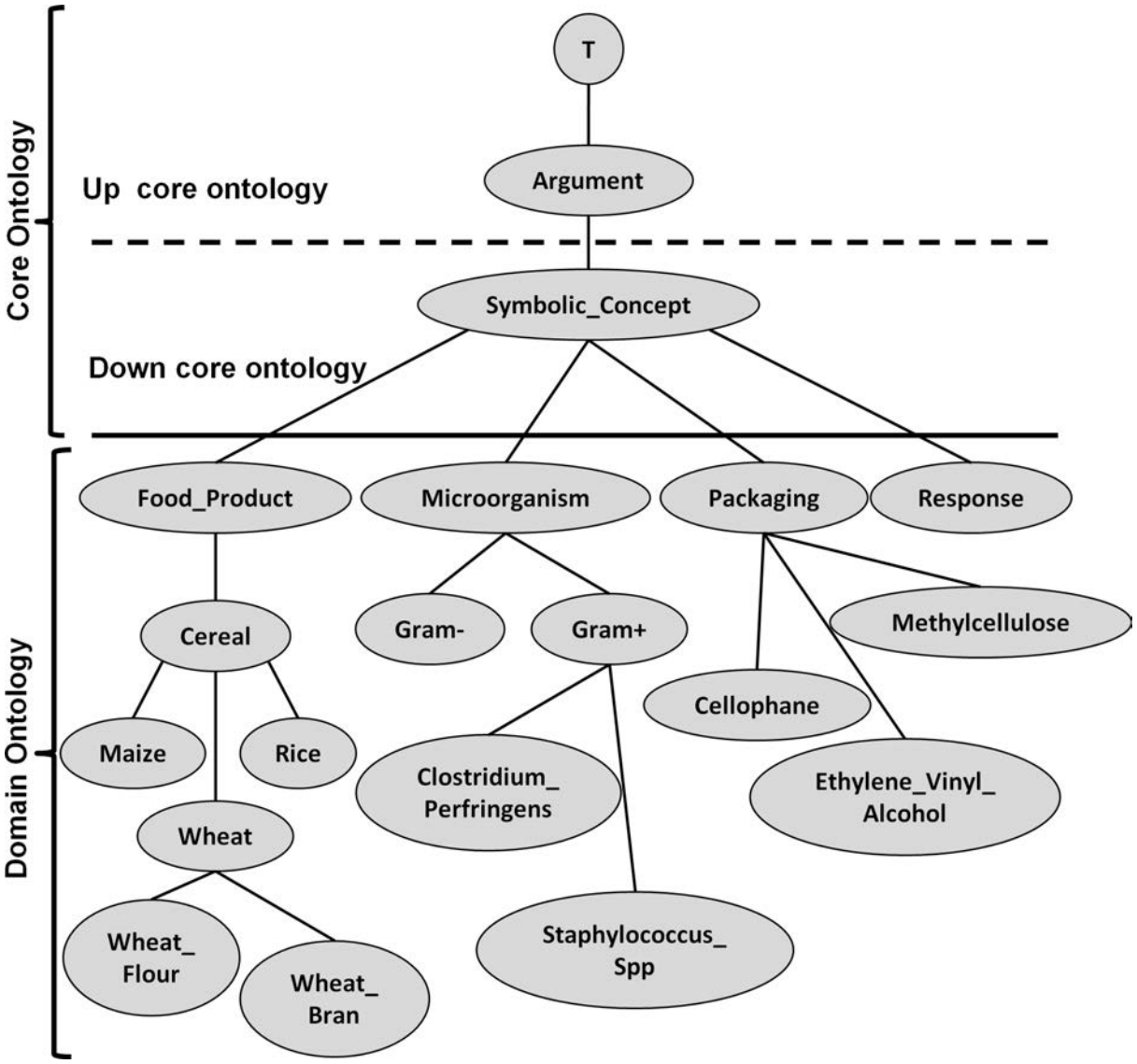


FIGURE 2.5: Un extrait de la hiérarchie des concepts symboliques de la RTO naRyQ_emb

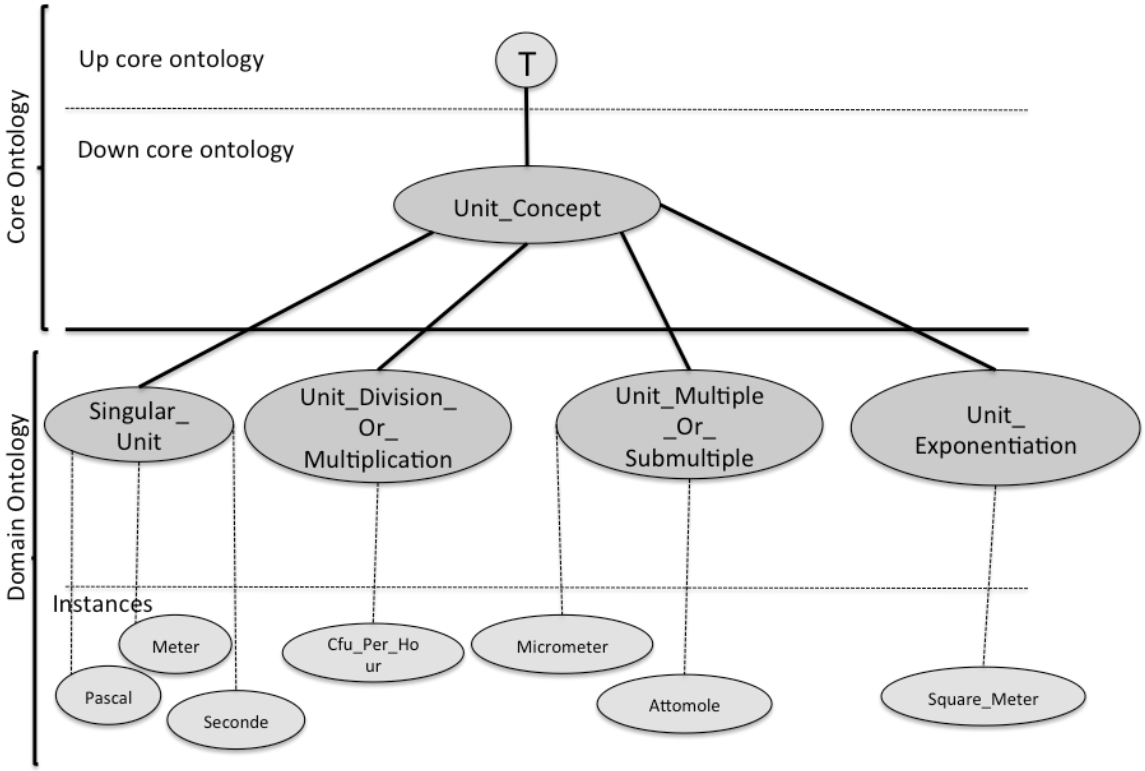


FIGURE 2.6: Un extrait de la hiérarchie de concepts des unités de mesure et quelques exemples d'instances dans naRyQ_emb

devient à la fois une instance de `skos :Concept` et un concept de type `owl :Class`. L'association du concept et du terme, représenté en `skos :Concept`, se fait par mapping de type 1 :1. Ainsi, *Ethylene Vinyl Alcohol* bénéficie de toutes les propriétés associées à la représentation SKOS.
Cette composante terminologique est particulièrement importante pour nos travaux, notamment pour compléter la terminologie du domaine d'informations jugées pertinentes, e.g. étiquettes syntaxiques. L'exemple 3 montre un extrait d'analyse syntaxique sur une phrase extraite du corpus des emballages. Le triplet de dépendance restitue le lien grammatical qu'entretiennent les deux entités de la relation selon leur position dans la phrase. Ainsi, le triplet *prep_ of(thickness-10, μm-13)* montre une relation de type prépositionnel entre l'unité de mesure μm, à la position 13 dans la phrase, et l'argument quantitatif *thickness*, à la position 10 dans la phrase. Ce type d'information peut se révéler pertinent à conserver dans les propriétés des termes de la RTO.

Exemple 3 *The film used in the reference work had a thickness of 54 µm and its OP was measured at 24 °C and 0% relative humidity.*
Extrait de triplets de dépendances syntaxiques restitués :
det(thickness-10, a-9)
nsubjpass(measured-18, thickness-10)
num(µm-13, 54-12)
prep_of(thickness-10, µm-13)
poss(OP-16, its-15)
prep_of(thickness-10, OP-16)
conj_and(µm-13, OP-16)

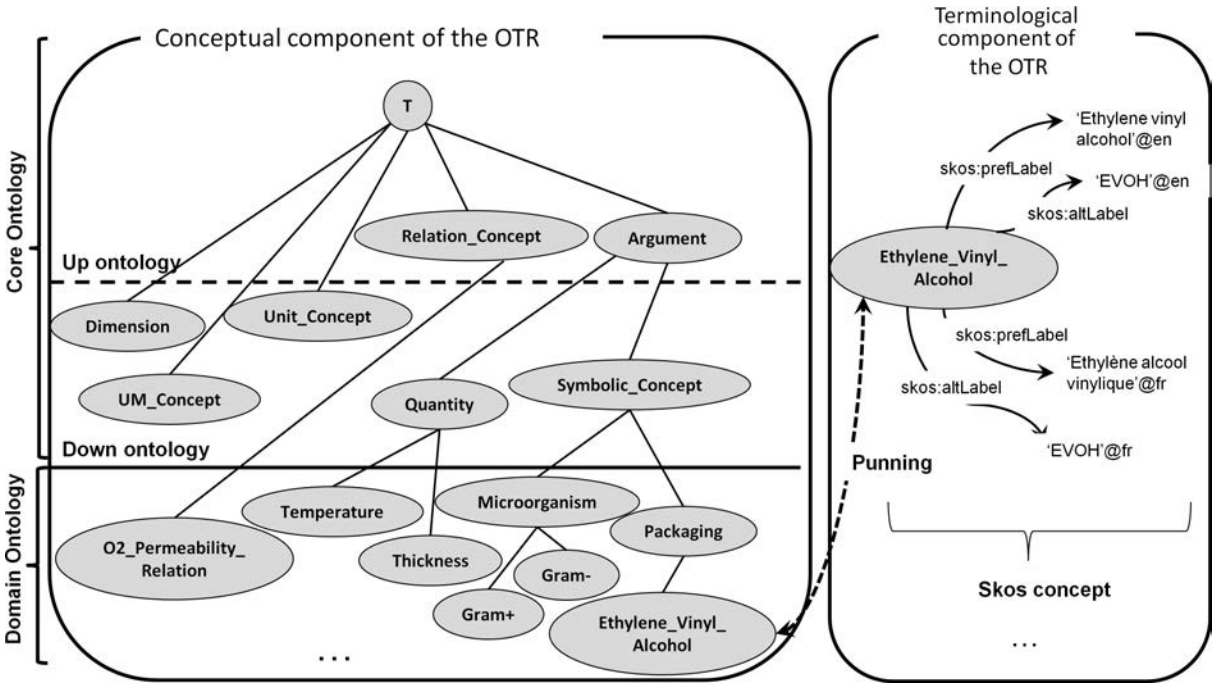


FIGURE 2.7: Un extrait de la composante terminologique dans naRyQ_emb

La modélisation en RTO permet de guider les travaux menés pour proposer des méthodes d'extraction de relations n-aires. Comme nous l'avons décrit, la RTO propose un niveau générique de représentation des relations n-aires et un niveau dépendant du domaine avec sa terminologie associée. Elle représente ainsi une ressource structurée fondamentale pour les travaux menés.

2.2.3 Représentation des données d'intérêt en relation n-aire

Ce modèle générique a été appliqué avec succès, par rapport à l'objectif de ré-utilisation des données modélisées, à deux domaines d'intérêt différents : les caractéristiques d'un emballage et le procédé de transformation d'une biomasse. La RTO naRyQ_emb est modélisée pour représenter les expérimentations permettant de caractériser des emballages alimentaires en termes de perméabilité/diffusivité à l'oxygène (O2), au dioxyde de carbone (CO2) et la solubilité (H2O) à partir de plusieurs paramètres de contrôle, e.g. épaisseur, pression partielle, comme représenté dans la figure 2.3.

Dans le cas de la RTO naRyQ_bioraf, la modélisation permet de représenter les expérimentations correspondant à un enchaînement d'opérations unitaires de prétraitements de la biomasse afin d'en extraire le glucose. Chaque opération unitaire est modélisée sous forme de relation n-aire. Par exemple, la figure 2.8 illustre l'opération unitaire de broyage *Milling Solid sty output relation* composée de paramètres d'entrée pouvant être des flux, i.e la quantité de biomasse brute, ou des paramètres de contrôle du procédé, i.e. durée du traitement. La sortie de l'opération représente le résultat de l'opération unitaire, i.e. la quantité de biomasse broyée.

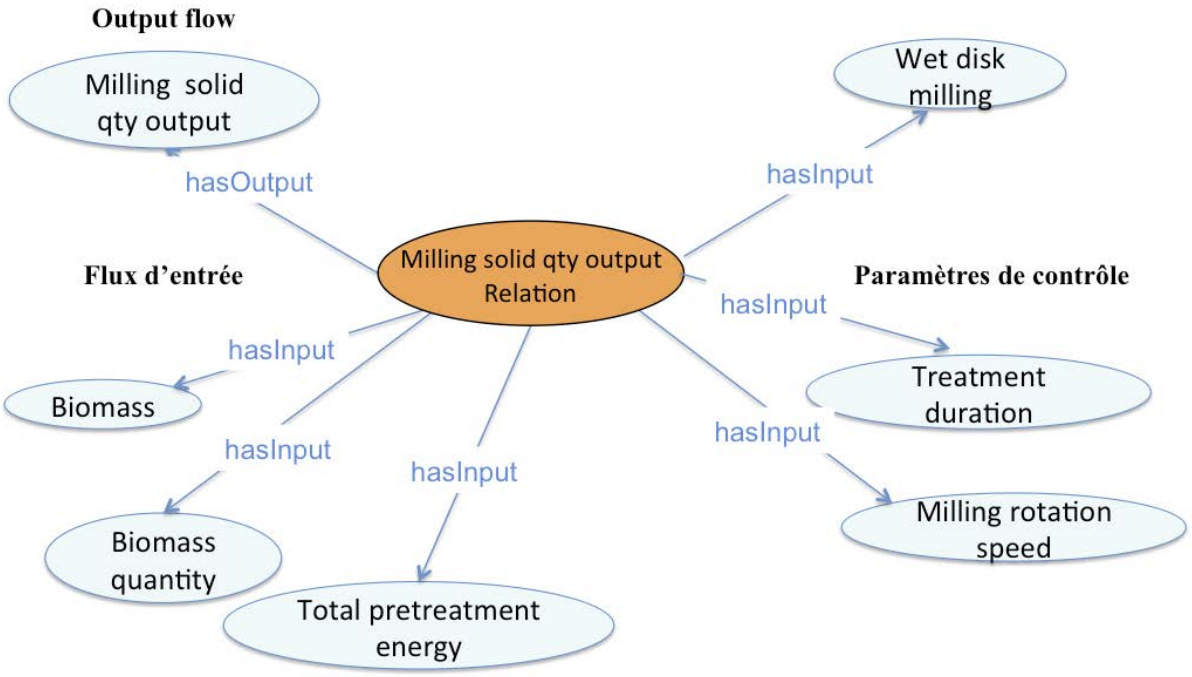


FIGURE 2.8: Représentation de la relation n-aire Milling solid qty output Relation

La robustesse du travail de construction de l'ontologie repose sur le modèle de relations choisi et sur des données expertisées. En effet, la construction de l'ontologie se réalise de manière itérative et collaborative, plusieurs experts confrontent leur opinion concernant les données d'intérêt à identifier et capitaliser. Puis, cette modélisation est instanciée au niveau des textes du corpus. L'instanciation du modèle sur un corpus regroupant les textes de la communauté internationale est une nouvelle confrontation du modèle à la réalité des besoins à grande échelle et, permet ainsi de définir un vocabulaire partagé par l'ensemble de la communauté. Au besoin le modèle est affiné pour permettre l'identification de toutes les données d'intérêt pour le domaine.
Dans le cas de l'annotation des données, la RTO intègre également un ensemble de consignes d'annotation afin de s'entendre sur les décisions à prendre concernant les cas présentant des ambiguïtés. De ce fait, la RTO représente une ressource fiable, de qualité et exhaustive dans le type de données d'intérêt pour le domaine étudié.
Le choix dans le travail de thèse est de s'appuyer sur cette connaissance experte en intégrant la RTO au coeur du travail de construction de méthodes répondant aux problématiques de découverte et d'extraction des relations n-aires recherchées. La RTO dissocie, comme décrit à la section 2.1, la manifestation linguistique (le terme) de la notion qu'elle dénote (le concept). Cela permet de considérer le niveau terminologique de manière plus pertinente et efficace, notamment pour les travaux d'extraction à partir de corpus textuels.

2.3 Définitions et hypothèses de travail

Nous souhaitons contribuer à la problématique de l'enrichissement de la RTO et à l'extraction des instances de relations n-aires dans les textes, en évaluant les méthodes d'apprentissage supervisé et de fouille de données. Si les méthodes proposées sont adaptées aux données structurées, i.e l'extraction de connaissance dans les bases de données, elles ne peuvent s'appliquer directement aux données textuelles non structurées. À partir de cette considération, nous avons défini des éléments méthodologiques, en étant guidés par la connaissance du domaine et en restant proches de la notion de relation n-aire de données quantitatives, afin de proposer une segmentation du texte en unités textuelles, plus adaptées à notre problématique et indispensables pour les méthodes d'apprentissage et de fouille de données. En effet, plutôt que de segmenter le texte selon sa structure naturelle en paragraphes, nous avons souhaité proposer une segmentation plus appropriée à la recherche des données quantitatives dans les textes. Dans ce but, nous présentons, dans la suite de la section, les notions de descripteur pertinent, de phrase pivot et de fenêtre textuelle dans le contexte de recherche des relations n-aires représentant des données expérimentales.

2.3.1 Sélection d'un descripteur pertinent au contexte

Dans le cadre de la représentation des données textuelles, nous avons cherché à sélectionner le descripteur le plus pertinent afin de segmenter les unités textuelles en restant proches de la relation n-aire. Pour cela, nous sommes repartis de la définition de la relation n-aire, dans la RTO naRyQ, afin de cibler ce descripteur pertinent sur lequel va reposer notre représentation textuelle des documents :
L'instanciation de la relation n-aire dans les textes associe des instances d'argument symbolique aux instances d'arguments quantitatifs. Les arguments quantitatifs instanciés associent la valeur numérique de l'instance à une unité de mesure. Par exemple, dans le domaine des emballages, l'instance d'argument symbolique *(EVOH, packaging : (Ethylene Vinyl Alcohol)*, qui est une instance de l'argument concept *Ethylene Vinyl Alcohol* est associée aux arguments quantitatifs, e.g. à sa valeur d'épaisseur *(25 µm, Thickness : (value : 25, unit concept : Micrometer))*. La situation la plus fréquente rencontrée dans les articles scientifiques pour exprimer la mesure d'argument quantitatif est qu'elle ne se fait pas explicitement par le terme dénotant le concept quantitatif, e.g. *thickness, temperature*, mais par l'expression de la valeur mesurée et de son unité de mesure. L'exemple 4 illustre cette situation fréquente, la phrase exprime implicitement la mesure de l'épaisseur de l'emballage. Nous sommes capables de déduire qu'il s'agit de la mesure de l'épaisseur parce que l'unité de mesure proche de la valeur numérique le précise.

Exemple 4
shiitake mushrooms were placed in bags of low density polyethylene film (0.04 mm).

À partir de ce constat, l'unité de mesure apparait clairement comme le descripteur le plus pertinent à choisir pour représenter les données quantitatives impliquant des valeurs numériques et rester ainsi au plus proche de l'instance de relation n-aire.

L'hypothèse de travail 1, que nous posons dans le cadre de la thèse, établit que, à partir des concepts d'unités de mesure de la RTO naRyQ, l'identification du contexte d'expression des instances est favorable. Ce contexte correspond à la découverte la plus probable des instances d'arguments de la relation n-aire modélisée dans la RTO et que nous cherchons à extraire.

Hypothèse 1
Les termes d'unités de mesure référencés dans la RTO représentent un contexte favorable à la découverte d'arguments de la relation n-aire.

Cette hypothèse permet d'ancrer les définitions proposées pour une nouvelle représentation des données afin de segmenter les unités textuelles, comme nous le décrivons dans les sections suivantes.

2.3.2 La phrase pivot

Une fois l'hypothèse 1 posée, en étant guidés par la RTO, nous identifions toutes les unités de mesure référencées dans la RTO. La phrase dans laquelle la (les) unité(s) sont identifiée(s) joue un rôle important car cette phrase, que nous explicitons dans la définition 2, va jouer le rôle de pivot pour déterminer un contexte favorable de découverte de connaissance proche de la RTO de domaine, que nous appelons fenêtre textuelle et que nous définissons dans la section suivante.

Définition 2
Une phrase pivot est une phrase où au moins un terme d'unités de mesure, référencé dans la RTO, est identifié. Elle devient le pivot à la sélection de la fenêtre textuelle, représentant un contexte favorable à la découverte d'arguments de la relation n-aire.

Nous étendons notre recherche au-delà de la phrase pivot car les arguments impliqués dans l'expression des données numériques, e.g. *thickness, temperature*, ne sont pas toujours explicitement présents dans cette même phrase. L'analyse du contexte proche se révèle souvent appropriée à la recherche de la terminologie associée à l'argument quantitatif de la relation n-aire impliqué dans l'expression du résultat numérique. Pour cela nous définissons une nouvelle notion, la fenêtre textuelle.

2.3.3 Sélection d'une fenêtre textuelle

La fenêtre textuelle, dont nous proposons la définition 3, représente une unité textuelle pertinente, c'est-à-dire que le document est représenté selon un ensemble de phrases autour de chaque phrase pivot identifiée. Chaque ensemble, identifié dans le document, représente un contexte phrastique proche de la RTO, favorable à la découverte de connaissances associées à l'expressivité des arguments de la relation n-aire.

Définition 3
La fenêtre textuelle, notée f_n, est l'ensemble des phrases composé de la phrase pivot à laquelle on ajoute les n phrases précédentes et/ou les n phrases suivantes.
n représente la dimension de la fenêtre.

Le sens de parcours, noté s, des phrases est représenté par le signe - pour représenter les phrases précédentes, le signe + pour représenter les phrases suivantes et le signe ± pour représenter les phrases précédentes et suivantes.
La notation choisie pour représenter la fenêtre étudiée est de la forme : f_{sn}

Comme nous le verrons dans le chapitre 3, la définition d'une fenêtre textuelle correspond à un contexte phrastique pertinent à l'extraction de nouveaux variants d'unités de mesure déjà référencées dans la RTO. Dans la section 4.2, nous verrons que la fenêtre textuelle correspond à un contexte phrastique pertinent à la découverte de l'expression des instances d'arguments. Cette approche de segmentation en unités textuelles, en étant guidés par la RTO de domaine, est un atout incontestable car elle permet de situer la recherche de connaissance au niveau le plus pertinent du document, c'est-à-dire au niveau des arguments de la relation n-aire.

Chapitre 3

Extraction des unités de mesure

Sommaire

3.1 Introduction

Nous avons discuté en introduction 1.2 des différents verrous associés à la tâche d'extraction des relations n-aires. Parmi ces verrous identifiés, les unités de mesures subissent de fortes variations terminologiques, comme cela est illustré dans l'Exemple 5 et, mis en évidence dans les travaux de (Ghersedine et al., 2012) sur la RTO du domaine du risque alimentaire microbiologique étendu aux emballages. Une même instance d'unités de l'ontologie peut donc être représentée par des termes d'unités très différents dans les documents, nous les nommons les variants d'unités. Cette section présente les travaux, publiés dans (Berrahou et al., 2013a,b, 2014, 2015) dont l'objectif est de proposer une nouvelle méthode de localisation et d'identification des variants d'unités de mesure.

Exemple 5
Les termes d'unités suivants sont utilisés indifféremment par les auteurs pour exprimer la perméabilité à l'oxygène des emballages :

- $cm^3.\mu m/cm^2.d.kPa$
- $cm^3.\mu m/m^2.d.kPa$
- $cm^3.um/m^2.d.kPa$
- $cm^3.\mu m.m^{-2}.d^{-1}.kPa^{-1}$
- $cm^3/\mu m/cm^2/d/kPa$

L'enrichissement terminologique de la RTO est une étape incontournable afin d'améliorer le processus d'annotation et extraire plus efficacement les instances numériques de la relation n-aire. Ces unités de mesure sont considérées comme de nouvelles unités, plus précisément comme des variants d'unités déjà référencées dans la RTO.

Nous avons ciblé deux verrous scientifiques associés à cette tâche d'extraction des variants d'unités de mesure :

(i) Localiser efficacement les variants dans les documents scientifiques,

(ii) Identifier le variant extrait avec l'unité déjà référencée dans la RTO.

Dans ce chapitre, nous proposons une méthode automatique reposant sur deux étapes afin d'extraire et identifier les variants d'unités de mesure dans le but d'enrichir la RTO. Ces deux étapes consistent à :

1. proposer une méthode s'appuyant sur l'apprentissage supervisé pour réduire l'espace de recherche des variants dans les documents.

2. proposer une nouvelle mesure de similarité adaptée à la problématique d'identification des nouveaux termes extraits.

Dans ce chapitre, nous présentons un état de l'art des travaux menés en extraction de données quantitatives dans des domaines connexes, puis, nous présentons notre proposition, en deux étapes successives, pour localiser et identifier les variants d'unités de mesure pour l'enrichissement d'une RTO.

3.2 État de l'art

L'extraction des données quantitatives est un enjeu majeur pour de nombreux domaines scientifiques dont l'objectif concerne la capitalisation et la pérennisation des connaissances du domaine. Nous nous sommes rapprochés des travaux effectués dans des domaines apparentés à celui des sciences du vivant et avons constaté que les auteurs sont également confrontés au problème d'extraction des données quantitatives, du fait de la forte variation d'écriture des unités de mesure.

Les travaux de (Jessop et al., 2011a,b), dans le domaine de la chimie, présentent une forte variation typographique des unités comme principale cause d'échecs dans le processus d'annotation des données expérimentales en utilisant un outil spécifique *Chemical-Tagger*. Cet outil est dédié aux pré-traitements des documents scientifiques comportant des résultats expérimentaux chimiques (Hawizy et al., 2011). L'outil traite chaque phrase des documents (normalisation, tokenisation, annotation...) pour produire en sortie un fichier XML d'annotation des résultats expérimentaux chimiques. Dans ce processus, l'outil annote les entités chimiques en utilisant OSCAR (Jessop et al., 2011b), un outil d'annotation des entités nommées en chimie. Ensuite, pour annoter certains termes associés aux entités nommées, i.e. certains verbes spécifiques du domaine, les données numériques et les unités de mesure, l'outil utilise des expressions régulières. Les auteurs sont confrontés,

dans cette dernière étape à une forte variation typographique des unités de mesure qui impacte la précision des résultats d'annotation des données quantitatives car, les expressions régulières choisies ne couvrent pas l'ensemble des variations rencontrées.

Dans une démarche consensuelle, le Système International (SI) (Thompson and Taylor, 2008) organise, en posant plusieurs définitions formelles, le système des quantités et des unités de mesure. Il définit ainsi des unités de base, i.e. unités simples comme *kilogram*, et des unités dérivées, i.e. unités plus complexes comme $kg.m^{-1}$. Ce standard pose les règles d'écriture de l'ensemble des unités de mesure mais n'intègre pas la notion de variants d'unités.

Ces principes sont repris dans des travaux récents (Rijgersberg et al., 2013) afin de modéliser formellement cette connaissance dans une ontologie dédiée à la représentation des données quantitatives et des unités de mesure. Les auteurs ont ainsi modélisé *OM* (Ontology of Units of Measure and Related Concepts) qui est, à l'état de l'art, l'ontologie la plus étendue (Rijgersberg et al., 2011). Le travail de modélisation de la partie conceptuelle dédiée aux unités de mesure dans la RTO étudiée dans cette thèse est inspiré de *OM*.

Dans les travaux de (Willems et al., 2012), les auteurs présentent une nouvelle méthode d'annotation des documents pour l'extraction des données quantitatives et des unités de mesure, en utilisant l'Ontologie *OM*. L'approche repose sur l'utilisation des balises disponibles dans les documents Latex. Cette méthode ne répond pas à la problématique d'extraction et d'identification des variants d'unités de mesure dans les documents non structurés.

Les travaux de (Van Assem et al., 2010) posent la problématique d'identification des données quantitatives présentes dans les cellules des tableaux représentés dans les documents. Pour répondre à la problématique d'identification des variants d'unités de mesure, les auteurs utilisent une mesure de similarité, Jaro-Winkler-TFIDF (Cohen et al., 2003). La localisation des variants d'unités ne constitue pas un verrou dans ces travaux car la méthode repose sur le format structuré des tableaux. Les variants étant directement identifiés dans les cellules du tableau étudié. La recherche des variants ne se poursuit pas dans le reste du document. Contrairement à ces travaux, les variants d'unités de mesure à localiser et identifier dans nos travaux sont contenus dans le texte libre.

Les travaux de (Grau et al., 2009) proposent des méthodes d'extraction des données expérimentales dans le domaine biomédical. L'identification des unités de mesure repose

sur les unités référencées dans l'ontologie *units.obo*[1], dans le format de *Gene Ontology*[2] sans problématique d'identification des variants d'unités de mesure. Comme évoqué précédemment, nous préférons nous appuyer sur la modélisation conceptuelle de l'ontologie *OM*, la plus étendue à l'état de l'art et proche de nos préoccupations de variants d'unités.

A notre connaissance, les méthodes décrites à l'état de l'art des domaines partageant l'objectif d'extraction de données quantitatives, ne permettent pas de résoudre la problématique de localisation et d'identification des variants d'unités de mesure dispersés dans les documents scientifiques au format textuel non structuré.
Dans ce chapitre, nous proposons une méthode automatique reposant sur deux étapes pour répondre à la problématique de localisation abordée en section 3.3, et d'identification des variants d'unités de mesure, abordée en section 3.4.

3.3 Localisation des unités de mesure

3.3.1 Méthodologie

3.3.1.1 Contexte et processus global

Nous avons montré que la forte variation d'écriture des unités de mesure dans les documents engendre des problèmes d'identification des instances numériques de la relation n-aire. L'enrichissement de la RTO avec des variants d'unités de mesure est une étape fondamentale dans le processus global d'extraction des instances de relations n-aires.
Dans cette section, nous présentons notre proposition qui tente de répondre à deux questions concernant l'extraction des variants d'unités de mesure dans les documents textuels non structurés :

- La question concernant la localisation des variants dans le document. Sachant que nous travaillons sur l'intégralité des documents, nous préférons l'approche par apprentissage afin de prédire la localisation des variants en favorisant une recherche exhaustive des variants par classification automatique ;

- La question de l'identification du variant une fois qu'il est localisé. A quel autre terme d'unité de mesure référencée dans la RTO peut-on le rapprocher, en sachant que les termes d'unités répondent à leurs propres règles d'écriture ? Les méthodes

1. OBO : Open Biomedical Ontologies
2. http ://www.geneontology.org/

existantes doivent être adaptées à ces nouvelles règles.

Pour répondre à ces questions, notre contribution repose sur une méthode automatique de localisation et d'identification des variants d'unités de mesure fondée sur deux étapes :

- La première étape s'appuie sur une méthode employant l'apprentissage supervisé pour prédire la localisation des variants d'unités de mesure. Pour ce faire, nous proposons une nouvelle représentation textuelle adaptée au contexte des données quantitatives. Cette nouvelle représentation des données est guidée par la RTO et propose une approche linguistique selon différentes fenêtres textuelles, favorables à la découverte de variants d'unités. Puis, nous utilisons la classification automatique pour prédire si les phrases représentées dans un certain contexte textuel (fenêtre textuelle) sont susceptibles de contenir des variants d'unités de mesure. L'apprentissage permet de déterminer quelle fenêtre textuelle et quel algorithme produit le meilleur modèle pour prédire la localisation des variants ;

- La seconde étape concerne l'identification des variants localisés, en proposant une nouvelle mesure de similarité adaptée aux règles spécifiques des unités de mesure. La nouvelle mesure permet de rapprocher le variant découvert à une autre unité déjà référencée dans la RTO à partir d'un score de similarité.

3.3.1.2 Représentation des données textuelles adaptée au contexte d'étude

La première étape de notre méthode consiste à prédire la localisation des variants d'unités de mesure dans les documents. Nous avons utilisé une méthode reposant sur l'apprentissage supervisé et s'appuyant sur des techniques de classification textuelle qui combine une approche linguistique et une approche statistique pour représenter les données.

Pré-traitements linguistiques. L'approche de pré-traitements linguistique consiste à préparer le corpus. Les pré-traitements choisis et exécutés séquentiellement pour préparer nos documents sont illustrés dans la phase 1 de la figure 3.1 :

— Segmentation des documents en phrases,
— Suppression de la ponctuation que nous avons adaptée sachant que la syntaxe des unités intègre de nombreux caractères communs aux signes de ponctuation, afin de

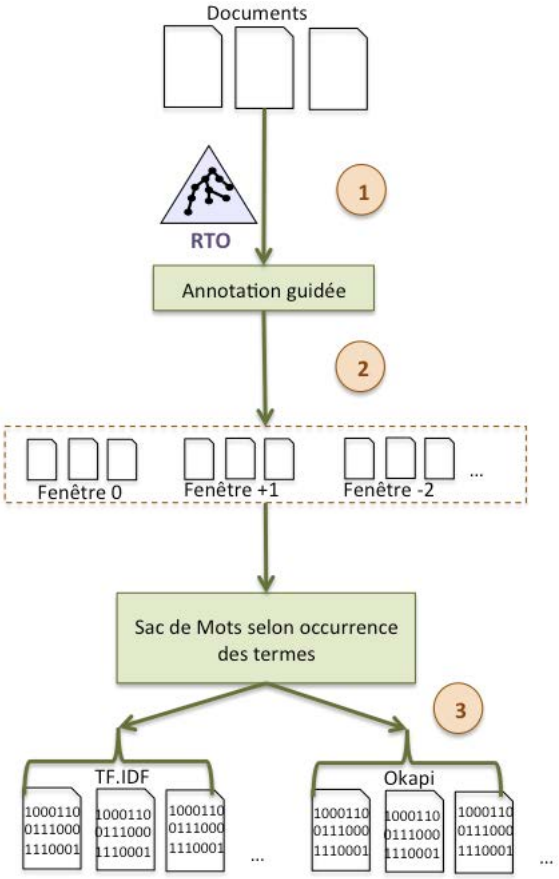


FIGURE 3.1: Représentation textuelle adaptée au contexte

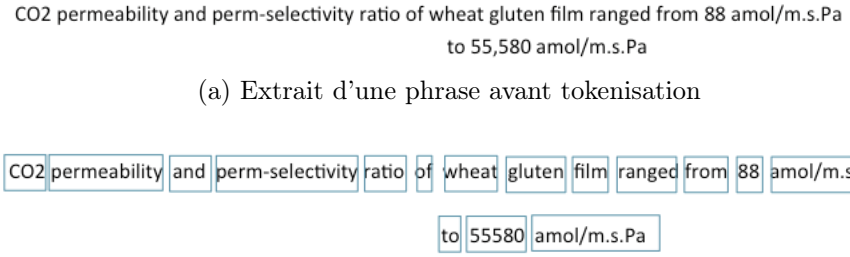
ne pas les endommager. Le résultat, illustré dans l'exemple 6, montre que les virgules et le point final sont supprimés, en revanche, les signes de ponctuation distinctifs des entités de domaine (i.e. *soybean-protein-isolate*) des valeurs numériques (i.e. 19.2) et des unités de mesure (i.e. $g/m \cdot d \cdot MPa$ et $cm\hat{}^3/m \cdot d \cdot MPa$) sont conservés,

Exemple 6
Avant traitement :
The soybean-protein-isolate films based on the optimal composite emulsifiers show their tensile strength of 908 MPa, percentage elongation at break of 25.8%, water vapor permeability of 19.2 $g/m{\cdot}d{\cdot}MPa$, and oxygen permeability of 0 $cm\hat{}^3/m{\cdot}d{\cdot}MPa$, being stronger than the control.

Après traitement :

The soybean-protein-isolate films based on the optimal composite emulsifiers show their tensile strength of 908 MPa percentage elongation at break of 25.8% water vapor permeability of 19.2 $g/m \cdot d \cdot MPa$ and oxygen permeability of 0 $cm\hat{}^3/m \cdot d \cdot MPa$ being stronger than the control

— Tokenisation des phrases, comme illustré dans la Figure 3.2,



(a) Extrait d'une phrase avant tokenisation

(b) Après tokenization

FIGURE 3.2: La tokenisation

— Suppression des mots fonctionnels de la phrase, e.g. les articles et prépositions,

— Annotation automatique des phrases en projetant les termes d'unités référencées dans la RTO : dans notre contexte d'étude, nous avons choisi cette annotation en nous reposant sur l'hypothèse 2 dérivant de l'hypothèse 1 pour déterminer l'élément pertinent à annoter automatiquement dans le texte. Nous n'avons pas souhaité utiliser d'autres pré-traitements linguistiques. En effet, les traitements consistant à étiqueter grammaticalement les mots du texte peuvent engendrer du bruit à partir de nos données scientifiques très spécialisées. Nous préférons nous appuyer sur la connaissance du domaine afin de produire l'annotation automatique adaptée au contexte d'étude des données quantitatives.

Hypothèse 2
Les termes d'unités de mesure référencés dans la RTO indiquent un contexte favorable de découverte de variants d'unités.

Définition du contexte phrastique. À partir de nos définitions 2 et 3, posées dans la section 2.3, nous définissons une nouvelle représentation des documents sous forme de fenêtres textuelles, représentant un contexte phrastique précis. Contrairement à l'approche classique, cette nouvelle représentation est pertinente pour notre contexte d'étude des données expérimentales. En effet, sachant que le processus est guidé par la RTO, nous

restons toujours au plus proche des définitions de la relation n-aire et donc des instances quantitatives que nous recherchons. Chaque fenêtre textuelle est évaluée indépendamment comme cela est représenté dans la phase 2 de la figure 3.1.
En sortie d'annotation automatique du corpus, à partir de chaque phrase des documents contenant une (ou plusieurs) unité(s) de mesure annotée(s) automatiquement, nous pouvons représenter le texte selon différentes fenêtres, comme illustré dans l'exemple 7, à partir d'une phrase (a) pré-traitée et annotée automatiquement.

Exemple 7
(b) system programmed 2 h waiting period 10 cycles readings 2 h allow films achieve equilibrium
(a)Oxygen permeability calculated dividing O2 transmission rate difference O2 partial pressure sides film multiplying average film thickness measured random places oxygen permeability reported $cm^{\hat{}3}$ *<um>µm</um>* $m^{\hat{}-2}$ $day^{\hat{}-1}$ $kPa^{\hat{}-1}$ *units*
(c)Three replicates film evaluated
L'unité de mesure µm est identifiée, la phrase (a) est, selon la définition 2, considérée comme une phrase pivot. La représentation textuelle est selon la définition 3 :
f_0 *contient (a) seulement*
$f_{\pm 1}$ *contient (b)(a)(c)*
f_{-1} *contient (b)(a)....*

Représentation *Sac de mots*. Dans notre contexte d'étude, pour chacune des fenêtres, nous avons sélectionné les termes apparaissant fréquemment (occurrence ≥ 2) dans le corpus dédié à l'apprentissage pour constituer le sac de mots. Ces termes sélectionnés correspondent aux descripteurs utilisés pour représenter les documents sous forme vectorielle. La constitution du sac de mots réduit sensiblement l'espace de représentation des textes sans avoir pour autant un impact sur les résultats de la classification. La sélection des descripteurs pour représenter nos documents se fait par rapport à une fenêtre textuelle, définissant un contexte phrastique précis, proche des instances quantitatives définies dans la relation n-aire recherchée. De ce fait et comme nous le discutons dans la section 3.3.2.3, les termes sélectionnés pour représenter les documents sont beaucoup plus pertinents.

Pondération des descripteurs et représentation vectorielle. Les termes ainsi présélectionnés sont projetés sur chaque document. Cette projection consiste à savoir si le document contient ou pas les descripteurs du sac de mots, et ainsi produire son vecteur de représentation. Elle est illustrée par la phase 3 dans la figure 3.1. L'ensemble des vecteurs constitue la matrice vectorielle 3.3, nouvelle représentation des documents, adaptée aux

algorithmes d'apprentissage supervisé.

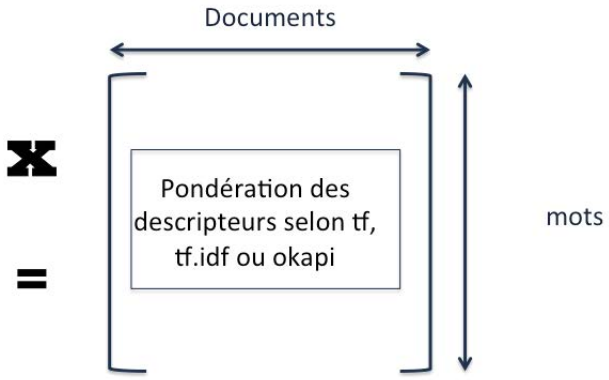


FIGURE 3.3: Représentation vectorielle

Dans notre contexte d'étude, la phase 3 consiste à projeter les descripteurs sélectionnés dans un contexte phrastique précis, proche de la RTO, sur les fenêtres textuelles évaluées. Nous obtenons autant de matrices vectorielles qu'il y a de fenêtres textuelles à évaluer. En effet, la matrice vectorielle décrite dans la figure 3.3 ne représente plus les documents mais les phrases contenues dans une fenêtre textuelle, représentant un contexte phrastique spécifique.

Hormis la représentation booléenne des descripteurs (présence/absence), les descripteurs du sac de mots peuvent être pondérés selon différentes approches statistiques pour leur représentation vectorielle. Les références bibliographiques de ces mesures sont données dans (Claveau, 2012). Nous présentons les principales mesures de la littérature que nous utilisons dans nos travaux :

- *Term Frequency, tf* considère que plus le terme est représenté dans le document, plus il est important
  Pour chaque terme t et le document d, le calcul de f_{dt} représente la fréquence d'occurrences du terme t dans le document d.
  n représentant le nombre total de mots dans le document d ;

$$tf_{dt} = \frac{f_{dt}}{n}$$

- *Term Frequency.Inverse Document Frenquency, tf.idf* : est une mesure établissant un compromis entre le *tf*, la fréquence locale du terme dans le document, et la fréquence globale du terme dans la collection de documents, ayant ainsi une meilleure représentativité (*IDF*). En d'autres termes, plus un terme apparait dans un grand nombre de documents moins il est intéressant à considérer (Salton et al., 1993) ;

$$w_{tf-idf}(t,d) = tf_{dt} * \log(N/df_t)$$

Avec :
tf_{dt} le nombre d'occurrences de t dans d
df_t le nombre de documents contenant le terme t
N le nombre total de documents

- *Okapi BM25* (Jones et al., 2000) : est également une mesure de pondération des termes des documents mais intégrant une normalisation en fonction de la taille des documents.

$$w_{BM25}(t,d) = \frac{tf_{dt} * (k_1 + 1)}{tf_{dt} + k_1 * (1 - b + b * dl_d/dl_{avg})} * \log(\frac{N - df_t + 0.5}{df_t + 0.5})$$

Avec :
$k_1 = 2$ et $b = 0.75$ des constantes choisies par l'auteur
dl_d la longueur du document d
dl_{avg} la longueur moyenne des documents

En somme, nos documents sont transformés en plusieurs matrices vectorielles, selon la représentation du sac de mots sous forme booléenne ou pondérés par les différentes mesures *tf, tf.idf* ou *okapi*. Ces matrices vectorielles utilisent le format adapté aux algorithmes d'apprentissage présentés dans la section suivante.

3.3.1.3 Prédiction des localisations par apprentissage supervisé

Présentation des principaux algorithmes Nous choisissons d'évaluer les fenêtres textuelles, représentant un contexte phrastique spécifique, en utilisant les algorithmes de classification supervisée. On attribue automatiquement à chaque vecteur une classe selon que le contexte phrastique contient une unité de mesure annotée, classe **unit**, ou qu'il n'en contient pas, classe **non-unit**. Notre objectif, à partir des résultats de classification obtenus, est de produire un modèle d'apprentissage, selon la meilleure fenêtre évaluée, pour prédire si de nouvelles instances testées par le modèle peuvent contenir des variants d'unités ou pas.
Nous avons choisi d'évaluer les fenêtres avec plusieurs algorithmes d'apprentissage supervisé, connus de la littérature, afin d'obtenir une base comparative suffisamment exhaustive. Les algorithmes sont jugés par rapport à leur comportement et leur performance de classification sur les données textuelles. Ils sont implémentés dans l'environnement weka [3] :

— Naive Bayes classifier (John and Langley, 1995) est un algorithme efficace en termes de classification et de performance de calcul. Il est qualifié de "naïf" car son principe repose sur l'indépendance des variables. Il utilise des conditions de probabilité observées sur les données pour déterminer la classe à attribuer aux jeux de données, parmi un ensemble d'exemples. À partir de l'exemple 8, nous présentons les paramètres pris en considération par l'algorithme Naive Bayes afin d'attribuer une classe :

Exemple 8
Considérons deux ensembles : un ensemble de 20 objets classés a et un ensemble de 40 objets classés b.
Supposons que nous souhaitions catégoriser un nouvel objet et savoir si cet objet appartient à la classe a ou à la classe b.
Sachant qu'il y a plus d'objets de la classe b, il est fort probable que le nouvel objet soit de la classe b. Cette supposition est appelée la probabilité antérieure (pa). Elle est issue de l'expérience précédente et utilisée pour prévoir les résultats.
On peut ainsi établir que :

$$pa_b = \frac{b}{a+b} = \frac{40}{60}$$
$$pa_a = \frac{a}{a+b} = \frac{20}{60}$$

3. http ://www.cs.waikato.ac.nz/ml/weka/

Puis, on affine cette supposition, en disant que la probabilité que l'objet soit de la classe a ou b dépend des objets qui se trouvent dans son entourage immédiat. Pour cela, l'algorithme considère un ensemble quelconque d'objets, défini préalablement, autour du nouvel objet à catégoriser, et va calculer la vraisemblance qu'il soit de la classe a ou de la classe b :

Soit x le nombre d'objets a (dans l'exemple x=3) dans son entourage et y le nombre d'objets b (dans l'exemple y=1) dans son entourage,

$$vrai_a = \frac{x}{a} = \frac{3}{20}$$
$$vrai_b = \frac{y}{b} = \frac{1}{40}$$

Il apparait clairement que la probabilité antérieure et la vraisemblance se contredisent, on calcule alors la probabilité finale, qui combine les deux mesures pour former la probabilité postérieure pp :

$$pp_a = pa_a * vrai_a = \frac{1}{3} * \frac{3}{20} = 0.05$$
$$pp_b = pa_b * vrai_b = \frac{2}{3} * \frac{1}{40} = 0.0167$$

Le nouvel objet est finalement catégorisé en classe a.

Plus généralement pour la classification automatique des documents, l'hypothèse à poser est la catégorie C à prédire pour le document D à classer. Le classifieur doit choisir à partir d'un ensemble de catégories candidates, la catégorie $c \in C$ la plus probable à affecter au document $d \in D$

$$P(C|D) = \frac{P(D|C) * P(C)}{P(D)}$$

Au cours de l'apprentissage, il peut y avoir plusieurs hypothèses à choisir (plusieurs catégories). Dans ce cas, cela signifie que l'algorithme fait un choix ayant la probabilité maximale, comme illustré dans l'exemple 8, pour attribuer la catégorie à un nouvel objet, en calculant la probabilité postérieure.

Sachant que ce sont les mots qui constituent un document, l'algorithme utilise la probabilité de tous les mots du document en question étant donné une catégorie :

$$P(c|D) = P(m_1|c) * P(m_2|c)....P(m_n|c) \ (1)$$

Dans notre nouvelle approche de représentation des données, il récupère en entrée l'ensemble des descripteurs choisis pour constituer le sac de mots à partir des fenêtres textuelles. Pour chaque mot, il calcule le nombre de fois où il apparait dans le document ou la fenêtre textuelle dans une classe donnée, la classe "unit" ou "non-unit". Il calcule cette fréquence pour chaque classe considérée.

Dans une première phase, l'algorithme apprend à classer à partir d'un ensemble d'apprentissage :

Apprend_NB($\mathcal{E}$, $\mathcal{C}$)

Soit $\mathcal{E}$ un ensemble de documents ayant une classe attribuée.
Soit $\mathcal{C}$ l'ensemble des classes considérées.

L'algorithme Naive Bayes calcule $P(w_k|c_j)$, la probabilité qu'un mot du document ayant pour classe c_j soit le terme w_k. Il calcule également la probabilité $P(c_j)$.

1. Soit $\mathcal{S} \leftarrow$ l'ensemble des descripteurs du sac de mots définis à partir de $\mathcal{E}$
2. Pour chaque classe c_j de $\mathcal{C}$:
   doc$s_j \leftarrow$ est le sous-ensemble de $\mathcal{E}$ ayant pour classe c_j.
   $P(c_j) \leftarrow \dfrac{|docs_j|}{|\mathcal{E}|}$
   Pour tous les termes w_k de $\mathcal{S}$:

   - $n_k \leftarrow$ est le nombre d'occurrences du terme dans la classe c_j
   - n est le nombre total de mots dans le corpus d'entrainement

$$P(w_k|c_j) \leftarrow \frac{n_k + 1}{n + |\mathcal{S}|}$$

Cette estimation est préférée au calcul de la probabilité de tous les mots du document étant donné la catégorie, définie par l'équation (1) :
En effet, cette équation présente un inconvénient majeur : si un attribut n'apparait pas dans chacune des catégories dans la phase d'entrainement, sa probabilité est nulle et cela engendre une probabilité globale nulle.
Pour éviter cette problématique de nullité de l'équation, on ajoute la valeur 1 au numérateur et la valeur n au dénominateur.

Dans une seconde phase, l'algorithme utilise le modèle et classe de nouvelles instances :

Classe_ NB(Doc)

L'algorithme retourne la classe pour le document *Doc*

1. a_i correspond au terme contenu dans le sac de mots et trouvé à la ième position dans le *Doc*

2. $p \leftarrow$ toutes les positions des termes de *Doc* trouvés dans $\mathcal{S}$

$$c_{NB} = argmax_{c_j \in \mathcal{C}} P(c_j) \prod_{i \in p} P(a_i|c_j)$$

— Discriminative Multinominal Naive Bayes (DMNB) classifieur (Su et al., 2008) est un algorithme de classification textuelle qui étend et améliore l'efficacité de l'algorithme Naive Bayes précédemment décrit. L'algorithme DMNB est un algorithme performant, rapide en exécution et compétitif par rapport à d'autres classifieurs discriminants, e.g. SMO, et qui supporte les contraintes liées aux données textuelles : des données de grandes dimensions, souvent éparses, causant des difficultés en termes d'efficacité et de rapidité de calcul. Il fonctionne de manière itérative sur les documents d'entrainement :
Pour tout document d^t, il calcule la probabilité postérieure ou conditionnelle $\hat{p}(c|d^t)$, puis l'algorithme met à jour la fréquence correspondante en utilisant la différence

entre la probabilité vraie $P(c|d^t)$ (la vraisemblance) et la probabilité postérieure $\hat{p}(c|d^t)$.
Cette différence est appelée la perte prédictive, notée $L(d^t)$, pendant l'entrainement des documents d^t :

$$L(d^t) = P(c|d^t) - \hat{p}(c|d^t)$$

En résumé, l'algorithme DMNB va mettre à jour itérativement la fréquence des mots en s'appuyant sur la probabilité conditionnelle mais en tenant compte de la perte prédictive au cours de l'entrainement des documents :

1. DMNB crée d'abord une table de fréquences vide, puis

2. Pour tout t de 1 à $|T|$:

   - Il sélectionne aléatoirement un document d^t du corpus d'apprentissage T

   - Il estime la probabilité des paramètres en utilisant $\hat{p}(w_i|c) = \frac{f_{ic}}{f_c}$, où f_{ic} est le nombre d'occurrences de w_i dans les documents de la classe c, et f_c est le nombre total d'occurrences de mots dans les documents de la classe c

   - Il calcule les fréquences f^t_{ic}

   - Il calcule la probabilité postérieure $\hat{p}(c|d^t)$

   - Il calcule la perte prédictive $L(d^t)$ en utilisant $L(d) = P(c|d) - \hat{p}(c|d)$

   - Pour tout mot w_i ayant un nombre d'occurrences différent de 0 dans le document d^t :

- Il considère la f_{ic}^{t}, calculée précédemment (la fréquence du mot w_i dans le t^{ieme} document d^t)

- Il met à jour la fréquence f_{ic}^{t+1}, en tenant compte de la perte prédictive sur le calcul de la fréquence dans le document précédent $f_{ic}^{t+1} = f_{ic}^{t} + L(d^t) * f_{ic}^{t}$

— J48 Decision tree classifier est l'implémentation de l'arbre de décision C4.5 (Quinlan, 1993) qui étend l'algorithme initial ID3 (Quinlan, 1979) du même auteur. Cet algorithme est reconnu comme une référence dans la littérature de fouille de données (Kohavi and Quinlan, 2002). L'arbre est une structure permettant d'obtenir un résultat à partir de décisions successives. Pour trouver un résultat, le parcours de l'arbre s'effectue depuis la racine. Chaque noeud correspond à une décision. On descend de proche en proche sur chaque noeud pour atteindre la feuille qui correspond à la réponse de l'arbre au cas testé.

Dans le cas de la classification, il faut construire un arbre par catégorie (ou classe). Chaque nouveau document à classer est soumis à l'arbre de chaque catégorie et l'algorithme prend une décision. Pour déterminer à quelle(s) catégorie(s) appartient un nouveau document, on utilise l'arbre de décision de chaque catégorie auquel on soumet le document à classer. Chaque arbre répond Oui ou Non. Chaque exemple est un ensemble attributs/valeurs, i.e. un terme peut être un attribut et sa valeur peut-être un booléen (absence ou présence) de type 0 ou 1, selon que le terme appartient ou non au document.
Pour construire l'arbre de décision, un attribut est testé à chaque noeud selon un processus récursif et pour déterminer quel attribut tester à chaque étape, on utilise un calcul statistique qui détermine dans quelle mesure cet attribut sépare bien les classes. Initialement, l'algorithme ID3 utilise la mesure statistique appelée *Information Gain*, fondée sur l'entropie :

$$Entropie(S) = -\frac{p}{N} * log_2(\frac{p}{N}) - \frac{n}{N} * log_2(\frac{n}{N})$$

où :
S est l'ensemble des N exemples

p est le nombre d'exemples classés Oui par l'arbre de l'ensemble S
n est le nombre d'exemples classés Non par l'arbre de l'ensemble S

L'entropie permet de mesurer l'homogénéité des exemples. Si l'entropie vaut 0, tous les exemples appartiennent à la même classe, si l'entropie vaut 1, il existe autant d'exemples positifs que d'exemples négatifs. La mesure *Information Gain* calcule la réduction de l'entropie si un attribut particulier est utilisé. L'algorithme calcule cette mesure pour chaque attribut, pouvant avoir une valeur différente selon la classe considérée. Il choisit ensuite celui qui permettra de séparer les exemples le plus nettement possible, c'est-à-dire celui qui réduit le plus l'entropie.

$$InformationGain(A,S) = Entropie(S) - Entropie(A,S)\ InformationGain(A,S) = Entropie(S) - \sum_{v \in \mathcal{D}_A}(|S_v| * \frac{Entropie(S_v)}{|S|})$$

où :
S est l'ensemble des exemples
$\mathcal{D}_A$ est le domaine de valeurs de A et représente l'attribut choisi
S_v est le sous-ensemble de S dont l'attribut A a la valeur v

ID3 présente néanmoins quelques problèmes car il nécessite des données exhaustives avec une quantité de calcul assez importante. Il crée de la redondance dans les tests des différents sous-arbres et produit des sous-arbres encombrants qui ne sont pas forcément exploités.
Des améliorations ont été apportées dans l'algorithme C4.5 et tiennent compte des points suivants, particulièrement importants s'agissant de classification de données textuelles :

- La prise en compte des attributs à valeurs continues,
- La prise en compte des attributs manquants ou à valeurs inconnues,
- La capacité à prendre en compte des attributs ayant des poids différents,
- L'élagage de l'arbre en cas de besoin, en remplaçant tout un sous-arbre par une feuille.

— Sequential Minimal Optimization (SMO) (Platt, 1999) est un algorithme appartenant à la famille des Machines à Vecteurs de Support (SVM).
Les classifieurs des SVM utilisent une fonction qui construit une séparation optimale des données, un hyperplan.

Considérons une base d'exemples $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$
où x_i est un vecteur dans un espace $\mathcal{X} \in \mathbb{R}^N$ et $y_i \in \{-1, +1\}$

Lorsque les données sont linéairement séparables, l'algorithme construit un hyperplan qui sépare les exemples positifs des exemples négatifs selon la fonction :

$f(x) = a.x + b = 0$
où $a \in \mathbb{R}^N, x \in \mathcal{X}, b \in \mathbb{R}$ et . est le produit scalaire

Les SVM ont pour avantage de construire un hyperplan optimal qui maximise la distance entre les exemples et le classifieur. Les points les plus proches sont utilisés pour la détermination de l'hyperplan optimal et sont appelés vecteurs de support (SV). La figure 3.4 montre que l'hyperplan séparateur est défini par la fonction

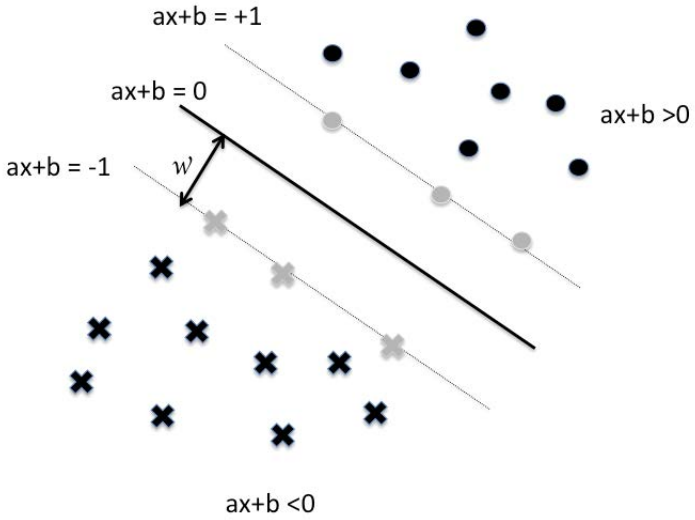


FIGURE 3.4: Hyperplan séparateur et marge

$(a.x + b = 0)$ et la fonction de décision pour attribuer la classe est le signe de $(a.x + b)$. La marge $\mathcal{W}$ est la distance minimale des points avec l'hyperplan. Les points situés sur les lignes pointillées sont les SV.

Pour traiter des exemples non séparables linéairement, l'algorithme utilise une fonction noyau $K(t, u)$, qui permet de projeter les exemples dans un espace de grande dimension et ramener le problème de séparation à un cas linéairement séparable. L'algorithme d'optimisation séquentielle minimale SMO a été développé spécifiquement afin de pouvoir manipuler des ensembles de très grande dimension. Il existe plusieurs fonctions noyaux, i.e. noyau linéaire, noyau polynomial de degré p, noyau gaussien.
SMO, utilisé dans le cadre de nos expérimentations, fonctionne avec un noyau polynomial. Ce noyau est efficace pour donner de bons résultats de classification sur les données textuelles car il a la capacité de prendre en charge leurs dimensions élevées, leur grand nombre d'attributs et leurs vecteurs fréquemment clairsemés.

Il existe d'autres algorithmes en apprentissage supervisé, nous avons choisi de présenter les approches classiques de la littérature que nous utilisons au cours des expérimentations.

Critères d'évaluation. Les critères d'évaluation permettent de comparer les différents classifieurs testés. Cependant, sachant qu'il n'existe pas de définition formelle de l'appartenance d'un document à une catégorie, du fait du caractère subjectif relevant de la sémantique des textes, les méthodes d'évaluation s'appuient sur l'expérience pour évaluer les documents. Nous utilisons le principe de validation croisée, en séparant le corpus de textes en plusieurs ensembles afin d'évaluer les décisions prises par le classifieur automatique :
On divise l'ensemble des textes en k groupes égaux. À chaque tour, l'algorithme utilise les $k - 1$ groupes pour s'entrainer et valide le résultat sur le dernier groupe. Il effectue autant de validations qu'il y a de groupes et, à chaque nouvelle étape, un nouveau groupe joue le rôle pour la validation des résultats.
La table de contingence, représentée dans le tableau 3.1, permet alors de comparer les décisions prises par le classifieur automatique en calculant plusieurs paramètres, i.e. la précision, le rappel et la F-mesure.

- La précision correspond au nombre de résultats correctement extraits par rapport au nombre de résultats extraits, soit :

$$P = \frac{a}{(a + b)}$$

	Documents appartenant à la catégorie	Documents n'appartenant pas à la catégorie
Documents assignés à la catégorie par le classifieur	a	b
Documents rejetés de la catégorie par le classifieur	c	d

Tableau 3.1: Table de contingence.

- Le rappel, que l'on peut considérer comme mesurant l'exhaustivité des résultats de l'extraction, correspond au calcul du nombre de résultats correctement extraits par rapport au nombre de résultats à extraire, soit :

$$R = \frac{a}{(a+c)}$$

- La F-mesure correspond à une fonction qui est maximisée lorsque la précision et le rappel sont proches et qui est donc intéressante à suivre afin de connaitre la qualité du classifieur déterminée par l'équilibre que l'on recherche entre la précision et le rappel. Elle est donnée par la moyenne harmonique pondérée :

$$F_\beta = \frac{(1+\beta^2) * R * P}{\beta^2 * R + P}$$

Communément, β=1 pour donner le même poids à la précision et au rappel

Nous avons vu que la RTO permet de guider la représentation des données sous forme de fenêtres textuelles, représentant un contexte à évaluer par les algorithmes de classification présentés. Le but des expérimentations proposées dans la section suivante est de proposer un modèle permettant de prédire, pour un nouveau texte du domaine, si le contexte validé est susceptible de contenir un variant d'unité de mesure.

3.3.2 Expérimentations

3.3.2.1 Protocole expérimental

Les expérimentations ont été menées sur deux corpus différents en anglais, sélectionnés selon une recherche par mots clés définis par les experts, à partir de bibliothèques en ligne (e.g. Elsevier, Springer) : un corpus de 115 articles scientifiques issus du domaine des emballages alimentaires en microbiologie prévisionnelle et un corpus de 243 articles scientifiques issus du domaine de la bioraffinerie. Les expérimentations s'appuient sur une liste de 211 termes dénotant les différents concepts d'unités de mesure pour le domaine des emballages alimentaires et une liste de 36 termes pour le domaine de la bioraffinerie. Nous restituons en résultats les expérimentations sur les fenêtres d'étude les plus pertinentes[4] dans le tableau 3.2 :

— f_0 : représente la fenêtre comportant la phrase où au moins un terme d'unité dénotant un concept de la RTO est identifié,

— f_{+2} : représente la fenêtre comportant la phrase où au moins un terme d'unité dénotant un concept de la RTO est identifié ainsi que les deux phrases suivantes,

— f_{-2} : représente la fenêtre comportant la phrase où au moins un terme d'unité dénotant un concept de la RTO est identifié ainsi que les deux phrases précédentes.

Ces différentes fenêtres correspondent à des sous-ensembles du corpus représentant 5 000 phrases (e.g. f_0) à 15 000 phrases (e.g. f_{+2}). Le corpus complet comportant plus de 35 000 phrases pour le corpus des emballages. Le corpus complet de bioraffinerie comporte un peu plus de 115 000 phrases dont environ 11 000 sont susceptibles de contenir des variants d'unités. Le sac de mots représente un ensemble de 3 000 à 4 800 descripteurs selon les différentes représentations. Concernant les critères d'évaluation des résultats, nous avons utilisé les critères décrits dans le paragraphe 3.3.1.3 et réalisé une 10-validation croisée. Dans le but de tester la généricité du modèle produit, nous avons établi un protocole en deux étapes : la première étape consiste à produire le modèle appris à partir des données du premier corpus et dans une seconde étape, d'appliquer ce modèle appris sur les données du deuxième corpus.

4. Les résultats sur les fenêtres f_{+1}, f_{-1}, $f_{\pm 1}$ sont très proches de la fenêtre f_0, le résultat sur la fenêtre $f_{\pm 2}$ n'apporte pas d'éléments supplémentaires à ceux présentés et les résultats sur les fenêtres supérieures ne se sont pas révélés appropriés

3.3.2.2 Résultats

Le corpus des emballages a servi, dans un premier temps, à apprendre un modèle performant pour découvrir la localisation des variants d'unités, en fonction des différentes fenêtres textuelles proposées. Une fois ce modèle obtenu, nous l'avons appliqué sur le corpus de bioraffinerie afin d'évaluer sa généricité et sa performance à découvrir des variants d'unités dans un domaine différent mais dont la problématique d'identification des variants d'unités est similaire. Les tableaux 3.2 et 3.3 restituent les résultats obtenus sur le corpus des emballages. Le tableau 3.4 restitue les résultats obtenus sur le corpus de bioraffinerie en appliquant le modèle obtenu et validé en phase d'apprentissage sur le corpus des emballages.

Le tableau 3.2 restitue les résultats de classification sur différentes fenêtres textuelles en fonction des algorithmes choisis, toutes mesures de pondération confondues. Cette première évaluation permet de juger quelle fenêtre textuelle est la plus pertinente et efficace à localiser les variants d'unités de mesure. L'analyse des résultats montre que Naives Bayes produit une F-mesure allant de 0.85 à 0.88, l'arbre de décision établi sur l'algorithme C4.5 (J48) produit de meilleurs résultats autour de 0.93 à 0.96. DMNB et SMO produisent les meilleurs résultats (0.95 à 0.99) conformément à ce qui est souligné dans la littérature du domaine. Outre ces résultats analytiques, nous remarquons qu'un plus large contexte, à partir des fenêtres f_{+2} et f_{-2}, n'améliore pas les résultats d'apprentissage. Nous pouvons donc en déduire que la plus petite fenêtre textuelle, c'est-à-dire celle où au moins un terme d'unité référencé dans la RTO apparait, est le contexte le plus favorable à la découverte de variants d'unités. L'application du modèle d'apprentissage sur le corpus emballage permet de réduire l'espace de recherche des variants de 86%.
Le tableau 3.3 restitue les résultats selon les différentes mesures de pondération et la matrice booléenne pour la fenêtre optimale f_0. Notre objectif étant d'évaluer, à partir de ce nouveau tableau, quel algorithme produit le modèle le plus constant sur les différentes mesures. Nous constatons que la F-mesure sur Naive Bayes chute de 0.88 (modèle booléen) à 0.76 (mesure okapi). Elle s'effondre de 20% avec le modèle SMO, en chutant de 0.99 à 0.82. Elle reste plutôt stable avec l'arbre de décision, autour de $0.92 - 0.93$. La F-mesure, ainsi que les valeurs de précision et de rappel restent stables et élevées avec le modèle DMNB, en restituant une valeur constante autour de 0.95.

Les expérimentations sur le corpus des emballages ont montré que la fenêtre textuelle f_0, c'est-à-dire celle où au moins un terme d'unité référencé dans la RTO est identifié, est la fenêtre optimale de découverte de variants d'unités. Les expérimentations ont également permis de conclure que le modèle le plus constant et le plus performant sur les résultats est obtenu avec l'algorithme DMNB. Nous pouvons, pour tout nouveau document, considérer uniquement les phrases du document contenant au moins une unité

	Dec. Tree J48			**Naive Bayes**			**DMNB**			**SMO**		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
f_0	0.99	0.87	0.93	0.83	0.93	0.88	0.95	0.96	**0.95**	0.99	0.99	**0.99**
f_{+2}	0.99	0.92	0.96	0.95	0.77	0.85	0.93	0.96	**0.95**	0.99	0.97	**0.99**
f_{-2}	0.99	0.92	0.95	0.77	0.98	0.86	0.94	0.96	**0.95**	0.99	0.97	**0.98**

Tableau 3.2: Résultats des instances de la classe "Unit" : Précision (P), Rappel (R), F-mesure (F) restitués pour chaque fenêtre textuelle.

	Dec. Tree J48			**Naive Bayes**			**DMNB**			**SMO**		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Boolean	0.99	0.87	0.93	0.83	0.93	0.88	**0.95**	**0.96**	**0.95**	0.99	0.99	0.99
TF	0.99	0.86	0.92	0.69	0.85	0.76	**0.95**	**0.96**	**0.95**	0.84	0.90	0.87
TF.IDF	0.99	0.86	0.92	0.69	0.85	0.76	**0.95**	**0.96**	**0.95**	0.84	0.90	0.87
Okapi	0.99	0.86	0.92	0.69	0.86	0.76	**0.95**	**0.96**	**0.95**	0.77	0.88	0.82

Tableau 3.3: Résultats des instances de la classe "Unit" sur f_0 : Précision (P), Rappel (R), F-mesure (F) restitués pour chaque mesure de pondération et le modèle booléen.

référencée dans la RTO et classer ces phrases, à partir du modèle d'apprentissage produit, comme contenant ou pas un variant d'unité de mesure.
Nous avons appliqué le modèle appris à partir du corpus des emballages sur le corpus de bioraffinerie. Ce modèle appris nous permet de classer les nouvelles phrases extraites du nouvel ensemble de bioraffinerie. Nous avons donc recueilli toutes les phrases contenant au moins une unité référencée de la RTO naRyQ_Bioraf et nous cherchons à prédire si ces phrases comportent ou pas des variants d'unités de mesure. Les résultats de la classification des nouvelles phrases sont restitués dans le tableau 3.4. Les résultats montrent que malgré une légère chute des résultats, le modèle d'apprentissage reste suffisamment générique pour être performant sur un autre domaine.

	Boolean			**TF**			**TF.IDF**			**Okapi**		
	P	R	F	P	R	F	P	R	F	P	R	F
DMNB f_0	0.92	0.86	0.89	0.92	0.86	0.89	0.92	0.86	0.89	0.92	0.86	0.89

Tableau 3.4: Résultats des instances de la classe "Unit" : Précision (P), Rappel (R), F-mesure (F) restitués sur les données de bioraffinerie à partir du modèle appris et validé sur le corpus des emballages.

3.3.2.3 Discussion

Au cours de cette première étape, nous avons donc proposé une nouvelle représentation des données permettant d'évaluer des fenêtres textuelles afin d'apprendre le contexte phrastique optimal à la découverte des variants d'unités de mesure. De plus, guider le processus en s'appuyant sur la RTO s'avère efficace et permet de proposer une méthode affranchie de la coûteuse tâche d'annotation manuelle des données par les experts. En effet, en projetant les termes d'unités référencés dans la RTO, nous avons annoté automatiquement les contextes phrastiques en deux classes, "unit" ou "non-unit", i.e. exemples positifs ou exemples négatifs. Puis, nous avons utilisé les méthodes d'apprentissage supervisé, en comparant le comportement de différents algorithmes sur les données dont les descripteurs choisis, ceux apparaissant plus d'une fois dans les documents, ont été pondérés par plusieurs mesures, *tf, tf.idf et okapi*. Les résultats de la première étape permettent d'optimiser sensiblement le processus de localisation des variants d'unités en réduisant le champs de recherche de 86% pour le corpus des emballages et de près de 90% pour le corpus de bioraffinerie. Une des originalités de l'approche est d'avoir appris un modèle sur un corpus d'un domaine et de l'appliquer sur un nouveau corpus d'un domaine différent

sans réitérer de phase d'apprentissage sur ce nouveau corpus. Les résultats ont montré que le modèle est suffisamment générique pour produire de bons résultats sur le nouveau corpus. Ces résultats permettent d'optimiser l'espace de recherche des variants dans les documents. Cette étape préalable est fondamentale car elle permet d'accroitre la performance d'identification des variants par les mesures de similarité, comme cela est présenté dans la deuxième étape de la méthode dans la section 3.4.3.
Au-delà de ces résultats, nous avons réalisé des expérimentations supplémentaires autour des descripteurs choisis (apparaissant au moins une fois dans les documents étudiés) afin d'observer leur importance dans le processus de décision durant la classification. La question que nous nous posons est de savoir si l'ensemble d'apprentissage repose sur les descripteurs d'unités de mesure contenus dans le sac de mots ou si les algorithmes utilisent d'autres termes, c'est-à-dire le contexte, pour prendre les décisions de classification.

Nous avons réalisé deux types d'expérimentations :

- Nous avons retiré les descripteurs de type terme d'unité du sac de mots et nous avons réitéré une phase d'apprentissage sur le corpus des emballages, à partir du modèle validé,
- Nous avons utilisé une fonction de rang, *desc ranking*, permettant d'ordonnancer l'ensemble des descripteurs en fonction de leur poids dans les différents vecteurs.

Les résultats de l'apprentissage sans les descripteurs d'unités de mesure sont restitués dans le tableau 3.5. Les nouveaux résultats de précision, de rappel et F-mesure sont notés en gras afin d'être comparés aux précédents résultats, en non gras. L'annotation f_0^{-um} correspond à la fenêtre considérée pour laquelle le sac de mots ne comporte pas de termes d'unités de mesure $(-um)$. Notre première observation est que le modèle produit par

	Boolean			**TF**			**TF.IDF**			**Okapi**		
	P	R	F	P	R	F	P	R	F	P	R	F
f_0	0.92	0.86	0.89	0.92	0.86	0.89	0.92	0.86	0.89	0.92	0.86	0.89
f_0^{-um}	**0.76**	**0.81**	**0.79**	**0.76**	**0.81**	**0.79**	**0.76**	**0.81**	**0.79**	**0.76**	**0.81**	**0.79**

Tableau 3.5: Résultats des instances de la classe "Unit" : Précision (P), Rappel (R), F-mesure (F) restitués en retirant du sac de mots les termes d'unités référencés dans la RTO.

DMNB restitue toujours des résultats équilibrés avec les mesures appliquées. Les résultats

d'apprentissage montrent une baisse de la précision et du rappel avec des taux respectifs de 0.76 et 0.81. Ces résultats permettent de déduire que le descripteur unité de mesure reste un descripteur majeur. Il contient une information sémantique et lexicale fondamentale que nous exploitons dans la section suivante pour l'identification des nouvelles unités. Néanmoins, cette baisse de la précision et du rappel reste acceptable suggérant que le contexte joue également un rôle pour prédire la localisation des variants d'unités.

Nous avons défini une fonction de rang, *desc ranking*, afin d'ordonnancer les descripteurs selon leur valeur discriminante octroyée par l'algorithme, à partir des résultats obtenus par la mesure de pondération *okapi* et dans le contexte d'un sac de mots vidé des termes d'unités de mesure référencés dans la RTO. Le tableau 3.6 montre que les descripteurs les plus discriminants pour l'apprentissage sont des descripteurs en rapport avec les données et les conditions expérimentales du domaine d'intérêt, en particulier des verbes expérimentaux de type descriptif (i.e. *modify, store*) ou analytique (i.e. *range*). Parmi les 10 premiers descripteurs discriminants, nous retrouvons deux termes dénotant les concepts quantité de la relation n-aire (i.e. *rh, thickness*). Cet ordonnancement montre que les descripteurs liés au domaine étudié, et présents dans la RTO, correspondent à des attributs pertinents et discriminants. Les descripteurs les moins discriminants sont effectivement des termes plutôt génériques même s'ils font néanmoins partie du vocabulaire expérimental.

Dans cette section, nous avons montré que la première étape de notre méthode, s'appuyant sur l'apprentissage supervisé, permet de localiser automatiquement les variants d'unité dans un contexte phrastique optimal de recherche.
La méthode repose sur une nouvelle représentation des données, sous forme de fenêtres textuelles d'étude, que nous obtenons en étant guidés par la RTO.

La section suivante décrit la deuxième étape de la méthode : une fois l'espace de recherche réduit aux variants d'unités, elle propose une nouvelle mesure de similarité permettant d'identifier automatiquement les variants découverts par rapport à un terme d'unité déjà référencé dans la RTO.

3.4 Identification des unités de mesure

La première étape de la méthode proposée permet, à partir du modèle d'apprentissage, de classer les contextes favorables à la découverte de nouvelles unités de mesure et ainsi en réduire l'espace de recherche.

Rang	terme	valeur discriminante
1	modified	1.02
2	rh	0.95
3	stored	0.77
4	thickness	0.61
5	room	0.56
6	glycerol	0.47
7	range	0.43
8	packages	0.41
9	solutions	0.40
10	samples	0.40
...	...	...
37	humidity	0.299
46	experimental	0.289
47	op	0.286
...	...	...
594	chlorine	0.0999
595	pouree	0.0994
596	hour	0.098
...	...	...
2622	analysis	-0.40
2623	methods	-0.49
2624	food	-0.53
2625	dipped	-1.4

Tableau 3.6: Ordonnancement des descripteurs avec *desc ranking*

La deuxième étape concerne l'identification des variants d'unités. En effet, lorsque la phrase est classée comme susceptible de contenir une nouvelle unité, l'approche consiste à comparer le variant extrait aux termes d'unités de mesure référencées dans la RTO et identifier de quelle unité il est le variant. L'identification des termes de variants d'unités concerne la problématique de comparaison des chaînes de caractères et de calcul de la similarité existante entre ces deux chaînes.
Nous présentons, dans la section 3.4.1, les mesures principales à l'état de l'art, puis, nous détaillons notre proposition de mesure de similarité adaptée à la problématique de comparaison des unités de mesure.

3.4.1 Les mesures de similarité

Nous nous sommes intéressés aux différentes mesures de similarité à l'état de l'art. Il existe trois types de méthodes pour comparer des chaînes de caractères :

Méthodes basées sur les caractères. La mesure de similarité de Levenshtein (Levenshtein, 1966) calcule le coût minimal pour transformer une première chaîne de caractères en une seconde chaîne de caractères en considérant les opérations de remplacement de caractères entre les deux chaînes, l'ajout d'un caractère ou la suppression d'un caractère. Le coût est ensuite normalisé pour obtenir une valeur de la distance entre les deux chaînes entre 0 et 1.
Cette mesure est étendue par Damerau (Damerau, 1964) qui inclut dans celle de Levenshtein la notion de transposition de caractères d'une chaîne à une autre, i.e. *litre* et *liter* où il y a transposition entre les caractères "e" et "r".
En considérant deux chaînes de caractères u_1 et u_2, on obtient la formule normalisée suivante pour le calcul de la distance de Damerau-Levenshtein (DL) (Maedche and Staab, 2002) :

$$SM_{DL}(u_1, u_2) = max[0; \frac{\min(|u_1|, |u_2|) - DL(u_1, u_2)}{min(|u_1|, |u_2|)}] \in [0; 1] \tag{3.1}$$

Un résultat proche de 1 conclut que les deux chaînes u_1 et u_2 sont similaires.

La mesure de Jaro (Jaro, 1989) est définie par la formule suivante :

$$Jaro(u_1, u_2) = \frac{1}{3}(\frac{m}{|u_1|} + \frac{m}{|u_2|} + \frac{m-t}{m}) \tag{3.2}$$

où :

- $|u_i|$ est la longueur de la chaîne de caractères u_i ;

- m est le nombre de caractères correspondants, c'est-à-dire que leur éloignement (la différence entre leur position dans les chaînes u_1 et u_2) ne dépasse pas : $[\frac{max(|u_1|,|u_2|)}{2}] - 1$;

- t est le nombre de transpositions, obtenu en comparant le i-ème caractère dans u_1 et le i-ème caractère dans u_2. Le nombre de fois où ce caractère est différent divisé par 2 donne le nombre de transpositions.

La méthode de Winkler (Winkler, 1999) permet d'étendre la mesure de Jaro en introduisant la notion de préfixe commun P qui favorise les chaînes commençant par un même préfixe de longueur l avec $l \leq 4$. La mesure est définie par la formule :

$$JaroWinkler(u_1, u_2) = Jaro(u_1, u_2) + (l_p(1 - Jaro(u_1, u_2))) \tag{3.3}$$

où :

- l est la longueur du préfixe commun avec un maximum de 4 caractères ;

- p est un coefficient qui favorise les chaînes ayant un préfixe commun, fixé à 0.1 par Winkler.

Les méthodes fondées sur les caractères permettent de prendre en compte la position de caractères identiques dans deux chaînes à comparer, mais ne peuvent pas distinguer les mots lorsqu'une chaîne en comporte plusieurs.

Méthodes fondées sur les tokens. Lorsque la chaîne de caractères est découpée en unités lexicales élémentaires, correspondant souvent à un symbole bien précis (article, nom, verbe, chiffre...), nous appelons l'unité élémentaire découpée un lexème ou une entité lexicale ou encore un token. Un token est donc, en généralisant, utilisé pour définir un mot.
Les méthodes basées sur les tokens permettent de prendre en compte les tokens communs dans des chaînes en les considérant comme un ensemble de mots plutôt que comme un ensemble de caractères.

L'indice de Jaccard consiste à diviser le nombre d'objets communs par le nombre d'objets distincts dans les deux ensembles, autrement dit le cardinal de l'intersection divisé par le cardinal de l'union, dont la formule est définie par :

$$Jaccard(u_1, u_2) = \frac{|u_1 \cap u_2|}{|u_1 \cup u_2|} \tag{3.4}$$

L'indice de Jaccard est utilisé pour comparer des ensembles de mots ou tokens équivalents. Il est particulièrement efficace pour comparer des chaînes de caractères comportant des éléments indépendants et équiprobables.

La mesure statistique TF.IDF, que nous avons présentée dans la section 3.3 concernant les mesures de pondérations utilisées en Recherche d'Information, est également utilisée pour mesurer la similarité entre deux chaînes de caractères. La mesure TF.IDF va attribuer une valeur de similarité entre deux chaînes en considérant les tokens communs dans les chaînes à comparer et la fréquence de chaque token dans les chaînes. Elle va donc octroyer un poids plus important aux termes plus discriminants et comparer plus finement les deux chaînes lorsque cela est nécessaire. Cette mesure est intéressante pour comparer des chaînes de caractères contenant des mots vides comme "et", "le", "la".

Les méthodes fondées sur les tokens permettent de prendre en compte les mots communs entre les chaînes de caractères comparées sans tenir compte de l'ordre des mots dans les chaînes. En revanche, elles ne sont pas en mesure de traiter les mots similaires, par exemple les mots "mole" et "mol" seront considérés comme des mots différents.

Méthodes hybrides. Les méthodes hybrides permettent de comparer les chaînes de manière plus flexible par rapport aux deux types de mesures précédentes. En effet, avec les mesures fondées sur les caractères, la position des caractères, donc des mots, est prise en compte alors que les méthodes fondées sur les tokens permettent de comparer des chaînes en s'affranchissant de l'ordre des mots. Ces dernières ne permettent néanmoins pas de comparer des mots similaires, par exemple, le terme *mole* sera considéré comme différent de *mol* même si seulement une lettre les différencie. Les méthodes hybrides permettent de calculer la similarité en tenant compte à la fois des tokens et des caractères.
La méthode de (Monge and Elkan, 1996) fractionne les chaînes en tokens et calcule la similarité de chaque token d'une chaîne par rapport à la deuxième chaîne en utilisant une mesure basée sur les caractères. La mesure calcule ensuite un score global de similarité en considérant la moyenne des scores obtenus.
La méthode de (Cohen et al., 2003) permet d'améliorer la mesure par TF.IDF en pro-

posant des mesures comme *SoftTFIDF* ou *JaroWinkler-TFIDF*. La mesure ne considère pas uniquement les tokens de l'intersection entre les deux ensembles mais va également prendre en compte les tokens similaires.

Dans la section suivante, nous exposons les spécificités à prendre en compte pour comparer des unités de mesure et nous nous appuyons sur les méthodes à l'état de l'art afin de proposer une nouvelle mesure de similarité.

3.4.2 Comparer des unités de mesure

Contrairement aux variations terminologiques considérées pour évaluer la similarité entre deux chaînes de caractères, les unités de mesure possèdent leurs propres règles d'écriture établies librement par l'auteur du document. Par exemple, l'unité de mesure *amol/(m.s.Pa)* définie dans la RTO naRyQ_emb peut être écrite à l'aide de différents variants dans les documents scientifiques selon les cas :

- d'insertions de caractères comme dans *amol/m.sec.Pa* ou *amol.m-1.s-1.Pa-1*,
- de suppressions de caractères comme dans *mol/m.s.Pa*,
- d'inversions de certains blocs dans l'unité comme dans *amol.s-1.m-1.Pa-1*,
- d'écriture non plus ponctuée mais comme un ensemble composé de blocs indépendants comme dans *amol m-1 s-1 Pa-1*.

D'autre part, nous avons constaté, à partir de la hiérarchie des unités de mesure établie dans la RTO, que les variants posant des problèmes d'identification sont ceux rattachés au concept *Unit_Division_Or_Multiplication* référençant des unités complexes. Les unités dénotant ce concept sont formées en fonction de blocs d'unités dénotant les concepts *Singular_Unit, Unit_Multiple_or_Submultiple ou Unit_Exponentiation* de la hiérarchie décrite dans la figure 3.5. Par exemple, l'unité *amol/m.s.Pa* est formée par les blocs d'unités *amol* dénotant une instance du concept *Unit_Multiple_Or_Submultiple* et les unités *m, s, Pa* dénotant les instances du concept *Singular_Unit*.

Nos travaux proposent d'extraire de telles variations terminologiques dans les documents afin d'enrichir la RTO de ces variants d'unités de mesure. Une telle extraction ne peut pas reposer sur des méthodes utilisant des expressions régulières sur les unités de mesure car elles ne couvrent pas suffisamment les expressions possibles sachant que l'écriture est librement choisie par les auteurs. Nous proposons d'utiliser et adapter les mesures de proximité aux spécificités des unités de mesure afin de répondre à la problématique d'identification des variants d'unités.

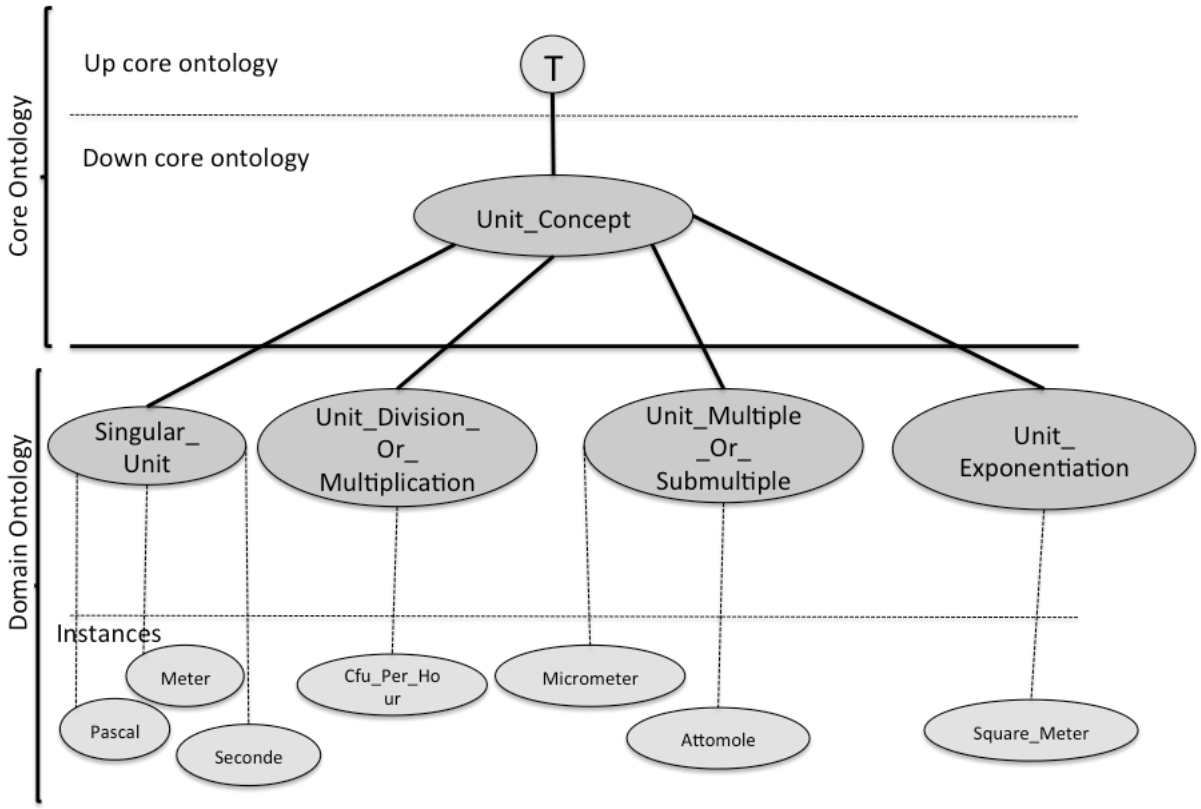


FIGURE 3.5: Hiérarchie de concepts des unités de mesure de la RTO naRyQ.

Dans notre approche, il est fondamental de pouvoir prendre en considération les particularités d'écriture des unités de mesure, en notant que chaque bloc est indépendant dans l'écriture de l'unité. De ce fait, l'ordre des blocs n'est pas important à prendre en compte, en revanche, la comparaison des blocs entre eux nous semble plus pertinente et plus adaptée dans le calcul de la similarité, comme nous le présentons en détails dans la section suivante.
En somme, à partir des méthodes de l'état de l'art, nous n'avons pas pu nous appuyer sur les mesures proposant d'utiliser la mesure de TF.IDF, car ce calcul est intéressant et pertinent à utiliser dans le contexte de comparaison entre deux chaînes pour lesquelles le nombre de termes communs, assimilés à des mots vides, tels que les conjonctions de coordination ou les pronoms par exemple, sont susceptibles de fausser le résultat de la similarité. Le TF.IDF, dans ce contexte, permet de pondérer leur poids, en considérant que, plus le terme apparait fréquemment, moins il est intéressant à prendre en compte. Dans notre cas, les termes sont tous spécifiques, il n'est donc pas pertinent d'utiliser ce type de pondération.
Nous n'avons pas pu non plus nous appuyer sur la mesure de Jaro-Winkler car celle-ci prend en compte un préfixe commun pour comparer les chaînes de caractères. Or, comme nous l'avons présenté, les unités de mesure peuvent varier considérablement dans l'ordre d'écriture des blocs sans que cela n'impacte la similarité réelle entre deux chaînes de caractères d'unités de mesure.
Dans notre contexte de variants d'unités de mesure, deux mesures nous sont apparus pertinentes : la mesure de Damereau-Levenshtein (formule 3.1) car elle prend en compte toutes les variations rencontrées dans l'écriture des unités de mesure (insertion, suppression, transposition...) et la mesure de Jaccard (formule 3.4) car elle permet de s'affranchir de l'ordre des blocs dans l'unité complexe en considérant ceux-ci comme indépendants entre eux, comme illustré dans l'exemple 10. Nous avons testé, dans un premier temps, la mesure de Damereau-Levenshtein comme définie à l'état de l'art. Un résultat est restitué dans l'exemple 9. Nous remarquons, à partir de cet exemple, que malgré la forte similarité des deux unités, le score restitué par la mesure de Damereau-Levenshtein n'est que de 0.69. Ce résultat illustre le fait que la mesure de Damereau-Levenshtein, basée sur la comparaison de chaque caractère dans les chaînes, nécessite une adaptation afin de prendre en compte les variations typographiques des unités de mesure et restituer un score fidèle à la réalité.

Exemple 9
Considérons l'unité référencée dans la RTO naRyQ_ Emb amol/(m.s.Pa) et amol/m.sec.Pa extraite d'un document.

Nous avons deux caractères supprimés "(" et ")" et nous avons deux caractères insérés "e" et "c".
La distance D_c (Damereau-Levenshtein fondée sur les caractères) est donc égale à 4.
La mesure de similarité normalisée est donc de :

$$SM_{Dc}(amol/(m.s.Pa), amol/m.sec.Pa) = max[0; \frac{13-4}{13}] = 0.69$$

La mesure de Jaccard permet, dans l'exemple 10, de comparer les deux chaînes de caractères sans tenir compte de l'ordre des blocs.

Exemple 10
Considérons les unités amol m^{-1} s^{-1} Pa^{-1} et amol s^{-1} m^{-1} Pa^{-1}.
Nous avons le bloc s^{-1} qui se trouve à une position différente dans les deux unités.
La similarité prend en compte uniquement les blocs communs sans tenir compte de leur ordre dans les chaînes comparées :

J(amol m^{-1} s^{-1} Pa^{-1}, amol s^{-1} m^{-1} Pa^{-1}) = 1

3.4.3 Nouvelle mesure d'identification adaptée aux unités de mesure

Dans le contexte des unités de mesure, nous savons que les unités complexes possèdent une syntaxe particulière, composée de blocs indépendants, correspondant à des unités dénotant d'autres concepts de la RTO et comprenant des caractères spécifiques, librement choisis pour exprimer l'unité complexe finale. Nous proposons de calculer la similarité entre une unité variante candidate et une unité référencée dans la RTO en deux temps et en s'appuyant sur les caractères spécifiques (/, (,), ., ×, ^...) utilisés :

1. Dans un premier temps, les candidats sont pré-sélectionnés selon la mesure de Jaccard :

   - Soit un couple composé du variant candidat u_i et d'une unité référencée dans la RTO u_j. $J(u_i, u_j)$ (formule 3.4) calcule le score de similarité entre l'ensemble u_i et l'ensemble u_j par rapport aux blocs communs sans tenir compte de leur ordre.

- Le couple (u_i, u_j) est alors pré-sélectionné comme étant pertinent à être comparé si $J(u_i, u_j) > K'$, K' étant le seuil minimal défini préalablement par l'utilisateur.

2. Dans un second temps, après cette phase de pré-sélection, les candidats sont sélectionnés selon une nouvelle mesure, appelée SM_{Db} , qui est une mesure étendue de Damereau-Levenshtein (formule 3.5). La mesure adaptée à notre contexte ne considère plus la comparaison des caractères mais des blocs de caractères, correspondant à des unités simples. Le variant et l'unité de référence composant le couple présélectionné lors de la première phase sont dans cette seconde phase, comparés bloc à bloc pour déterminer leur similarité finale.

$$SM_{Db}(u_i, u_j) = max[0; \frac{\min(|u_i|, |u_j|) - D_b(u_i, u_j)}{min(|u_i|, |u_j|)}]; SM_{Db}(u_i, u_j) \in [0; 1] \quad (3.5)$$

où :

- (u_i, u_j) représente le couple présélectionné à partir de la mesure de Jaccard,
- Chaque bloc de u_i est comparé aux blocs de u_j pour calculer la nouvelle distance D_b,
- u_i est sélectionné comme un variant de l'unité u_j si $SM_{Db} > K$, avec K un seuil de similarité défini préalablement,

Cette nouvelle mesure combinée est plus efficace pour comparer les variants et leurs unités de référence dans la RTO, comme illustré dans les exemples 11 et 12 extraits des résultats que nous avons obtenus en expérimentations.

Exemple 11
Prenons l'exemple du couple composé d'un variant localisé et extrait à partir d'un document kg m $Pa^{\hat{}-1}s^{\hat{}-1}m^{\hat{}-2}$ *et son référent dans la RTO* $lb.m.m^{-2}.s^{-1}.Pa^{-1}$.
L'utilisateur a fixé un seuil de sélection pour K et K' supérieur à 0.5.
Dans un premier temps, la mesure de Jaccard calcule la pertinence à comparer ces deux unités :

J(kg m $Pa^{\hat{}-1}s^{\hat{}-1}m^{\hat{}-2}$, $lb.m.m^{-2}.s^{-1}.Pa^{-1}$*)* $= \frac{4}{6} = 0.7$

Le couple est pré-sélectionné et, dans un second temps, SM_{Db} *calcule la similarité du couple en comparant chaque bloc dans les unités :*

SM_{Db}(*kg m Pa*$^{\hat{}-1}$*s*$^{\hat{}-1}$*m*$^{\hat{}-2}$, *lb.m.m*$^{-2}$*.s*$^{-1}$*.Pa*$^{-1}$) = $max[0; \frac{5-1}{5}] = 0.8$

Le variant est sélectionné et peut être proposé à l'expert pour ajout dans la terminologie de la RTO dénotant l'instance appropriée du concept Division_ Or_ Multiplication.

Exemple 12

Prenons un nouvel exemple avec le couple composé d'un variant localisé et extrait à partir de nos documents cm$^{\hat{}3}$ *mm day*$^{\hat{}-1}$ *m*$^{\hat{}-2}$ *et son référent dans la RTO cm*^{3}*.µm.m*$^{-2}$*.-day*$^{-1}$*.atm*$^{-1}$*.*
Le seuil de pré-sélection et sélection (K et K') est toujours supérieur à 0.5.
La mesure de Jaccard restitue la pertinence à comparer ces deux unités :

J(cm$^{\hat{}3}$ *mm day*$^{\hat{}-1}$ *m*$^{\hat{}-2}$*, cm*^{3}*.µm.m*$^{-2}$*.day*$^{-1}$*.atm*$^{-1}$*)* = $\frac{3}{6}$ = *0.5*

Nous sommes à la limite du seuil de pré-sélection défini pour comparer le couple. En effet, les deux unités ne comportent pas le même nombre de blocs. Il n'existe pas de référent dans la RTO, il s'agit d'une nouvelle unité découverte, à intégrer dans la RTO. Elle dénote une nouvelle instance du concept Division_ Or_ Multiplication.
Le processus d'identification et d'enrichissement étant un processus itératif, de nouveaux couples de variants et unités de la RTO sont à nouveau évalués. Les nouvelles unités intégrées permettent de nouvelles validations, e.g. cm$^{\hat{}3}$ *mm day*$^{\hat{}-1}$ *m*$^{\hat{}-2}$ *permet au variant cm*^{3} *mm m*$^{-2}$ *day*$^{-1}$ *d'être évalué.*

J(cm^3 *mm m*$^{-2}$ *day*$^{-1}$*, cm*$^{\hat{}3}$ *mm day*$^{\hat{}-1}$ *m*$^{\hat{}-2}$*)* = $\frac{4}{4}$ = *1*

et la restitution avec la nouvelle mesure SM_{Db} *sera donc de :*

SM_{Db}*(cm*^{3} *mm m*$^{-2}$ *day*$^{-1}$*, cm*$^{\hat{}3}$ *mm day*$^{\hat{}-1}$ *m*$^{\hat{}-2}$*)*= $max[0; \frac{4-0}{4}] = 1$

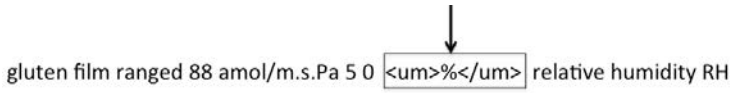
Le variant cm^3 *mm m*$^{-2}$ *day*$^{-1}$ *est validé et est intégré dans la terminologie de la RTO dénotant une instance du concept Division_ Or_ Multiplication.*

Un tel processus fondé sur ces deux phases consécutives de présélection et de sélection finale permet la découverte de nouveaux variants à intégrer dans la RTO. Nous présentons dans la section suivante les résultats de nos expérimentations sur les deux corpus étudiés.

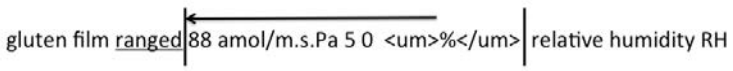
3.4.4 Expérimentations

3.4.4.1 Protocole expérimental

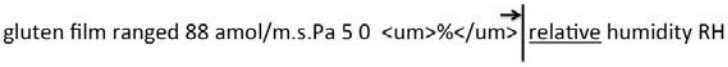
À la fin de la première étape, présenté en section 3.3, nous obtenons les phrases classées comme contenant potentiellement un variant d'unités à identifier. Ces phrases sont d'abord nettoyées afin d'isoler le variant d'unité à identifier. À partir de l'unité de mesure référencée dans la RTO, nous balayons la phrase de part et d'autre de l'unité jusqu'à trouver un terme du dictionnaire. Le dictionnaire est celui utilisé par l'étiqueteur de *Brill* (Brill, 1992) contenant près de 100 000 mots. Une fois le mot du dictionnaire identifié, nous récupérons ainsi uniquement la fraction de la phrase contenant le variant à identifier, comme cela est illustré dans la figure 3.6 sur un variant considéré comme un token et sur la figure 3.7 sur un variant considéré comme plusieurs tokens. L'unité référencée et les valeurs numériques sont également éliminées.



(a) Unité de mesure référencée dans la RTO



(b) Un premier terme reconnu du dictionnaire



(c) Un second terme reconnu du dictionnaire

Unit fragment: 88 amol/m.s.Pa 5 0 %

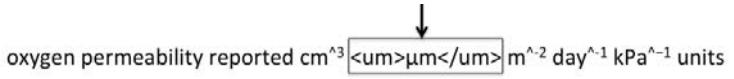
(d) Le variant à identifier

FIGURE 3.6: Isoler le variant considéré comme un token

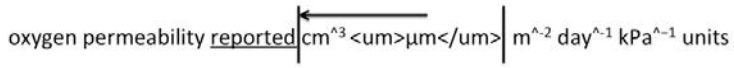
Dans cette deuxième étape, les termes d'unités de mesure référencées dans la RTO des domaines servent de référents dans le processus d'identification. Les expérimentations sont menées sur les mêmes corpus que ceux décrits en section 3.3.2.1.

3.4.4.2 Résultats et discussion

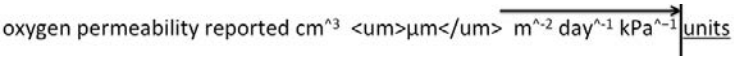
Nous nous appuyons sur les résultats obtenus précédemment afin d'identifier les variants d'unités. Dans ce contexte, nos mesures doivent sélectionner les variants les plus



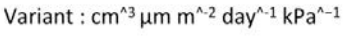
(a) Unité de mesure référencée dans la RTO



(b) Un premier terme reconnu du dictionnaire



(c) Un second terme reconnu du dictionnaire



(d) Fraction contenant le variant à identifier

FIGURE 3.7: Isoler le variant considéré comme plusieurs tokens

pertinents à présenter aux experts, ce qui revient à minimiser le bruit. Ainsi, nous privilégions la mesure de précision pour évaluer nos propositions. La précision des résultats est restituée après validation manuelle. Le tableau 3.7 restitue les résultats d'évaluation obtenus avec les mesures de Jaccard et SM_{Db} sur un ensemble de 11 articles restituant 25 variants d'unités à identifier. Les couples d'unités formés du variant et son référent dans la RTO sont comparés par seuil de similarité (à partir de 0.4) :

1. La précision reste plus élevée avec le processus complet (c'est-à-dire présélection et sélection avec SM_{Db}), confirmant ainsi la pertinence et l'intérêt de la mesure SM_{Db},
2. Au-delà de 0.6, les validations montrent que le processus propose une extraction avec des taux de précision supérieurs à 0.7, après application des deux étapes pour l'essentiel des variants identifiés,
3. En dessous de 0.5, les résultats d'extraction restent très bons avec des F-mesures autour de 0.7 et 0.8, ce qui montre que les extractions plus nombreuses à ces seuils ne provoquent pas de bruit gênant pour les validations manuelles. En choisissant de ne considérer que les seuils au-dessus de 0.5, nous créons forcément du silence, comme le montre les taux de rappel entre 0.4 et 0.6 mais un silence "contrôlé". En effet, le processus d'extraction et d'identification des variants étant un processus itératif, les nouvelles unités intégrées dans la RTO favorisent la découverte d'autres variants qui s'expriment dans cette plage de silence. Les prochaines itérations permettent ainsi d'améliorer le rappel, en particulier pour les seuils supérieurs à 0.7.

Dès l'enrichissement de nouvelles unités dans la RTO, la similarité des nouveaux couples est restituée à de meilleurs seuils, ayant de nouveaux référents dans la RTO, comme cela est illustré dans l'exemple 12

Le processus est ensuite appliqué sur le corpus complet. Chaque variant peut former un couple avec plusieurs unités de référence dans la RTO. En effet, prenons l'exemple du variant *amol / m s Pa*, sa comparaison avec les unités de référence *amol/m/s/Pa, amol/(m.s.Pa), amol/(m s Pa)...* est considérée comme pertinente. Pour tous les couples pertinents validés, nous n'intégrons qu'une seule fois le variant *amol / m s Pa*. Considérant cette remarque, sur les 267 couples identifiés comme pertinents par SM_{Db} dans le processus d'identification et validés, nous obtenons 121 variants d'unités uniques à intégrer dans la RTO naRyQ_Emb.

Les tableaux 3.8 et 3.9 restituent quelques exemples d'unités extraites sur nos deux corpus, à partir de la nouvelle mesure combinée et compare ces résultats avec ceux de la mesure d'origine SM_{Dc}. La nouvelle mesure SM_{Db} restitue des résultats de similarité plus pertinents et s'adapte plus efficacement à la syntaxe des unités de mesure. En effet, dans un premier temps, la mesure de Jaccard permet de présélectionner efficacement les couples à comparer, y compris les couples où les unités sont écrites comme un ensemble de tokens distincts. Cette dernière particularité ne peut être prise en charge par la mesure classique SM_{Dc}, qui calcule la similarité uniquement sur les unités considérées comme un seul mot. Nous avons représenté cette incapacité avec le symbole "/" dans le tableau. Au cours des tests, nous avons remplacé les blancs séparateurs par un *underscore* qui est un caractère n'intervenant pas dans la syntaxe des unités de mesure. Ce traitement permet à la mesure SM_{Dc} de calculer une similarité, mais le résultat restitué reste non pertinent par rapport à SM_{Db}. Par exemple, reprenons une des identifications effectuées dans le tableau 3.8, *(amol/(Pa m s), amol/m/s/Pa)*. Remplaçons les blancs séparateurs par des _ et comparons *(amol/(Pa_m_s), amol/m/s/Pa)* par SM_{Dc}. La mesure SM_{Dc} comptabilise 4 caractères supprimés, 2 "_", "(" et ")", 2 caractères insérés, 2"/" et les 4 transpositions puisque les blocs sont désordonnés. Elle comptabilise le nombre de caractères total dans l'unité la plus grande (13) et la similarité $SM_{Dc} = \frac{13-10}{13} = 0.23$, par rapport à $SM_{Db} = 1$.

En analysant de plus près les résultats obtenus sur le corpus de bioraffinerie, pour lequel la RTO du domaine est en cours d'enrichissement, nous voyons qu'il est particulièrement intéressant d'analyser également les faibles seuils de similarité. En effet, dans ce contexte, nous identifions de nouvelles unités à intégrer dans la RTO. N'ayant pas encore de référents spécifiques dans la RTO naRyQ_bioraf, elles sont comparées à des référents potentiels mais, logiquement, avec de faibles résultats de similarité. Sachant que nous

avions dans la RTO un faible nombre d'unités de référence, nous avons abaissé les seuils K et K' à 0.2 afin de découvrir de nouvelles unités à intégrer dans la RTO. Le processus d'enrichissement de la RTO étant un processus itératif, une nouvelle phase d'extraction et d'identification permettrait alors de comparer d'autres variants avec ces nouvelles unités intégrées dans la RTO, qui deviennent des référents.

Seuil de similarité	**Pré-sélection par Jaccard**			**Sélection par** SM_{Db}		
	Précision	*Rappel*	*F-mesure*	*Précision*	*Rappel*	*F-mesure*
[0.9-1]	0.7	0.4	0.5	0.8	0.4	0.5
[0.8-1]	0.8	0.5	0.6	0.8	0.6	0.7
[0.7-1]	0.8	0.7	0.7	0.8	0.6	0.7
[0.6-1]	0.7	0.7	0.7	0.7	0.8	0.7
[0.5-1]	0.7	0.8	0.7	0.7	0.8	0.7
[0.4-1]	0.5	0.8	0.6	0.6	1	0.8

Tableau 3.7: Résultats obtenus avec la nouvelle mesure combinée

Variants d'unités	**Référent**	SM_{Dc}	SM_{Db}
10e10(cm3.m-1.sec-1.Pa-1)	10e10.cm3.m-1.sec-1.Pa-1	0.87	1
10e-14(cm3/m.s.Pa)	10e-14.cm3/(m.s.Pa)	0.89	1
10e18(mol.m/Pa.sec.m2)	10e18.mol.m/(Pa.sec.m2)	0.87	1
amol.m-1.s-1.Pa-1	amol.s-1.m-1.Pa-1	0.88	1
amol/m.s.Pa	amol/(m.s.Pa)	0.84	1
amol/m.sec.Pa	amol/(m.s.Pa)	0.69	0.75
cm3.um/m2.d.kPa	cm3.μm/(m2.d.kPa)	0.77	0.8
$cm\hat{}3.(Pa.s.m)\hat{}-1$	cm3/(m.s.Pa)	0.44	0.8
amol/(Pa m s)	amol/m/s/Pa	/	1
g/($s \times m \times Pa$)	mol/m/s/Pa	0	0.75
g.(Pa s m)$\hat{}-1$	mol/m/s/Pa	/	0.6
kg m Pa-1 s-1 m-2	m3.m.m-2.s-1.Pa-1	/	0.8
fl m/s $m\hat{}2$ Pa	m3.m/(m2.s.Pa)	/	0.6

Tableau 3.8: Un extrait des variants identifiés selon la mesure classique SM_{Dc} et la nouvelle mesure SM_{Db} sur des couples sélectionnés sur le corpus des Emballages

Variants d'unités	**Référent**	SM_{Dc}	SM_{Db}
kW h $kg\hat{}-1$	kW.h.kg^{-1}	/	1
kg $h\hat{}-1$	kg m	/	0.5
mol $L\hat{}-1$	mol/m/s/Pa	/	0.25
g glucose/100g	g/100g	/	0.7
kW MJ	kW.h	/	0.5
g $mL\hat{}-1$	g/kg	/	0.5
kcal $kg\hat{}-1$ °$C\hat{}-1$	kg^{-1}	/	0.33

Tableau 3.9: Un extrait des variants identifiés selon la mesure classique SM_{Dc} et la nouvelle mesure SM_{Db} sur des couples sélectionnés sur le corpus de Bioraffinerie

3.5 Conclusion

Cet enrichissement est une étape fondamentale dans les travaux menés dans la thèse car les variants d'unités de mesure représentent une des problématiques d'extraction des données quantitatives dans les textes. Dans cette section, nous avons proposé une nouvelle mesure de similarité, SM_{Db}, adaptée aux spécificités des unités de mesure, qui s'appuie sur la mesure de Damereau-Levenshtein, appropriée à notre contexte car elle prend en charge toutes les variations constatées pour les unités de mesure. De plus, associée à l'indice de Jaccard, la nouvelle mesure permet de rapprocher les couples variant - unité référencée de la RTO de manière plus pertinente en octroyant un premier score global de similarité qui ne tient pas compte de l'ordre des blocs dans la construction de l'unité complexe. Dans un second temps, la nouvelle mesure SM_{Db} affine ce rapprochement en comparant chaque bloc du variant et de l'unité référencée sélectionnée.
La méthode proposée, guidée par la connaissance de la RTO, permet à partir d'un processus complet de localisation et d'identification des variants d'unités de mesure, d'enrichir la RTO de nouveaux termes d'unités. De plus, nous avons sélectionné l'unité de mesure comme descripteur pertinent à la représentation des données de type expérimental. De ce fait, leur identification est un élément fondamental de nos propositions.

Dans le chapitre suivant, nous présentons notre contribution à la problématique d'extraction des instances de relations n-aires pour lesquelles les unités de mesure tiennent un rôle majeur.

Chapitre 4

Localisation et extraction des arguments de relations n-aires

Sommaire

4.1 Introduction

Dans le chapitre précédent, nous avons présenté notre contribution pour lever le verrou concernant la localisation et l'identification des variants d'unités de mesure dans les documents. En effet, l'unité de mesure est un descripteur ayant une sémantique forte dans la définition des données quantitatives de type expérimental, sur laquelle nous nous appuyons pour proposer notre contribution au travail d'extraction des instances d'arguments de relations n-aires. La contribution concernant les variants d'unités de mesure a permis d'enrichir une RTO de domaine, étape fondamentale pour la suite de nos travaux. Dans ce chapitre, nous nous intéressons aux problématiques d'extraction des instances de relations n-aires dans les documents. En effet, nous avons vu que la tâche d'extraction des relations n-aires est complexe car les arguments engagés dans la relation sont dispersés dans le document ou dans des tableaux, et que l'expression de ces arguments est souvent effectuée dans un discours implicite. De plus, nous avons vu que dans l'ensemble du document, de nombreux résultats expérimentaux sont restitués, et que parmi ces résultats, seuls ceux représentés dans la modélisation de la RTO sont pertinents à être extraits. Ces nouvelles instances serviront à peupler la RTO à des fins de capitalisation de connaissance dans des outils d'aide à la décision. Par exemple, dans le projet EcoBioCap dans le domaine des emballages alimentaires, l'un des objectifs fixés est de permettre aux experts de répondre à des questions telles que :
" Quel emballage choisir afin de conditionner l'aliment A, conservé pendant X jours, à la température de $T\,°C$ et à une perméabilité optimale P ?"
Les instances extraites des documents et intégrées dans les systèmes sont interrogées et analysées par les experts afin de les aider dans leur prise de décision.
Dans ce chapitre, nous proposons de contribuer à cette tâche d'extraction en adoptant une approche hybride fondée sur des méthodes de fouille de données et d'analyse de relations syntaxiques extraites des documents en étant guidés par la RTO. En effet, nous

souhaitons tirer profit du fort potentiel des méthodes de fouilles de données à découvrir des relations implicites et des régularités au sein des données, et de l'analyse syntaxique, en étudiant les relations syntaxiques entre mots de la phrase.
Nous avons décrit, en introduction du mémoire, les méthodes proposées à l'état de l'art concernant l'extraction des relations binaires et les travaux apparentés à notre problématique d'extraction des relations n-aires. Dans ce chapitre, nous présentons dans la section 4.2 notre contribution, fondée sur les approches de fouille de données afin de faire émerger des relations implicites entre les arguments de la relation n-aire. Dans la section 4.3, nous présentons notre approche hybride fondée sur l'analyse des relations syntaxiques afin d'étendre les motifs séquentiels, émergeant de l'étape de fouille de données tout en s'appuyant sur la connaissance du domaine, la RTO.

4.2 Extraction d'arguments corrélés

4.2.1 Introduction

Dans cette section, nous présentons notre travail fondé sur la recherche de régularités et de relations implicites entre les arguments de la relation n-aire. Notre questionnement ne se porte pas sur les données elles-mêmes, car la RTO qui représente les relations n-aires à extraire permet de décrire ces données recherchées. En revanche, notre questionnement porte sur l'expression des arguments et les relations qu'ils entretiennent dans les documents :

— De quelle manière les arguments s'expriment-ils ? Existe-t-il un schéma fréquent d'expression ? Si cela est le cas, quelle en est la structure sémantique ou quelle en est la structure lexicale ?

— Existe-t-il une (des) cooccurrence(s) fréquente(s) entre arguments de la relation n-aire ?

— Existe-t-il des liens spécifiques entre arguments autres que ceux connus dans la RTO ? Si oui, quelle est la typologie de ces liens ?

— Sachant que l'expression de la relation n-aire est dispersée dans le document, peut-on prévoir dans quelle mesure les arguments s'expriment ?

— Existe-t-il un élément fréquent qui permette de lier l'ensemble des arguments de la relation n-aire ?

— Existe-t-il des règles implicites s'appliquant aux arguments de la relation n-aire pour la construction de patrons d'expressivité ?

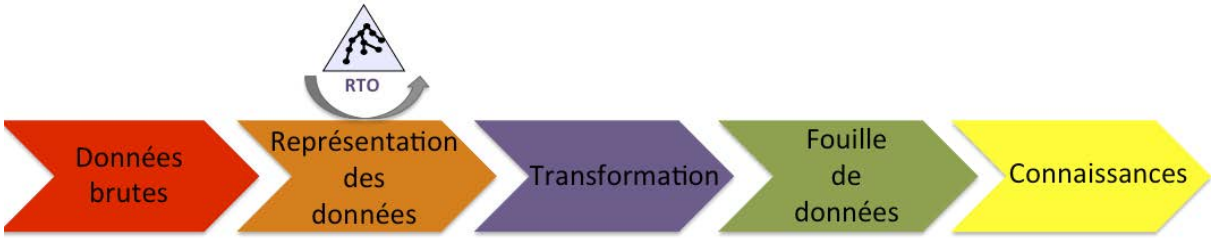


FIGURE 4.1: Schéma d'extraction des connaissances guidé par la RTO

Pour faire émerger et découvrir ces régularités dans l'expressivité des arguments de la relation, nous proposons d'utiliser les méthodes de fouille de données. La fouille de données est une étape fondamentale du processus d'extraction des connaissances correspondant à l'exploration et l'analyse d'un ensemble de données afin de découvrir de l'information utile et implicite. Un processus d'extraction des connaissances en 5 étapes fondamentales est proposé par (Fayyad et al., 1996) qui le définit comme *"un processus non trivial qui construit un modèle valide, nouveau, potentiellement utile et au final compréhensible, à partir de données"*.
Nous avons adapté ce processus à notre contexte de recherche d'expressivité des arguments de la relation n-aire. À partir du schéma d'extraction global, illustré dans la figure 4.1, nous proposons le processus d'extraction des connaissances guidé par la RTO de domaine, illustré dans la figure 4.2, afin de proposer de nouvelles représentations des données pour appliquer l'étape de fouille de données :

1. Les données brutes à étudier :
   Dans notre contexte d'extraction de connaissance à partir de données textuelles, il s'agit des articles scientifiques des domaines de spécialité que nous regroupons en corpus.

2. Le pré-traitement des données permet d'éliminer le bruit et traiter les données de qualité médiocre (données manquantes, erronées, incomplétude...) :
   Dans notre contexte, nous avons remplacé cette étape par trois différentes représentations des données, détaillées en section 4.2.3. Les représentations sont guidées par la RTO afin de rester fidèle à la modélisation des relations n-aires et permet d'appliquer les algorithmes de fouille de données sur des données textuelles non structurées. A partir de ces représentations, nous constituons notre ensemble de

données à fouiller, nommé base d'objets dont la constitution est détaillée dans la section 4.2.3.2. L'ensemble des items, que nous appelons également attributs ou, plus communément en fouille de textes, les descripteurs pertinents, sont sélectionnés comme détaillé dans la section 4.2.3.3.

3. La transformation ou la définition des structures optimales de représentation des données adaptée aux algorithmes, comme cela est illustré dans la figure 4.2 :
   Dans notre contexte, nous transformons chaque ensemble de données, à partir des attributs et selon les nouvelles représentations, en matrices.

4. La fouille de données et la tâche dévolue au processus (la classification, analyse, prédiction...). Dans notre contexte de représentations des données, nous verrons que nous appliquons les algorithmes sur des unités textuelles, que nous avons souhaité proches de la définition de la relation n-aire recherchée, sous forme de fenêtres pour préciser dans quelle séquence les relations implicites découvertes s'expriment. Les détails sont présentés dans la section 4.2.3.2.

5. Les connaissances extraites à capitaliser : l'analyse des résultats consiste à interpréter et évaluer les régularités intéressantes découvertes concernant l'expressivité des arguments impliqués dans la relation, à capitaliser en connaissance du domaine.

La section suivante présente les principaux algorithmes existant à l'état de l'art pour la fouille de données.

4.2.2 Principaux algorithmes de fouille données

Nous avons vu, à partir du schéma global 4.1, qu'une des étapes fondamentales est la fouille de données. La fouille de données a pour objet l'exploration et l'analyse d'un ensemble de données pour en extraire de la connaissance. Pour cela, plusieurs algorithmes existent pour construire des modèles à partir d'un ensemble de données. Ces modèles permettent de faire émerger des structures intéressantes qui vont permettre de dégager de la connaissance du domaine. Dans cette section, parmi les nombreux algorithmes existants, nous présentons les principaux algorithmes de la littérature pour la découverte de motifs et règles séquentiels et d'associations.

4.2.2.1 Quelques définitions utiles

Dans cette section, nous posons quelques définitions utiles à notre travail et initialement introduites dans (Agrawal and Srikant, 1995). Nous nous appuyons sur l'exemple

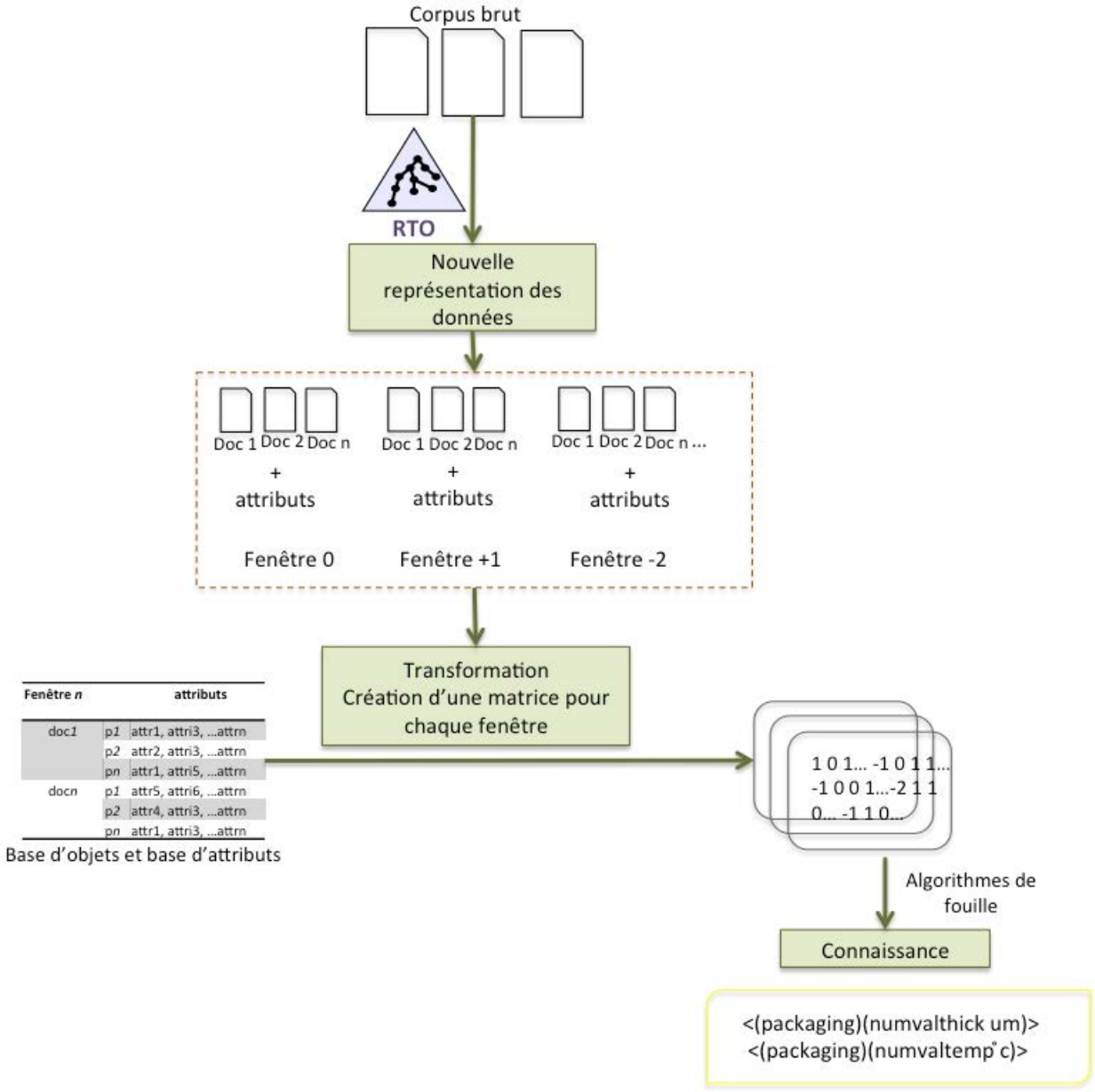


FIGURE 4.2: Processus d'extraction des connaissances guidé par la RTO

emprunté à (Fabrègue et al., 2012) pour illustrer ces définitions. La base de données $\mathcal{DB}$, représentée dans le tableau 4.1, regroupe l'ensemble des évènements ayant eu lieu dans plusieurs villes. Chaque transaction correspond aux événements d'une ville à une date donnée en considérant le triplet (id-ville, id-date, item) : l'identifiant de la ville, la date, l'ensemble des évènements (items) observés.

Ville	Mois	items
Nîmes	01/2011	Humidité=Faible, Soleil
Montpellier	02/2011	Soleil
Nîmes	02/2011	Chaleur=Forte
Montpellier	03/2011	Humidité=Faible, Chaleur=Forte
Nîmes	04/2011	Chaleur=Faible, Vent
Orange	04/2011	Pluie
Orange	06/2011	Pluie, Vent

Tableau 4.1: Base de données $\mathcal{DB}$.

Pour chaque ville, la base de séquences $\mathcal{S}$ est générée comme illustré dans le tableau 4.2

Ville	Séquence
Nîmes	<(Soleil, Humidité=Faible)(Chaleur=Forte)(Chaleur=Faible, Vent)>
Montpellier	<(Soleil)(Humidité=Faible, Chaleur=Forte)>
Orange	<(Pluie)(Pluie, Vent)>

Tableau 4.2: Base de séquences $\mathcal{S}$.

Définition 4 *Un **itemset** est un ensemble non vide d'items de $\mathcal{I} = \{i_1, i_2, ..., i_k\}$ noté $(i_1, i_2, ..., i_k)$. On dit qu'une transaction supporte l'itemset si l'itemset se trouve dans la transaction. Chaque itemset est caractérisé par un **support**. Le support est le nombre d'occurrences d'itemsets.*
L'extraction des itemsets fréquents est déterminée par un seuil $\sigma_s \in [0; 1]$. Les itemsets ayant un support supérieur au seuil sont extraits.

L'extraction des itemsets permet de construire des connaissances basées sur des corrélations appelées **règles d'associations** (Agrawal and Srikant, 1994).

Définition 5
Une règle d'association est de type $\mathcal{A} \longrightarrow \mathcal{B}$ *où* $\mathcal{A}$ *est un itemset tel que* $\mathcal{A} \subseteq \mathcal{I}$*, le conséquent* $\mathcal{B}$ *est également un itemset tel que* $\mathcal{B} \subseteq \mathcal{I}$ *et* $\mathcal{A} \cap \mathcal{B} = \emptyset$*.*
Le support de la règle $A \longrightarrow B$ *est* $S = support(A \cup B)$
La confiance de la règle $A \longrightarrow B$ *est* $C = \frac{support(A \cup B)}{support(A)}$
La confiance d'une règle indique la proportion des objets qui possèdent à la fois A et B parmi les objets qui possèdent A

Exemple 13 *Dans le tableau 4.2, une règle possible est Soleil => Chaleur=Forte. La règle est supportée par deux villes Montpellier et Nîmes. La confiance de la règle est de 1 car parmi les villes qui contiennent Soleil, on retrouve toujours Chaleur=Forte.*

L'extraction de motifs séquentiels peut être vue comme une extension de la problématique d'extraction des itemsets et des règles d'associations. Les itemsets et les règles d'associations restituent une connaissance sans la notion de temporalité. Les items ne sont pas ordonnés dans les motifs restitués. La notion de motifs **séquentiels** intègrent la notion de temporalité induisant la notion de séquence qui peut exister entre différents itemsets dans le motif découvert.

Définition 6 *Une* ***séquence*** *est une liste ordonnée, non vide, d'itemsets notée* $<(it_1), ...(it_n), ..., (it_k)>$ *où* (it_j) *est un itemset. La taille de la séquence est définie par le nombre d'items qu'elle contient.*

Définition 7 *La recherche de sous-séquences fréquentes, à partir des séquences, correspond à l'extraction des* ***motifs séquentiels****.*

Exemple 14 *Dans le tableau 4.2, la sous-séquence (i.e. motif séquentiel) s'=<(Soleil)-(Chaleur=Forte)> est supportée par les séquences des villes Nîmes et Montpellier. Son Support(s') = 2.*

4.2.2.2 Présentation générale de quelques algorithmes de fouille de données

Dans cette section, nous présentons le principe général de quelques algorithmes de la littérature de fouille de données pour l'extraction des règles d'association et des motifs et règles séquentiels.

Apriori (Agrawal and Srikant, 1994) L'algorithme *Apriori* peut être considéré comme l'algorithme générique, précurseur de tous les algorithmes de fouille de données. L'algorithme *Apriori* suit le principe de recherche en largeur d'abord, *breadth-first*, pour extraire les itemsets fréquents et règles d'associations.

Algorithm 1 Apriori

Entrées : une base de données $\mathcal{D}$, $\sigma_s \in [0; 1]$
Sortie : l'ensemble des motifs fréquents $\mathcal{F}$ extraits à partir du seuil fixé σ_s
Début
$i \leftarrow 1$
$C_1 \leftarrow$ ensemble des motifs de taille 1 (i.e. les items)

tant que $C_i \neq 0$ faire
Calculer le support de chaque motif $m \in C_i$
$F_i \leftarrow \{m \in C_i | support(m) \geq \sigma_s\}$
$C_{i+1} \leftarrow$ générer-candidats(F_i)
$i \leftarrow i + 1$
fin-tant que

Retourner $\bigcup_{i \geq 1} F_i$
Fin

L'algorithme effectue une extraction par niveau, selon le principe suivant :

1. Il recherche d'abord les motifs fréquents de taille 1.
2. Il combine les motifs de taille 1 pour obtenir des motifs de taille 2 pour lesquels on ne conserve que les fréquents.
3. Il combine les motifs de taille 2 pour obtenir des motifs de taille 3 pour lesquels on ne conserve que les fréquents.
4. Il poursuit ainsi jusqu'à l'obtention de la taille maximale définie dans $\mathcal{I}$, en conservant les fréquents.

De très nombreux algorithmes permettant de générer des itemsets fréquents et leurs règles d'associations dérivées existent, par exemple la recherche des itemsets fermés fréquents. Un itemset fermé fréquent est un ensemble maximal d'items communs à certains objets.

Exemple 15 *Dans l'exemple illustré dans le tableau 4.3, l'itemset (B, C, E) est un itemset fermé car il est l'ensemble maximal d'items communs aux objets {2, 3, 5, 6}. L'itemset seul (B, C) n'est pas un itemset fermé, car tous les objets contenant cet itemset commun,*

oid	items
1	A C D
2	B C E
3	A B C E
4	B E
5	A B C E
6	B C E

Tableau 4.3: Extraction d'itemsets fréquents

contiennent également en commun l'item E.

La recherche des itemsets fermés fréquents est une méthode optimisée pour rechercher les itemsets fréquents en évitant la redondance, comme cela est également le cas pour la recherche de motifs séquentiels fermés, avec l'algorithme clospan que nous présentons plus loin.

La recherche des règles séquentielles, comme cela est proposé avec l'algorithme CM-Rules (Fournier-Viger et al., 2012), nous intéresse également car elle propose de tenir compte de l'ordre entre les évènements qui composent la règle. L'algorithme recherche d'abord les règles d'associations et permet de réduire l'espace de recherche des items qui co-occurent dans plusieurs séquences. Puis, il élimine toutes les règles dont le support et la confiance minimum ne respectent pas le seuil prédéfini et retourne toutes les règles séquentielles communes à plusieurs séquences dans la base. L'algorithme est particulièrement efficace sur des seuils bas de support et se montre robuste en termes de passage à l'échelle.

SPADE (Zaki, 2001) La génération des algorithmes à laquelle appartient SPADE constitue les algorithmes pionniers dans la recherche des motifs séquentiels. SPADE fonctionne en cherchant à réduire l'espace de recherche des motifs, en les regroupant par catégorie et en considérant que les motifs fréquents présentent des préfixes communs. Les motifs séquentiels sont ainsi regroupés selon des classes d'équivalences construites sur une relation d'équivalence prenant en compte le préfixe des séquences. Le calcul de F2, les fréquents de taille 2, passe par une inversion de la base selon une représentation verticale, en inversant la méthode d'indexation de la base de données.

Algorithm 2 CMRules

Entrées : une base de séquences, minSeqSup, minSeqConf
Sortie : l'ensemble des règles séquentielles

1. Considérer la base de séquences comme une base de transaction
2. Trouver l'ensemble des règles d'associations à partir de la base de transaction, en appliquant un algorithme de recherche de règles d'associations comme *Apriori*. Paramétrer *minsup* = *minSeqSup* et *minconf*=*minSeqConf*.
3. Balayer la base de séquences originale afin de calculer le support séquentiel et la confiance séquentielle pour chaque règle d'association trouvée au cours de l'étape précédente. Eliminer chaque règle de la manière suivante :
   a. seqSup(r) $<$ minSeqSup
   b. seqConf(r) $<$ minSeqConf
4. Retourner l'ensemble des règles

Exemple 16 *Si nous reprenons notre tableau 4.2 représenté au format horizontal, la transformation au format vertical, représentée dans le tableau 4.4 revient à créer plusieurs bases temporaires qui associent à chaque séquence le couple (Id séquence, transaction) qui lui correspondent dans la base. Par exemple, pour la séquence <(Soleil)>, la base de données au format vertical enregistre une transaction pour l'Id. seq de Nîmes et une transaction pour l'id. seq de Montpellier.*

séquence <(Soleil)>	
Id. seq	transactions
Nîmes	1
Montpellier	1

Tableau 4.4: La base de données au format vertical pour la séquence <(Soleil)>

Pour que cette opération soit simplifiée, il faut charger en mémoire la base. L'algorithme propose ensuite deux stratégies pour la recherche de séquences fréquentes :
Breadth-First Search (BFS) est la stratégie employée par l'algorithme Apriori ou GSP (Srikant and Agrawal, 1996). Elle procède, comme nous l'avons décrit plus haut, niveau

par niveau, ce qui a l'avantage de procéder à un élagage beaucoup plus intéressant mais qui prend beaucoup de place mémoire.
DepthFirst Search (DFS) effectue les recherches branche par branche, et a l'avantage de traiter de manière plus optimale les ensembles de séquences fréquentes importantes. La génération des séquences candidates se fait par jointures successives. Le principal avantage de SPADE est d'effectuer des jointures sur les listes temporaires verticales afin d'optimiser les temps de calcul.

PrefixSpan (Pei et al., 2001) La prise en compte de la temporalité dans les transactions a conduit de nombreux auteurs à privilégier les méthodes selon le principe *depth-first* ou recherche en profondeur d'abord pour extraire les motifs fréquents. Les algorithmes PSP (Masseglia et al., 1998), Freespan (Han et al., 2000) et Prefixspan adoptent cette stratégie en mettant en place et en exploitant un arbre de préfixes pour gérer les candidats. L'objectif premier des auteurs est de réduire le nombre de candidats générés. Pour parvenir à cet objectif, PrefixSpan propose d'analyser les préfixes communs que présentent les séquences de données de la base à traiter. A partir de cette analyse, l'algorithme construit des bases de données intermédiaires qui sont construites à partir des bases d'origine en considérant les préfixes identifiés. Les nouvelles bases constituées ne contiennent alors que les suffixes. Le travail de l'algorithme s'effectue ensuite sur ces nouvelles projections, Prefixspan applique un comptage du support des différents items dans le but d'augmenter la taille des motifs séquentiels découverts. En pratique, la première étape d'extraction de motifs consiste à compter le support des séquences de taille 1. Si un préfixe est présent dans un nombre de séquences supérieur à un support minimum, alors ce préfixe est considéré comme fréquent. Lorsqu'un préfixe fréquent ou plusieurs sont trouvés, la base de données est divisée de manière récursive. Il n'est plus nécessaire de conserver toute la base de données. Les séquences qui ne supportent pas le préfixe courant (ou motif courant) ne sont pas conservées dans les bases projetées. En procédant à un comptage de tous les items fréquents sur les bases de données projetées, PrefixSpan extrait ainsi toutes les séquences fréquentes de taille 2. Le processus est répété selon une procédure récursive, jusqu'à ce que les bases de données projetées soient vides, ou qu'il n'y ait plus de séquences fréquentes possibles.

Clospan (Yan et al., 2003) L'extraction de motifs séquentiels devient problématique selon la taille des motifs séquentiels extraits. Un des problèmes mis en évidence par les auteurs de l'algorithme concerne une base de données qui ne contiendrait qu'un seul motif $< (a_1)(a_2)...(a100) >$. Il faudrait malgré tout générer les $2^{100}-1$ sous-séquences fréquentes avec le même support minimum de 1. Ces sous-séquences seraient alors redondantes. Les auteurs définissent ainsi la problématique de recherche des motifs séquentiels fermés avec

l'algorithme Clospan. L'algorithme évite le parcours des motifs redondants en détectant par avance les motifs séquentiels non clos. L'algorithme est basé sur le principe *depth-first* et implémente l'algorithme PrefixSpan. En fait, il s'agit d'une optimisation de ce dernier, destinée à élaguer l'espace de recherche en évitant de parcourir certaines branches dans le processus de divisions récursives. Le principe de CloSpan repose sur deux éléments essentiels : l'ordre lexicographique des séquences et la détection de liens systématiques entre deux items, par exemple être capable de détecter que β apparaît toujours avant α dans la base de données.

4.2.2.3 Choix des algorithmes pour les expérimentations

L'extraction de motifs a fait l'objet de nombreuses recherches dans le domaine de la fouille de données. De nombreux algorithmes ont été développés depuis le précurseur *Apriori*. Chacun apporte sa brique d'optimisation dans le processus d'extraction. Les algorithmes de la génération Prefixspan s'appuient sur de nouvelles structures de données et une génération de candidats efficace afin d'optimiser les résultats dans des espaces de recherche importants ou lorsque les données sont fortement corrélées ou encore lorsque le seuil de recherche de support est faible. Ils représentent actuellement les algorithmes les plus performants pour extraire les motifs séquentiels, que ce soit en termes de calcul ou bien en consommation de mémoire.
Dans ce contexte, nous cherchons à découvrir dans nos données, des schémas, des patrons intéressants pour le domaine de spécialité, impliquant les arguments de la relation n-aire. Nous nous intéressons plus particulièrement à découvrir des motifs séquentiels, des règles d'associations et des règles séquentielles. S'agissant de la découverte des règles d'associations, nous avons choisi, dans un premier temps, l'algorithme *Apriori* afin de parcourir l'ensemble complet des itemsets fréquents et d'obtenir des résultats exhaustifs. Pour les motifs séquentiels, nous avons choisi l'algorithme *Clospan* qui présente les mêmes avantages que l'algorithme PrefixSpan tout en permettant d'obtenir un ensemble concis de motifs sans perte d'information. Puis, nous choisissons d'affiner les règles obtenues avec l'algorithme *CMRules* qui permet d'extraire toutes les règles séquentielles de notre base de séquences. Les règles séquentielles, construites à partir des motifs séquentiels fréquents, ont l'avantage d'indiquer à la fois la fréquence et la fiabilité de la règle.

4.2.3 Nouvelles représentations des données textuelles guidée par la Ressource Termino-Ontologique

Dans cette section, nos propositions portent sur l'extraction de connaissances guidée par l'Ontologie d'un domaine et consistent à découvrir des relations implicites entre certains arguments de la relation n-aire. Dans cette tâche, nous établissons de nouvelles

représentations des données textuelles dont les définitions sont posées pour répondre à la problématique d'extraction de connaissances à partir de données textuelles de type expérimental, intégrant des concepts quantité et des concepts unités de mesure.
Les données textuelles de type expérimental correspondent aux résultats expérimentaux restitués dans les documents scientifiques du domaine étudié, suite à différentes expériences menées. Ces expériences font intervenir des paramètres de contrôle et des mesures représentés par une valeur numérique et leur unité de mesure. Nous reprenons notre hypothèse de travail 1 qui considère le concept Unité de mesure de l'Ontologie comme fondamental pour localiser les arguments quantitatifs du résultat expérimental.

4.2.3.1 Représentations des données

Nous proposons trois niveaux de représentation des données guidée par la Ressource Termino-Ontologique dans l'objectif de structurer nos textes pour les algorithmes de fouille de données. Le principe consiste à évaluer les trois types de représentation et conclure sur celle qui permet de découvrir des régularités intéressantes dans l'expressivité des arguments de la relation n-aire.
La première représentation des données, TDAT, se situe au niveau terminologique. Les données y sont représentées selon la terminologie exprimée dans les textes. Seuls les termes dénotant des instances d'unités de mesure sont représentés par le concept *Unit_ Concept*, ce dernier est remplacé par *um* (Unit of Measure) pour plus de simplicité d'écriture. Il permet, comme posé dans notre hypothèse 1, de considérer plus particulièrement un contexte favorable à l'expression des arguments de la relation n-aire. À partir de ce contexte favorable, nous cherchons à déterminer si la variabilité terminologique des textes suffit à restituer l'expressivité des arguments dans la relation.
Les deuxième et troisième représentations nommées respectivement GDAT et RDAT reposent sur la définition 8 qui s'appuie sur la définition des relations n-aires représentant les données expérimentales.

Définition 8
Représentation des données textuelles de type expérimental :
Etant donné une relation n-aire d'intérêt définie dans la RTO avec n arguments, les représentations GDAT et RDAT s'appuient sur les concepts correspondant aux arguments de la relation n-aire.

Dans la définition 8, les arguments de la relation n-aire peuvent être des arguments symboliques ou des arguments quantitatifs. Les arguments symboliques présentent une variabilité terminologique importante dans les textes. Par exemple, le sous-concept packaging est un argument symbolique dénoté par plusieurs termes. Nous proposons donc d'utiliser ces arguments symboliques, qui sont des concepts, pour représenter les termes

correspondant trouvés dans les données textuelles. En revanche, les arguments quantitatifs, e.g. les concepts *thickness, temperature...* (sous-concepts du concept *Quantity*) ont très peu de variabilité terminologique, ils ne s'expriment pas suffisamment dans les textes ou sont trop éloignés de leur valeur numérique. Nous avons donc choisi de représenter toute la terminologie associée aux arguments quantitatifs par le concept *Quantity* de la RTO. En utilisant le concept *Quantity* de l'Ontologie dans les nouvelles représentations, nous regroupons la terminologie associée aux arguments quantitatifs de manière à créer une meilleure variabilité terminologique et permettre ainsi de générer une meilleure fréquence d'expressivité des arguments quantitatifs de la relation n-aire dans les motifs restitués. Les données référencées dans la RTO et les valeurs numériques sont marquées par les concepts correspondants aux arguments de la relation n-aire, symbolisés par le sigle <>. Ces représentations octroient un niveau d'expression des données intéressant à exploiter avec les algorithmes de fouille de données afin de créer de la sémantique dans notre ensemble de données. En prenant l'exemple de la relation n-aire O2Permeability représentée dans la figure 2.3, les termes dénotant les arguments quantitatifs Thickness, Temperature, Partial_Pressure, Relative_Humidity et O2Permeability seront représentés avec le concept Quantity, alors que les termes dénotant des sous-concepts du concept Symbolic_Concept seront représentés par les arguments de la relation n-aire d'intérêt qu'ils dénotent (e.g. packaging). Dans notre proposition, les représentations des valeurs numériques jouent un rôle fondamental dans l'expression des résultats quantitatifs. En effet, comme nous l'avons expliqué, nous avons une très faible variabilité terminologique des arguments quantitatifs. La valeur numérique associée à l'unité de mesure permet fréquemment de déterminer à quel type d'argument quantitatif le résultat expérimental est associé.
Dans la représentation GDAT les valeurs numériques sont représentées en intégrant leur catégorie générique, *numval* (pour *valeur numérique*). La nouvelle représentation textuelle est illustrée dans l'exemple 17, à partir d'une phrase extraite du corpus des emballages.

Exemple 17
(a) Eight apple wedges were packaged into polypropylene trays and wrap-sealed using a 64 µm thickness polypropylene film with a permeability to oxygen of 110 cm^3 O[2] m^-2 bar^-1 day^-1 at 23 °C and 0% RH :

En suivant cette nouvelle représentation des données, les valeurs numériques et les descripteurs sont représentés par les concepts suivants *:*

(b) Eight apple wedges were packaged into polypropylene <packaging> trays and wrap-sealed using a 64 <numval> µm <um> thickness <quantity> polypropylene <packaging> film with a permeability to oxygen <quantity> of 110 <numval> cm^3 O[2] m^-2 bar^-1

day^-1 <um> at 23 <numval> °C <um> and 0 <numval> % <um> RH <quantity>

Dans la représentation RDAT, nous proposons d'affiner la représentation des valeurs numériques associées à des unités de mesure de la RTO. L'objectif est de représenter plus efficacement les valeurs numériques afin de permettre une meilleure expressivité des arguments quantitatifs de la relation n-aire. Au cours de nos expérimentations, il s'agit alors de représenter les valeurs numériques suivies d'unités de mesure rattachées à des concepts quantité définis dans la relation n-aire recherchée du domaine étudié : la température, l'épaisseur, l'humidité relative, la pression partielle et la perméabilité. Notre nouvelle représentation est illustrée dans l'exemple 18 en reprenant l'exemple 17.

Exemple 18
(a) Eight apple wedges were packaged into polypropylene trays and wrap-sealed using a 64 µm thickness polypropylene film with a permeability to oxygen of 110 cm^3 O[2] m^-2 bar^-1 day^-1 at 23 °C and 0% RH :

La nouvelle représentation des valeurs numériques permet de leur octroyer une valeur sémantique (éléments soulignés) correspondant aux concepts quantité de l'ontologie :

(c) Eight apple wedges were packaged into polypropylene <packaging> trays and wrap-sealed using a 64 <numvalthick> µm <um> thickness <quantity> polypropylene <packaging> film with a permeability to oxygen <quantity> of 110 <numvalperm> cm^3 O[2] m^-2 bar^-1 day^-1 <um> at 23 <numvaltemp> °C <um> and 0 <numvalrh> % <um> RH <quantity>.

Selon l'unité de mesure rattachée, Il existe trois types de valeurs numériques répondant à l'une des trois définitions suivantes :

Définition 9
Valeur numérique non ambiguë :
Une valeur numérique est non ambiguë si elle est suivie d'une unité de mesure dépendant uniquement du domaine étudié.

Par exemple, au cours de nos expérimentations, les valeurs numériques suivies de l'unité de mesure de perméabilité des emballages sont non ambiguës car elles dépendent uniquement du domaine étudié.

Définition 10
Valeur numérique ambiguë :
Une valeur numérique est ambiguë si elle est suivie d'une unité de mesure dénotant un seul concept quantité mais partagé par plusieurs domaines.

Par exemple, au cours de nos expérimentations, les valeurs numériques suivies de l'unité de mesure de température (i.e. °C) sont ambiguës car le concept quantité Température est partagé par plusieurs domaines.

Définition 11
Valeur numérique très ambiguë :
Une valeur numérique est très ambiguë si elle est suivie d'une unité de mesure dénotant plusieurs concepts quantité, ne dépendant pas nécessairement du domaine étudié.

Par exemple, au cours de nos expérimentations, les valeurs numériques suivies des termes % sont très ambiguës car cette unité de mesure dénote plusieurs concepts quantité dont celui de l'humidité relative.

À partir de nos documents, nous avons extrait certaines phrases comportant des valeurs numériques attribuées à des mesures de paramètres expérimentaux sans intérêt dans les instances de relations n-aires recherchées. L'exemple 19 illustre concrètement les problématiques soulevées concernant la représentation des valeurs numériques.

Exemple 19
The specimens were first fixed with 2.5% glutaraldehyde in phosphate buffer (pH 7.0) for more than 4 h.

L'unité de mesure % n'est pas associée à la mesure de l'humidité relative, elle dénote un autre concept quantité. La valeur est donc très ambiguë selon la définition 11

Films were conditioned at 25 °C and relative humidity of 45%.

Cette phrase présente une valeur ambiguë de température et très ambiguë d'humidité relative qui ne définissent pas les conditions de mesure de la perméabilité, mais les conditions d'emballage des aliments. Les paramètres à inclure dans l'instance de relation n-aire sont exprimés dans la phrase suivante du même article *:*

WVP of SPI films was evaluated at 25 °C and relative humidity of 54% according to ASTM standard method (ASTM E 96-93 1993).

The optimal composite emulsifier was beeswax, Span 20, and glycerol with tensile strength of 908 MPa.

La valeur numérique suivie de l'unité de mesure référencée dans l'ontologie (Mpa) répond à la définition 10 car elle ne définit pas la mesure de la pression partielle, liée à la perméabilité, mais la mesure de pression liée aux propriétés mécaniques de l'emballage. Le point de rupture du film est évalué à 908 Mpa. La pression partielle, liée à la perméabilité, est restituée dans la phrase suivante *:*

In the lower chamber, a controlled stream of pure O[2] flowed, thus maintaining constant the O[2] in this compartment (0.1 MPa).

Ces quelques exemples suffisent à illustrer la difficulté de ce processus automatique de représentation des données, du fait de la variabilité des paramètres expérimentaux existant dans les documents. Pour pallier ces ambiguïtés, nous avons établi une heuristique afin de mieux cibler les données d'intérêt :

Définition 12
Les valeurs numériques suivies d'une unité de mesure référencée dans la Ressource Termino-Ontologique sont sélectionnées si l'une des conditions suivantes est vérifiée :

- *La valeur numérique est non ambigüe car elle est suivie d'une unité de mesure dépendant uniquement du domaine étudié (e.g. une unité de perméabilité) ;*
- *La valeur numérique ambigüe est proche d'un terme d'unité dépendant uniquement du domaine étudié ou, est proche, en explorant le contexte des phrases environnantes, d'un concept quantité, dépendant du domaine et référencé dans la Ressource Termino-Ontologique (e.g. concept de perméabilité, d'épaisseur, d'emballage) ;*
- *La valeur numérique très ambigüe reprend les conditions de sélection de la valeur numérique ambigüe, à laquelle on ajoute qu'elle doit être proche, en explorant le contexte des phrases environnantes, du concept quantité (e.g. la valeur numérique suivie de l'unité de mesure % doit être proche d'un terme dénotant le concept quantité d'humidité relative, relative humidity ou RH).*

Exemple 20
The mushrooms were transported to the laboratory within 1 h of picking, then stored in darkness at 4 °C and 95% relative humidity (RH).

Cette phrase présente des valeurs de température et de perméabilité qui ne définissent pas nos instances recherchées. L'heuristique permet de les exclure du processus automatique de représentation des données. En effet, même si le terme de quantité "relative humidity (RH)" est présent, les conditions de proximité d'une unité dépendant du domaine et référencée dans l'Ontologie ou d'un descripteur ou concept quantité dépendants du domaine et présents dans l'Ontologie ne sont pas vérifiées, y compris dans la phrase précédente ou suivante, comme nous pouvons le constater en prospectant le contexte environnant :

Shiitake mushrooms (Lentinula edodes) used in this study were harvested in march from a local farm in Hangzhou, China.
The mushrooms were transported to the laboratory within 1 h of picking, then stored in darkness at 4 °C and 95% relative humidity (RH).
The next day the mushrooms were screened for uniform size and maturity and absence of mechanical damage.

La notion de proximité définie dans l'heuristique intègre la phrase où la valeur numérique à représenter est ciblée mais également la phrase qui précède ou qui suit. En effet, si dans la phrase précédente et/ou suivante, une des conditions est vérifiée, alors les valeurs numériques sont capturées dans le processus automatique. Ceci reste totalement cohérent car, même si les données ne se retrouvent pas dans la même phrase, mais dans un contexte proche, alors il existe une forte probabilité pour que ces données soient liées dans l'instance recherchée, comme le montre l'exemple 21.

Exemple 21
Analyses were performed at 25 °C and 70% RH.

A partir de cette phrase, étudiée isolément de son contexte, il est impossible de conclure si ces valeurs numériques sont des données d'intérêt pour nos instances de relations n-aires recherchées. En effet, comme nous l'avons mis en évidence dans les exemples précédents, la présence des unités de température et d'humidité relative ne permet pas cette conclusion. L'étude du contexte environnant, de la phrase précédente et/ou suivante, permet de produire suffisamment d'éléments pour permettre de conclure.

(a) Oxygen and carbon dioxide permeability were then evaluated.
(b) Analyses were performed at 25 °C and 70% RH.
(c) The permeation cell consisted in two stainless steel chambers separated by a film

sample.

La phrase (b) comporte une valeur numérique ambigüe et très ambigüe. La phrase précédente de celle-ci est la phrase (a). La phrase suivante est la phrase (c). Dans ce contexte, l'heuristique nous permet de conclure que les valeurs sont d'intérêt puisque les conditions se vérifient. Elles sont proches d'un descripteur dépendant du domaine et référencé dans l'Ontologie "Carbon dioxyde permeability" et l'unité de mesure % est proche du descripteur RH définissant son unité.

Nos propositions dans cette section permettent de :
- Définir trois nouvelles représentations des données textuelles.
- Définir une représentation RDAT qui s'appuie sur la définition des arguments des relations n-aires définies dans l'ontologie étudiée.
- Poser les définitions concernant l'ambiguïté des valeurs numériques suivies d'une unité de mesure référencée dans la Ressource Termino-Ontologique.
- Définir une heuristique pour désambiguïser la représentation des valeurs numériques suivies d'une unité de mesure référencée dans l'Ontologie.

4.2.3.2 Constitution de la base d'objets

A partir des trois nouvelles représentations des données, nous reprenons nos définitions de la section 2.3.2 et 2.3.3 s'appuyant sur l'hypothèse de travail 1 afin de constituer la base d'objets pour les algorithmes de fouille de données. Comme illustré dans la figure 4.2, chaque document du corpus est structuré en unités textuelles correspondant aux phrases sélectionnées selon la fenêtre étudiée. Ces phrases représentent les objets à fouiller, comme dans l'exemple 22 pour la fenêtre f_0.

Exemple 22 *Soit la fenêtre f_0 étudiée : elle permet de structurer tous les documents du corpus à partir des phrases pivot identifiées selon la définition 2 posée en section 2.3.2. Pour cette fenêtre, les phrases pivot constituent les objets à fouiller (les transactions).*

Propriété 1
La dimension et le sens de parcours des phrases de la fenêtre textuelle expriment la temporalité de l'expression des instances d'arguments de la relation n-aire découverte.

L'originalité de nos propositions est d'une part, de restreindre la recherche, en prospectant dans un contexte favorable à l'expression des instances d'arguments de la relation

n-aire, et d'autre part, en intégrant une notion de séquence plus précise, selon la propriété 1, dans les motifs et règles restitués par les algorithmes de fouille de données. En effet, dans les approches classiques de fouille de données, la notion de temps est effacée, les motifs et règles séquentiels n'expriment plus qu'une séquence indéfinie (i.e. β est suivi de α sans pouvoir exprimer quand exactement), les règles d'associations n'expriment plus que les corrélations. Notre proposition permet de restituer les motifs dans une fenêtre exprimant une séquence temporelle plus précise (i.e. β est suivi de α dans le contexte d'une fenêtre $f_{\pm 1}$, soit dans la fenêtre de la phrase précédente et suivante dans le contexte favorable d'expression des instances d'arguments).

4.2.3.3 Constitution de la base d'attributs

Dans cette section, nous décrivons la constitution de la base d'attributs. La base d'attributs (ou items pour les algorithmes de fouille de données) correspond à la sélection des descripteurs pertinents. Chaque objet à fouiller, décrit dans la section précédente et représenté dans la figure 4.2, contient un ensemble d'attributs associé.

Sélection des attributs TDAT. La base d'attributs est constituée des mots apparaissant au moins deux fois dans les documents. Elle sert de référence puisqu'elle suit l'approche classique *Sac-de-mots* utilisée en fouille de textes.

Sélection des attributs GDAT et RDAT. La base d'attributs est constituée en sélectionnant les mots apparaissant autour des concepts identifiés à partir des nouvelles représentations proposées. La sélection des mots autour des concepts identifiés est déterminée par des paramètres définis dans la section 4.2.3.4.

Exemple 23

Eight apple wedges were packaged into polypropylene <packaging> trays and wrap-sealed using a 64 <numvalthick> µm <um> thickness <quantity> polypropylene <packaging> film with a permeability to oxygen <quantity> of 110 <numvalperm> cm^3 O[2] m^-2 bar^-1 day^-1 <um> at 23 <numvaltemp> °C <um> and 0 <numvalrh> % <um> RH <quantity>.

Prenons l'exemple du concept identifié <packaging> et sélectionnons les 1-term autour de ce concept identifié afin de constituer la base d'attributs. Nous obtenons la base d'attributs suivante : <packaging>, polypropylene, trays, film.

Notre proposition permet de sélectionner les attributs, en s'appuyant sur les concepts représentant les arguments définis dans les relations n-aires, afin de découvrir l'expressivité des instances dans les textes. Cette expressivité restituée dans les motifs issus de la fouille de données s'effectue dans une fenêtre étudiée. Cette expressivité varie en fonction de la dimension de la fenêtre de la base d'objets et du nombre d'attributs sélectionnés.

Ces paramètres sont définis dans la section suivante.

4.2.3.4 Paramètrer les nouvelles représentations

Nous avons défini deux paramètres spécifiques aux nouvelles représentations des données textuelles afin de mieux maîtriser le nombre de motifs générés et l'observation, à partir des fenêtres, de l'expressivité des arguments de la relation n-aire :

- La dimension n de la fenêtre textuelle détermine le contexte phrastique autour de la phrase pivot, selon la définition 2, tout en restant dans un contexte favorable à la découverte de l'expressivité des instances d'arguments de la relation n-aire recherchée ;
- Le *n-term* détermine la sélection des attributs pertinents autour du concept identifié. Par exemple, n paramétré à 0 signifie que seuls les concepts sont pris en considération comme descripteurs.

Notre proposition, fondée sur les nouvelles représentations des données guidées par la RTO, permet de structurer les données pour les algorithmes de fouille de données. Notre base d'objets restitue les documents en considérant des fenêtres textuelles à fouiller, à partir de la base d'attributs, constituée de la terminologie et des concepts de la RTO représentés dans la définition de la relation n-aire.
La fenêtre permet de préciser la notion temporelle dans la séquence des motifs restitués.

4.2.3.5 Critères de sélection et d'évaluation

Nous proposons deux critères de sélection et d'évaluation supplémentaires afin de conserver les motifs et règles les plus proches de l'expression des instances de relations n-aires :

- Les motifs ou règles comportant au moins un argument de la relation n-aire, associés à une valeur du seuil de support permettant de générer des motifs de qualité et plus précis, au sens interprétables par l'expert ;

- Les motifs et règles issus de l'intersection entre différentes fenêtres étudiées et comportant au moins un argument de la relation n-aire.

Exemple 24
<(min)(water)> n'est pas un motif pertinent car il ne comporte aucun argument ciblé de la relation n-aire. <(packaging)(numvalthick um)> est un motif extrait avec une valeur du seuil de support de 0.5 dans une fenêtre $f_{\pm 1}$. Il est jugé pertinent car il comporte au moins un argument de la relation n-aire et est interprétable par l'expert.

Ces nouveaux critères permettent de réduire sensiblement le nombre de motifs et règles restitués, tout en conservant leur qualité et précision à l'analyse.

4.2.4 Expérimentations

4.2.4.1 Protocole expérimental

Nombre de documents. Nous avons évalué la méthode sur le corpus des emballages alimentaires constitué de 115 documents issus de la littérature du domaine. Nous n'avons pas pu tester la méthode sur le corpus de bioraffinerie car au moment des expérimentations, la RTO de domaine pour représenter les relations n-aires pertinentes était en cours de modélisation.

Nombre d'objets. Selon les fenêtres textuelles étudiées (e.g. $f_{\pm 1}$, $f_{\pm 2}$), le nombre de phrases à fouiller varie entre 5 000 et plus de 35 000. Ces phrases sont issues du corpus évalué.

Nombre d'attributs (ou items). Selon la fenêtre choisie et la sélection du paramètre *n-term* effectuée, le nombre d'attributs des différentes expérimentations varie entre 2 000 et plus de 10 000.

4.2.4.2 Résultats

Résultats sur TDAT.
Nous avons étudié chaque fenêtre du protocole et effectué une intersection entre elles afin de restituer les itemsets fréquents communs à toutes les fenêtres étudiées.

Le tableau 4.5 montre que la taille des itemsets fréquents ne dépasse que rarement la taille de 3 items et est majoritairement de taille 1. Dans la fenêtre $f_{\pm 2}$, l'extraction d'itemsets fréquents de taille 6 est observée mais ils ne présentent aucun intérêt pour

Fenêtre	Nombre	Taille
f_{-2}	39	3
f_{-1}	188	3
f_0	36	2
f_{+1}	90	3
f_{+2}	176	3
$f_{\pm 1}$	171	3
$f_{\pm 2}$	624	6
$\bigcap f_n$	36	2

Tableau 4.5: Nombre et taille des itemsets fréquents extraits selon les fenêtres textuelles étudiées.

Fenêtre	Itemsets fréquents	Support
f_{-2}	[°C]	0.507
	[films starch %]	0.053
	[films glycerol %]	0.055
	[films properties %]	0.052
f_{-2}	[%]	0.534
	[°c temperature]	0.106
f_0	[rh %]	0.1
	[°C stored]	0.057
$\bigcap f_n$	[°C %]	>0.05

Tableau 4.6: Exemples d'itemsets fréquents obtenus selon les fenêtres textuelles étudiées.

notre étude.
Le tableau 4.6 restitue quelques itemsets fréquents selon certaines fenêtres étudiées. Nous avons restitué uniquement les exemples révélateurs pour notre propos d'analyse sur ce cas d'étude, mais cette analyse est valable pour l'ensemble des autres fenêtres observées. Rappelons que, dans ce cas d'étude, la base d'items est constituée de tous les termes apparaissant plus d'une fois dans les documents. La base d'items est conséquente mais le premier constat est que les itemsets fréquents extraits sont assez pauvres et de faible taille, quelque soit la fenêtre étudiée. En effet, si nous analysons l'itemset fréquent *[films starch %]* de la fenêtre f_{-2}, nous pouvons déduire une forte corrélation entre les termes *films*, *starch*, désignant le nom d'un emballage alimentaire et l'unité de mesure *%*. Nous ne pouvons toutefois nullement conclure sur cette dernière unité. Il nous est en effet impossible de comprendre la signification de cette unité. Il nous est donc impossible de savoir si le terme *%* représente l'unité de mesure caractérisant l'humidité relative (autre argument impliqué dans la relation n-aire recherchée), ou un autre paramètre expérimental utilisant la même unité, ou bien encore le pourcentage de composant *starch* utilisé dans l'emballage étudié.
A cette granularité d'étude, qui consiste à extraire les itemsets fréquents à partir de la

terminologie, avec les termes apparaissant plus d'une fois dans le texte, ce cas de figure où les conclusions à apporter restent ambigües, se révèle majoritaire dans l'analyse des itemsets fréquents restitués.
Un autre cas de figure se révèle majoritaire également à cette granularité, il s'agit des itemsets fréquents évidents n'apportant aucune connaissance supplémentaire exploitable (e.g. *[rh %]*). Dans notre domaine d'étude, il est évident que le terme *rh* pour relative humidity est fortement corrélé à son unité de mesure *%*.
L'intersection des fenêtres étudiées révèle un 2-itemset fréquent intéressant. Si l'item *%* représente l'unité de mesure de l'humidité relative, alors cet itemset pourrait révéler une corrélation entre les arguments exprimant la température et l'humidité relative, à la condition que ces paramètres décrivent les conditions expérimentales de la mesure de perméabilité de l'emballage. Autrement dit, que ces unités de mesure soient rattachées aux instances de la Ressource Termino-Ontologique.
Un dernier point important à relever est l'absence totale d'expression des valeurs numériques dans les itemsets. En effet, rappelons que notre démarche vise à découvrir des régularités dans l'expressivité des instances de relations n-aires. L'instance va associer l'argument de la relation n-aire à sa valeur mesurée au cours des expérimentations menées sur les emballages alimentaires et restituées dans les articles scientifiques. A cette échelle, les valeurs numériques ont une grande variabilité qui inhibe leur expressivité dans les itemsets restitués. Ces valeurs ne ressortent donc pas dans les itemsets.
Nous pouvons donc conclure que l'échelle de travail sur les textes est à une granularité trop fine pour générer une expressivité intéressante à exploiter des instances représentées dans les documents. Dans la démarche suivante, nous intégrons les niveaux conceptuels de l'Ontologie afin de favoriser l'expressivité des instances d'arguments de la relation n-aire.

Résultats sur GDAT.
Au cours du protocole précédent, nous avons utilisé la terminologie apparaissant plus d'une fois dans les textes, comme base d'attributs. Les résultats montrent, malgré l'exhaustivité des descripteurs utilisés, une ambiguïté dans les itemsets fréquents générés et une expressivité des arguments faible voire inexistante, comme le cas des valeurs numériques.
Cette nouvelle approche exploite le niveau conceptuel de la RTO. En regroupant certains termes sous les concepts qu'ils dénotent, nous cherchons à générer une fréquence plus intéressante à exploiter dans l'expression des instances d'arguments, et produire ainsi un plus grand nombre de motifs exploitables. Au cours des expérimentations, nous n'avons pu expérimenter que la fenêtre f_0 (i.e. des phrases pivot), avec un 1-term d'attributs autour des concepts représentés et un seuil de support minimum au-delà de 0.6. Cette évaluation a généré une explosion de motifs (i.e. plus de 65 000) liée à la surexpression

des valeurs numériques avec la représentation *numval* considérée. En suivant l'approche TDAT, la forte variabilité des valeurs numériques ne permet pas de refléter leur expressivité dans les motifs. Puis, en intégrant un niveau conceptuel, un grand nombre de valeurs numériques s'expriment et le choix des n-term autour des concepts produit un grand nombre d'items dans la base d'attributs (i.e. plus de 10 000), générant ainsi un très grand nombre de motifs. Les expérimentations menées en suivant l'approche GDAT ne sont donc pas concluantes. L'expressivité des instances de relations n-aires reste pauvre à exploiter, générant des motifs peu précis voire impossibles à interpréter comme le restitue le tableau 4.7.

Exemples de motifs	Support
<(numval)(packaging)(numval um)(um)(um)(um)>	0.8
<(numval)(packaging)(numval)>	0.9
<(packaging)(quantity)(numval)(numval)(numval)>	0.69
<(packaging)(quantity)(numval)(numval)(quantity)>	0.7

Tableau 4.7: Motifs séquentiels de GDAT.

Résultats sur RDAT.
Comme présenté de manière détaillée dans la section 4.2.3.1, nous avons affiné la typologie de certaines valeurs numériques dans le but de les rapprocher des concepts quantité de l'Ontologie. L'appui de l'Ontologie, pour représenter les données en découvrant des motifs et règles séquentiels et des règles d'associations les plus proches de notre domaine d'application, a permis de réduire sensiblement le nombre de motifs et règles restitués. Ce nombre reste cependant encore conséquent, comme le montre un extrait des résultats obtenus à partir de 3-term dans le tableau 4.8. Dans le but de faciliter l'analyse des motifs et règles, restitués en très grand nombre au-delà de la fenêtre $f_{\pm 2}$, nous avons également procédé à l'obtention des motifs et règles communs à plusieurs fenêtres étudiées, quelque soit le support étudié, en effectuant une intersection entre fenêtres. Ces résultats sont restitués dans le tableau 4.8 avec $\bigcap_{f_n}$ RSeq et $\bigcap_{f_n}$ MS, pour représenter, respectivement, les règles séquentielles (RSeq) et les motifs séquentiels (MS) communs à plusieurs fenêtres.

Fenêtre	Nombre	Support
f_0	9958	>0.1
f_0	17987	>0.05
$f_{\pm 1}$	46926	>0.1
$f_{>\pm 2}$	>52000	>0.5
$\bigcap_{f_n}$ RSeq	575	>0.05
$\bigcap_{f_n}$ MS	404	>0.05

Tableau 4.8: Nombre de motifs séquentiels et règles extraits à partir de RDAT.

4.2.4.3 Discussion

Les résultats montrent que les motifs et règles séquentiels restitués sont plus pertinents et précis dans la découverte de régularités dans l'expressivité des instances d'arguments de la relation n-aire, comme le démontre l'extrait des résultats restitués dans le tableau 4.9. Nous remarquons que l'apparition de motifs d'expressivité des instances d'arguments de la relation n-aire se produit dès la fenêtre f_0, soit les phrases où au moins une unité de mesure reconnue de l'Ontologie apparait dans la phrase. En analysant de plus près les motifs et règles restitués, on découvre plusieurs cooccurrences entre arguments :

1. Le concept d'emballage, *packaging*, apparait fréquemment suivi d'une valeur numérique dénotant l'épaisseur, suivie de son concept d'unité de mesure ou d'un terme d'unité de mesure, e.g. *mm*, comme le suggère un de motifs du tableau 4.9, et ceci se produit dans une fenêtre maximale $f_{\pm 1}$. Cette corrélation se vérifie sur plusieurs motifs et règles restitués, y compris lors de la restitution des motifs et règles par intersection (e.g. packaging => numvalthick).

2. Le terme *materials* est fréquemment utilisé dans les documents et on le trouve souvent proche du concept *packaging*. Lorsqu'ils apparaissent ensemble, il apparait également un concept quantité, très probablement un concept d'épaisseur, comme cela se vérifie dans les autres motifs et règles.

3. Les termes *relative humidity* et *temperature* apparaissent fréquemment ensemble dans une fenêtre f_0. Cela est confirmé par la règle séquentielle *temperature* => *numvalrh*, qui suggère que lorsque le terme *temperature* est utilisé alors une valeur

Fenêtre	Motifs ou règles	Support
$f_{\pm 1}$	<(packaging)(numvalthick um)>	0.5
	<(numvalthick)(films)>	0.53
	<(films)(thickness values)>	0.1
	<(film thickness)(film)(observed)>	0.12
	<(film)(mm)(thickness)>	0.12
	<(film)(thickness)(determine)>	0.1
	<(film thickness)(rh)>	0.14
f_0	<(relative humidity)(humidity)(temperature)>	0.06
	<(pressure)(water permeability)>	0.053
	<(oxygen permeability)(pressure)>	0.05
	<(thickness)(films)(observed)>	0.052
	<(thickness)(water)(films)>	0.07
$\bigcap f_n$	packaging => numvalthick	>0.05
	quantity numvalthick => packaging	
	materials ==> packaging	
	packaging materials ==> quantity	
	temperature => numvalrh	
	quantity temperature ==> % numvalrh	
	packaging ==> quantity numvaltemp °c	
	packaging quantity => temperature numvalrh	
	<(packaging)(numvalthick)>	
	<(packaging)(numvaltemp °c)>	
	<(numvaltemp)(numvalrh%)>	

Tableau 4.9: Exemples de séquences obtenues à partir de RDAT.

numérique dénotant le concept quantité, humidité relative, est découvert. Cette corrélation est confirmée par le motif *<(numvaltemp)(numvalrh%)>* indiquant qu'une valeur de température est fréquemment suivie d'une valeur d'humidité relative et de son unité.

4. Les cooccurrences emballage - épaisseur et humidité relative - température apparaissent fréquemment dans un contexte textuel proche, comme le suggèrent les motifs *<(film thickness)(rh)>* et *<(packaging)(numvaltemp °c)>*, confirmé par la règle *packaging quantity => temperature numvalrh.*

5. Les motifs de la fenêtre f_0 montrent une cooccurrence entre les termes *water permeability* ou *oxygen permeability*, termes dénotant respectivement les concepts quantité *H[2]O permeability* et *O[2] permeability*, impliqués dans la relation n-aire, et le terme *pressure*. Ce dernier terme est inclus dans l'expression de l'argument *partial pressure*, ou pression partielle, impliqué également dans la relation n-aire.

Les points 1 et 2 permettent de conclure sur la cooccurrence de deux arguments impliqués dans la relation n-aire, l'emballage et son épaisseur. Cette cooccurrence apparait dans une fenêtre de phrases maximale $f_{\pm 1}$, donc ils sont très proches dans leur expression. De plus, une règle met en relief une corrélation fréquente entre le concept *packaging* et le terme *materials*, qui ne va pas sans rappeler la section *Materials and Methods*.
Le point 3 permet de découvrir la cooccurrence entre deux autres arguments impliqués dans la relation n-aire, la *température* et *l'humidité relative*, qui apparaissent très fréquemment dans un contexte textuel très proche.
Le point 4 montre clairement que l'expression co-occurrente de la température et de l'humidité relative s'exprime dans un contexte textuel très proche de l'expression co-occurrente de l'emballage et de son épaisseur.
Le point 5 suggère que l'expression de la pression partielle et de la perméabilité de l'emballage étudié (e.g. perméabilité à l'oxygène ou perméabilité à l'eau) sont co-occurrentes et toujours dans un contexte textuel très proche lorsqu'ils apparaissent. En effet, les motifs sont découverts dans une fenêtre f_0. Nous avons également évoqué la problématique d'analyse des résultats par fouille de données, liée au grand nombre de motifs restitués, notamment lorsque le seuil de support est bas. L'avantage d'analyser les motifs restitués à un niveau de support bas est qu'ils deviennent, en théorie, plus variés en proposant des niveaux mixtes, i.e. un niveau à la fois conceptuel et terminologique. Ils permettent ainsi une analyse plus pertinente et plus précise des régularités découvertes et sont interprétables par l'expert lors de la phase de validation. Dans notre cas, lorsque nous analysons plus précisément les motifs séquentiels restitués à des niveaux bas de support, nous re-

marquons que nous avons effectivement des motifs enrichis de la terminologie du domaine mais qui ne permettent néanmoins pas de préciser le contexte de manière beaucoup plus détaillée que les motifs extraits à des seuils de support plus haut. Reprenons des exemples extraits du tableau 4.9, le motif *<(packaging)(numvalthick um)* est extrait avec un support de 0.5 tandis que *<(thickness)(films)(observed)>* est extrait à un seuil de support de 0.052. Les deux motifs expriment la même tendance qui suggère que l'emballage co-occure avec l'épaisseur. Le deuxième motif restitué au niveau terminologique n'apporte pas d'éléments suffisamment complémentaires au premier motif restitué selon la nouvelle représentation guidée par la RTO. Cette observation nous permet de déduire deux points importants. D'une part, les représentations des données textuelles guidées par la RTO sont suffisamment pertinentes à la découverte de motifs variés à des seuils de support plus haut. Il est par conséquent pas forcément nécessaire de valider les motifs séquentiels à des niveaux plus bas restitués en très grand nombre. D'autre part, afin d'obtenir des informations complémentaires précisant et enrichissant les motifs extraits, l'analyse des relations syntaxiques est un bon compromis pour s'affranchir de l'analyse et de la validation d'un trop grand nombre de motifs, ce que nous verrons dans la section 4.3.

4.2.4.4 Conclusion

Dans cette section, nos travaux ont permis de poser des définitions dont les contributions s'adaptent à tout domaine exploitant des données textuelles de type expérimental formalisées sous forme de relation n-aire :

- En définissant de nouvelles représentations des données textuelles qui s'appuyent sur la RTO ;
- En définissant les valeurs numériques non ambiguë, ambiguë et très ambiguë ;
- En définissant une heuristique pour désambiguïser les valeurs numériques d'intérêt ;
- En définissant la phrase pivot et la fenêtre textuelle pour constituer la base d'objets à étudier à partir de laquelle découle la propriété de temporalité, propre à ces nouvelles représentations ;
- En définissant la sélection des attributs autour des concepts identifiés dans les relations n-aires d'intérêt définies dans la RTO ;
- En définissant de nouveaux paramètres de représentation afin de rechercher l'expressivité des instances en fonction de la terminologie et/ou des concepts de la RTO ;
- En définissant de nouveaux critères d'évaluation afin de restituer les motifs et règles les plus proches et les plus pertinents de l'expressivité des instances recherchées.

Les résultats des expérimentations menées avec les algorithmes de fouille de données montrent que nous avons pu réduire la problématique de recherche des instances de relation n-aire à la découverte de cooccurrences d'arguments et à la découverte de cooccurrences corrélées, dans un contexte textuel proche. Les résultats obtenus montrent que l'approche d'extraction des instances de relation n-aire doit être fractionnée en plusieurs étapes, dont la première correspond à nos contributions dans la recherche d'expressivité des instances d'arguments de la relation n-aire à partir des méthodes de fouille de données.
Une seconde étape, présentée dans la section suivante, consiste à enrichir ces motifs par l'analyse des relations syntaxiques pertinentes aux données expérimentales recherchées.

4.3 Vers une nouvelle approche hybride fondée sur l'analyse syntaxique

4.3.1 Introduction

Dans la section précédente, nous avons proposé trois représentations des données de type expérimental afin de découvrir, à partir des algorithmes de fouille de données, les relations implicites entre certains arguments de la relation n-aire. Nous avons montré que la problématique de recherche des instances de relation n-aire peut être réduite, en première étape, à la découverte de cooccurrences d'arguments (avec n arguments ≥ 2) de la relation n-aire. Puis, nous avons situé plusieurs cooccurrences entre elles dans une séquence déterminée, représentée par la fenêtre textuelle. Les résultats obtenus ont permis de découvrir des motifs d'expressivité des arguments de la relation n-aire. Nous avons également constaté que la nouvelle représentation des données textuelles guidée par la RTO du domaine permet de restituer des motifs séquentiels, suffisamment variés, à des seuils de support suffisamment élevés.

Dans cette section, nous avons initié un travail qui évalue le potentiel d'une approche hybride fondée sur l'analyse des relations syntaxiques. Notre objectif consiste, dans cette étape, à rechercher les relations syntaxiques (RS) pertinentes pour les données expérimentales et engageant les cooccurrences restituées dans les motifs séquentiels (MS), au cours de l'étape de fouille de données, et susceptibles de les étendre. S'il existe des relations syntaxiques spécifiques aux motifs d'expressivité des arguments de la relation n-aire, les RS pertinentes peuvent alors être utilisées pour étendre, enrichir les MS découverts et contribuer à l'extraction des instances engageant plus de deux arguments de la relation n-aire. La figure 4.3 présente le principe de la méthode hybride : la première étape représente l'approche par fouille de données et l'obtention de motifs d'expressivité des

arguments, la deuxième étape correspond à l'approche par analyse des triplets de dépendance syntaxique, la troisième étape propose de combiner les approches afin de construire des motifs étendus d'extraction des instances d'arguments corrélés dans les textes.

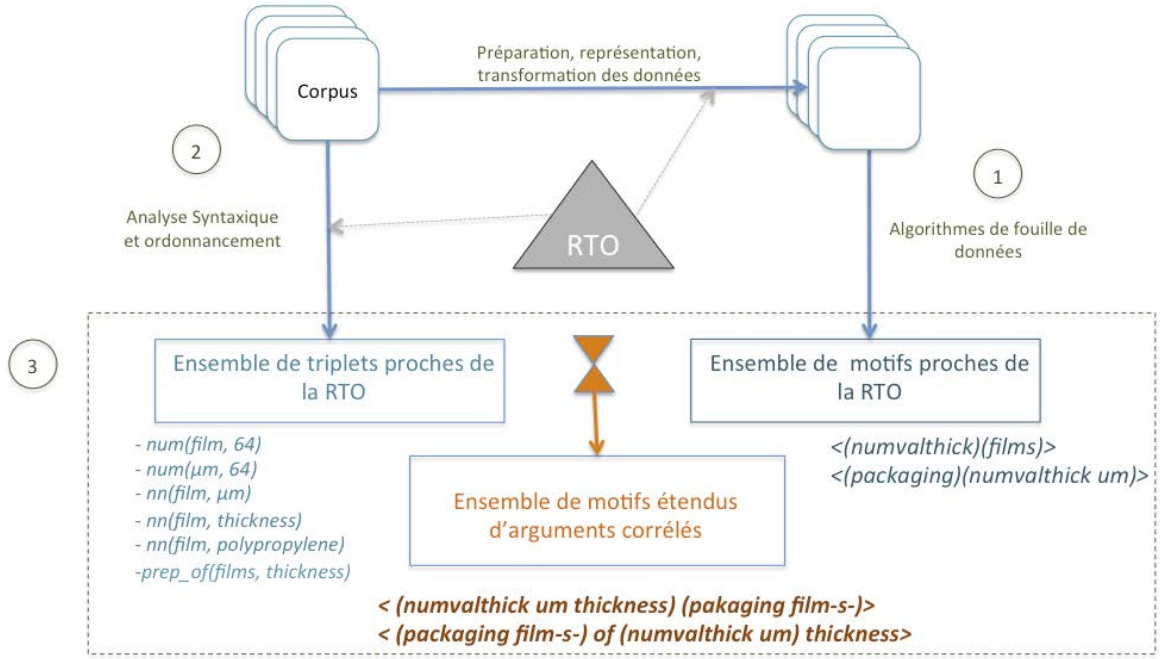


FIGURE 4.3: Architecture globale de l'approche hybride.

Principe de l'analyse syntaxique. L'analyse syntaxique permet de déterminer la structure grammaticale de la phrase selon deux méthodes généralement utilisées :

- L'analyse syntagmatique dans laquelle la phrase est représentée par des syntagmes imbriqués et la fonction syntaxique est attribuée aux groupes de mots (syntagmes). Il existe ainsi différents types de syntagmes, e.g. syntagme nominal, verbal, adjectival. Le but de l'analyse syntagmatique est de trouver les syntagmes simples, par une analyse superficielle, et, les syntagmes imbriqués, à partir de l'analyse syntagmatique profonde. La représentation de la décomposition de la phrase en ses constituants se fait le plus souvent sous forme d'arbre syntagmatique ;

- L'analyse des dépendances syntaxiques établit les relations grammaticales entre paires de mots de la phrase. Le résultat se présente sous forme d'arbre dont les

feuilles représentent les mots et les branches relient les mots par des relations grammaticales ou encore sous forme d'un ensemble de triplets de dépendances.

De nombreux analyseurs existent à l'état de l'art dont Minipar (Lin and Dekang, 1998), Link parser (Sleator and Temperley, 1995) et Stanford parser (Klein and Manning, 2003) pour l'anglais. L'analyseur de Stanford est un analyseur statistique en dépendances offrant un bon compromis entre rapidité et pertinence d'analyse. Il atteint jusqu'à 86% de précision, ce qui le place parmi les plus performants à l'état de l'art. Il restitue trois types de résultat correspondant chacun à un type d'analyse :

— **Les parties du discours** pour les mots de la phrase en utilisant l'annotation morpho-syntaxique du Penn Treebank [1] ;

— **L'arbre syntaxique** en constituants composé de regroupements grammaticaux sous forme d'arbres ;

— **Les dépendances** produites en s'appuyant sur une grammaire de dépendances et une hiérarchie de relations grammaticales développées par (de Marneffe et al., 2006). L'analyseur contient ainsi un peu plus de 50 relations grammaticales hiérarchisées qui permettent d'analyser la phrase et de produire une relation grammaticale entre chaque paire de mots de la phrase.

Le choix du type d'analyse dépend grandement de l'application attendue. Dans notre contexte, notre motivation concerne l'obtention d'une analyse de dépendances syntaxiques exhaustive afin d'étendre les motifs obtenus dans l'approche de fouille de données. Etant bien évalué à l'état de l'art avec un bon taux de précision, nous avons choisi de travailler avec l'analyseur syntaxique de Stanford.
La section 4.3.2 présente le principe d'une analyse classique. La section 4.3.3 présente nos propositions pour la combinaison des MS et RS. La sous-section 4.3.3.1 présente la méthode proposée afin d'extraire les relations syntaxiques proches de la RTO, la sous-section 4.3.3.2 présente une nouvelle fonction afin d'ordonnancer ces relations syntaxiques par rapport aux motifs extraits au cours de l'étape de fouille de données. La sous-section 4.3.3.3 propose de combiner les motifs et les relations syntaxiques pertinentes afin d'étendre les motifs pour l'extraction des instances d'arguments corrélés. Nous évaluons ce travail dans la section 4.3.4.

1. Le Penn Treebank est un corpus syntaxique étiqueté et analysé (http ://www.cis.upenn.edu/ treebank/)

4.3.2 Analyse syntaxique non guidée

Dans cette section, nous analysons les RS en sortie d'analyseur. L'analyseur produit, pour chaque phrase du document, l'ensemble des RS entre paire de mots de la phrase analysée. L'exemple 25 restitue l'ensemble des RS de la phrase proposée. Le triplet est composé du lien grammatical associant la paire de mots, des mots associés et de leur position dans la phrase.

Exemple 25
The dried films were peeled off from the glass plates and cut into test specimens.
det(films-3, The-1)
amod(films-3, dried-2)
nsubjpass(peeled-5, films-3)
nsubjpass(cut-12, films-3)
auxpass(peeled-5, were-4)
root(ROOT-0, peeled-5)
prt(peeled-5, off-6)
det(plates-10, the-8)
nn(plates-10, glass-9)
prep_from(peeled-5, plates-10)
conj_and(peeled-5, cut-12)
nn(specimens-15, test-14)
prep_into(cut-12, specimens-15)

Dans un premier temps, nous avons classé les triplets en fonction des étiquettes grammaticales apparaissant le plus fréquemment. Cependant, l'analyse du regroupement des étiquettes grammaticales les plus fréquentes ne permet pas de restituer celles pertinentes impliquant les instances d'arguments de la relation n-aire. En somme, les étiquettes grammaticales les plus représentatives en sortie brute de l'analyseur ne sont pas pertinentes pour le domaine étudié. Par conséquent, nous proposons dans la section suivante, d'utiliser la connaissance du domaine afin de guider la découverte des relations syntaxiques d'intérêt, les RS candidates.

4.3.3 Combinaison des MS et RS

4.3.3.1 Analyse syntaxique guidée par la RTO

Dans cette section, nous proposons dans un premier temps, d'identifier les RS en s'appuyant sur la définition de la relation n-aire. Cette première étape a pour objectif de

réduire les RS restituées par l'analyseur aux RS candidates selon la définition 13.

Définition 13 *Une relation syntaxique (RS) est définie comme candidate si elle comporte au moins un terme dénotant un argument des relations n-aires définies dans la RTO, i.e. symbolique, quantitatif, ou une unité de mesure d'un de ces arguments.*

Exemple 26
The thickness of resulting starch films ranged from 199.6 ± 22.6 *to* 271.4 ± 581 μm.
det(thickness-2, The-1)
nsubj(ranged-7, thickness-2)
amod(films-6, resulting-4)
nn(films-6, starch-5)
prep_of(thickness-2, films-6)
root(ROOT-0, ranged-7)
number(±-10, 199.6-9)
num(271.4-13, ±-10)
number(271.4-13, 22.6-11)
dep(271.4-13, to-12)
prep_from(ranged-7, 271.4-13)
xcomp(ranged-7, ±-14)
num(µm-16, 581-15)
dobj(±-14, µm-16)

Dans l'exemple 26, la définition 13 guide l'extraction des RS candidates en sélectionnant toutes celles comportant au moins un terme dénotant un des arguments ou leurs unités de mesure d'une relation n-aire de la RTO. Les RS candidates sont celles comportant les termes *thickness, starch* et μm. Le terme *starch* dénote le concept *packaging* défini dans la relation n-aire. À ce niveau de l'analyse et de l'extraction des RS candidates, les valeurs numériques associées à l'instance analysée ne sont pas extraites. Cette perte n'est pas dommageable, car notre intérêt dans l'analyse des RS candidates ne porte pas sur l'extraction des instances mais l'extraction des étiquettes grammaticales associant des arguments de la relation n-aire.

4.3.3.2 Une nouvelle fonction de rang

Dans un deuxième temps, nous proposons une fonction de rang *n-ary ranking*, illustrée dans la figure 4.4, qui permet d'ordonnancer les RS candidates identifiées par rapport aux

MS. Elle effectue d'abord une intersection entre les RS candidates et les MS obtenus au cours de l'étape de fouilles de données. Cette étape permet, parmi les RS candidates, de restituer celles partageant des points communs avec les MS. Ces RS sont alors les candidates à enrichir les MS pour l'obtention des MS étendus afin d'extraire les instances d'arguments corrélés comme cela est présenté dans l'exemple 27. Puis, la fonction restitue les RS et MS spécifiques.

Exemple 27 *Considérons le motif séquentiel extrait dans la fenêtre $f_{\pm 1}$, <(numvaltemp °c)(numvalrh%)>. Le motif met en relief une corrélation fréquente entre deux arguments de la relation n-aire, l'expression de la température et de l'humidité relative. Parmi les RS communes extraites, nous avons conj_and(°C, rh). Elle est commune car elle partage au moins un terme commun avec le MS, °C. Elle permet ainsi d'étendre, d'enrichir le motif extrait d'une manière plus fine. En effet, elle indique que dans la fenêtre d'extraction du motif $f_{\pm 1}$, l'expression des deux arguments sont très proches puisqu'ils sont associés par une conjonction de coordination de type "and". Cette information syntaxique complémentaire permet d'affiner la position des arguments entre eux dans la fenêtre de découverte du motif.*

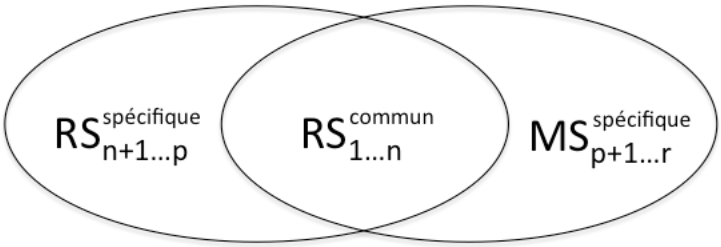


FIGURE 4.4: Fonction n-ary ranking

La fonction proposée permet d'obtenir un nouvel ordonnancement des RS et MS, non plus en considérant une fréquence classique mais, en considérant les différentes fréquences des RS et MS, en fonction des étiquettes grammaticales les plus représentatives :

- Les étiquettes grammaticales issues des relations communes, notées $RS_{1...n}^{commun}$, correspondent à l'ensemble des relations présentes dans les MS et RS. Nous utiliserons la fréquence des RS pour les classer ;

- Les étiquettes grammaticales issues des relations spécifiques, notées $RS_{n+1...p}^{specifique}$, correspondent à l'ensemble des relations qui ne s'expriment pas dans les motifs.

Nous utiliserons la fréquence des RS pour les classer ;

- Les motifs notés $MS_{p+1...r}^{specifique}$, correspondent à l'ensemble des motifs n'existant pas en tant que RS. Nous utiliserons le support pour les classer.

La fonction de rang *n-ary ranking* permet ainsi de rapprocher de manière plus précise les RS candidates et MS découverts tout en restant proches de la définition de la relation n-aire :

1. Elle considère les RS candidates pertinentes pour l'extension des MS en restituant celles partageant des points communs. Ces RS candidates sélectionnées, notées $RS_{1...n}^{commun}$, représentent des patrons syntaxiques intéressants à l'extension des MS. En reprenant l'exemple 26, la RS restitue le lien entre le film et l'épaisseur, ce lien est également représenté dans les MS découverts concernant la cooccurrence entre l'argument *thickness* et *packaging* de la relation n-aire.
2. Elle considère les étiquettes grammaticales intégrant des termes, toujours proches de la définition de la relation n-aire, tout en suggérant un contexte plus précis de l'expression des arguments de la relation n-aire, à partir de l'ordonnancement des $RS_{n+1...p}^{specifique}$. Par exemple, comme nous le verrons plus en détails dans les résultats en section 4.3.4, nous retrouvons dans les RS spécifiques des liens grammaticaux associant des arguments de la relation n-aire avec des verbes de type expérimental suggérant un contexte descriptif ou d'analyse des arguments de la relation n-aire.
3. Elle suggère un contexte plus global d'expression des MS découverts à partir de l'ordonnancement des $MS_{p+1...r}^{specifique}$ en intégrant la notion de temporalité des arguments déterminée dans des fenêtres textuelles.

4.3.3.3 Extension des MS par les RS

Dans un troisième temps, nous travaillons à partir des RS candidates communes aux MS restitués par la fonction *n-ary ranking*. Nous illustrerons le principe de la méthode hybride à partir des exemples de RS et MS communs extraits du corpus des emballages. À partir des MS restituant les cooccurrences entre les arguments *thickness, packaging* et entre *temperature, relative humidity*, illustrés dans la figure 4.5, nous recherchons parmi les RS candidates restituées par la fonction *n-ary ranking* celles comportant ces arguments afin d'étendre les motifs. Parmi les RS candidates communes restituées, nous en trouvons fréquemment entre le terme *thickness* et le terme *films*, e.g. lien grammatical de type prépositionnel. Ce lien est exprimé par le patron syntaxique *thickness [prep] films*. Nous en trouvons également fréquemment entre *thickness*, *films* et un terme dénotant le concept packaging, e.g. *starch, ldpe*, soit le patron syntaxique *[packaging] films*. En combinant

les patrons, nous obtenons le patron syntaxique *thickness [prep] [packaging] films*. Ce patron syntaxique est pertinent à l'extension de tous les MS restituant la cooccurrence entre l'épaisseur et l'emballage, e.g. *<(packaging)(numvalthick um)>* ou *<(numvalthick um)(packaging)>*. Le principe de l'extension du MS fonctionne en effectuant une jointure sur le concept commun *packaging* afin d'obtenir des motifs étendus pour l'extraction d'arguments corrélés, e.g. *<thickness [prep] (packaging) -film-s- (numvalthick um)>* ou *<thickness [prep] -film-s- (numvalthick um)(packaging)>*.

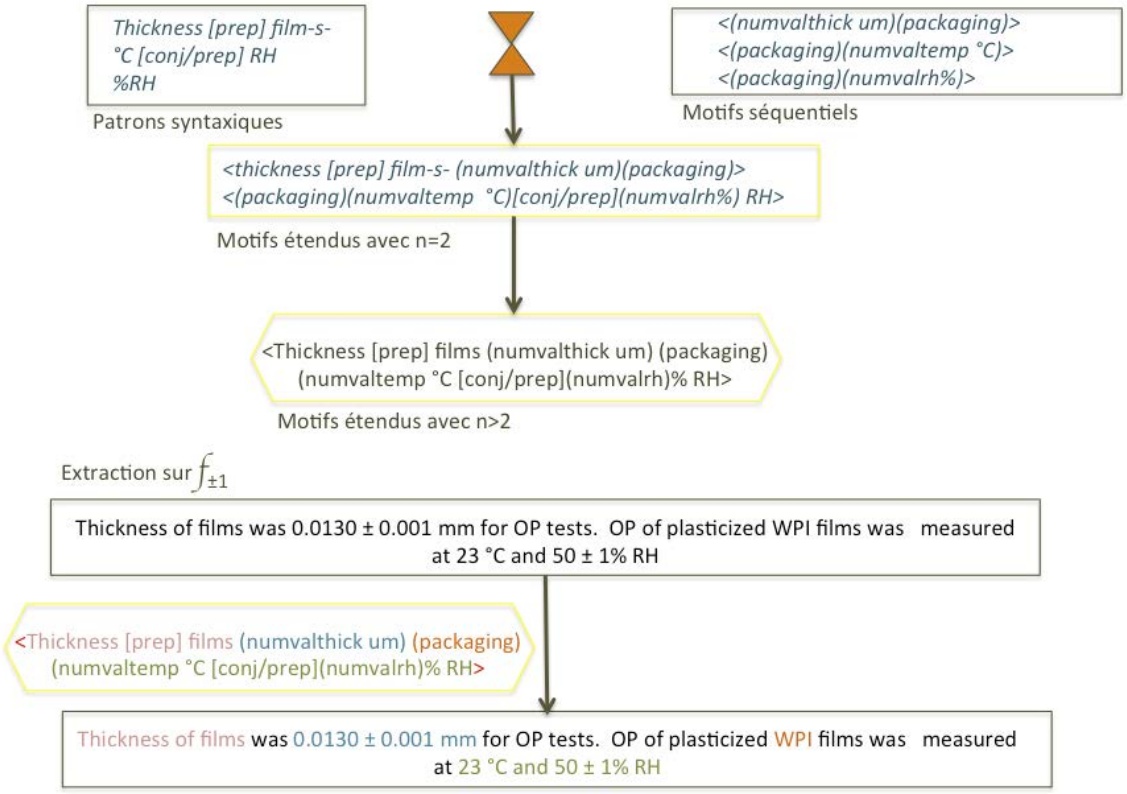


FIGURE 4.5: Motifs étendus de n arguments corrélés par approche hybride (exemple d'instance extraite).

L'approche hybride ne traite, dans un premier temps, que les cas des combinaisons entre RS communes aux MS restitués par la fonction *n-ary ranking*, car ce sont elles qui vont nous permettre d'étendre les MS communs restitués. Dans un deuxième temps, il serait intéressant d'analyser les RS spécifiques afin d'étendre les MS à des motifs plus complets pour couvrir une meilleure extraction. Ce travail n'a pas été effectué dans le cadre

des expérimentations et est envisagé en perspectives des travaux menés sur l'approche hybride.
La section suivante présente les résultats et discussion sur l'extraction et l'ordonnancement des RS candidates en étant guidés par la RTO. Puis, nous restituons les premiers résultats issus des tests d'évaluation de la méthode hybride sur quelques RS et MS combinés.

4.3.4 Expérimentations et résultats

4.3.4.1 Résultats de l'analyse syntaxique guidée par la RTO

Dans le cadre des expérimentations menées, à partir du corpus des emballages en s'appuyant sur la RTO naRyQ_emb, nous avons d'abord effectué une analyse syntaxique brute. Nous avons analysé les RS en sortie d'analyseur et nous avons effectué un classement des RS les plus fréquentes. Cette analyse produit plus de 50 000 RS, dont l'ordonnancement selon la fréquence ne nous a pas permis d'identifier de manière pertinente et adaptée les relations syntaxiques proches des arguments de la relation n-aire. En effet, les RS les plus fréquentes sont, par exemple les *prépositions*, les *déterminants* ou encore les *conjonctions de coordination* n'intégrant pas d'arguments de la relation n-aire, e.g. *det(vitamin, the), prep_in(colour, legend).*

Nous avons ensuite adopté notre approche d'analyse guidée par la RTO du domaine afin d'extraire les RS candidates au plus proches de la relation n-aire. Nous avons réduit les relations à 6613 RS candidates d'intérêt pour le domaine. La fonction de rang *n-ary ranking* présentée dans le tableau 4.10 permet d'analyser la diversité et le nombre d'occurrences des liens grammaticaux utilisés dans les documents afin d'exprimer les instances d'arguments. Les résultats obtenus montrent que les RS se retrouvant dans les MS, sont celles qui sont le plus proches de la définition de la relation n-aire et des propriétés définies dans la RTO. Ces RS sont à la base de notre travail, détaillé dans la section 4.3.3.3, de rapprochement et de combinaison avec les MS afin d'étendre les motifs découverts. Cette extension des MS permet de proposer des motifs étendus d'extraction des instances d'arguments corrélés définis dans la relation n-aire ayant plus de 2 arguments corrélés.
La fonction *n-ary ranking* permet également de mettre en évidence une proportion non négligeable de RS spécifiques, permettant d'appréhender avec plus de précision le contexte dans lequel est exprimé l'argument impliqué :

1. L'emploi de certains adjectifs et adverbes (e.g. *thin, thinner, thinnest, tickest, thicker*) associés en RS avec le terme *film*, montre que nous sommes dans un contexte de discussion des résultats obtenus sur l'emballage étudié et plus précisément concernant l'épaisseur de cet emballage.

2. L'emploi de certains verbes et noms (i.e. *measurement, measured, ranged, set*, associés à un terme de quantité (e.g. *thickness, temperature*), montre que nous sommes dans un contexte de mesure de certains arguments de la relation n-aire.

3. L'emploi de verbes de typologie expérimentale associés à un terme proche de la relation n-aire, est fréquemment observé. La fonction *n-ary ranking* nous permet de mettre en évidence que, parmi ces verbes, deux catégories se distinguent essentiellement :
   - La catégorie des verbes expérimentaux de type descriptif suggère que nous sommes dans un contexte de description des conditions expérimentales, e.g. *conduct, use, prepare, set, choose, keep* associés avec des termes de quantité comme *humidity, temperature* ou *film* ;
   - La catégorie des verbes expérimentaux de type analyse suggère que nous sommes dans un contexte de discussion des résultats obtenus sur les emballages évalués, e.g. *compare, increase, decrease, indicate, result* associés essentiellement à des termes de quantité.

Rang	n-ary ranking	nombre d'occurrences	Exemples
1	RS_1^{commun}	220	num(°c, <nombre>)
2	RS_2^{commun}	127	num(%, <nombre>)
3	RS_3^{commun}	84	number(%, <nombre>)
4	RS_4^{commun}	59	nn(thickness, film)
5	RS_5^{commun}	52	num(mm, <nombre>)
6	RS_6^{commun}	48	num(cm, <nombre>)
7	RS_7^{commun}	30	amod(rh, %)
8	RS_8^{commun}	29	num(°, <nombre>)
9	RS_9^{commun}	23	num(μm, <nombre>)
10	RS_{10}^{commun}	22	conj_and(°c, %)
...	...	...	...
21	RS_{21}^{commun}	10	prep_of(thickness, films)

22	RS_{22}^{commun}	10	conj_and($^{\circ}c$, humidity)
23	RS_{23}^{commun}	9	conj_and($^{\circ}c$, rh)
...	...	...	...
128	RS_{128}^{commun}	1	amod(thick, <nombre>-μm)
129	RS_{129}^{commun}	1	prep_of(overpressure, kpa)
130	RS_{130}^{commun}	1	rcmod(pressure, kpa)
131	RS_{131}^{commun}	1	nsubj(thickness, film)
132	RS_{132}^{commun}	1	amod(film, thickness)
133	$RS_{133}^{specifique}$	31	amod(film-s-, modified)
134	$RS_{134}^{specifique}$	28	nsubjpass(measured, thickness)
135	$RS_{135}^{specifique}$	13	nsubj(showed, films)
136	$RS_{136}^{specifique}$	12	nsubjpass(prepared, films)
137	$RS_{137}^{specifique}$	11	amod(films, investigated)
...	...	...	...
432	$MS_{432}^{specifique}$	111	<(packaging)(numvalrh%)>
433	$MS_{433}^{specifique}$	104	<(packaging)(numvalthick um)>
434	$MS_{434}^{specifique}$	103	<(packaging)(numvaltemp $^{\circ}c$)
...	...	...	...
471	$MS_{4710}^{specifique}$	71	<(numvalthick um)(packaging)>
...	...	...	...

Tableau 4.10: Ordonnancement des RS candidates avec *n-ary ranking*

4.3.4.2 Résultats de l'approche hybride

Les résultats des évaluations de l'approche hybride sont restitués selon les mesures de précision, de rappel et de F-mesure, à partir des motifs étendus comprenant 2 à 4 arguments corrélés découverts dans l'étape de fouille de données : *packaging, thickness* et *temperature, relative humidity*. Pour analyser les premiers résultats restitués par la méthode

hybride, nous avons effectué un échantillonnage des RS et MS communs, c'est-à-dire que nous avons conservé seulement 4 types de liens grammaticaux : nominal, numérique, les conjonctions et les prépositions dans lesquels apparaissent les 4 arguments de la relation n-aire étudiés et, 6 motifs fréquents : <(packaging)(numvalthick um)>, <(numvalthick um)(packaging)>, <(packaging)(numvaltemp °c)>, <(numvaltemp °c)(packaging)>, <(packaging)(numvalrh%)> et <(numvalrh%)(packaging)>. Ces motifs nous semblent particulièrement intéressants à évaluer par la méthode hybride car nous les retrouvons également dans l'ensemble des motifs restitués lors de l'intersection de toutes les fenêtres en approche de fouille de données, dans le tableau 4.8, notés $\bigcap_{f_n}$ MS. Il est entendu que cette stratégie d'échantillonnage restreint grandement l'exhaustivité et le choix des RS et MS mais nous permet déjà d'évaluer la méthode hybride sur l'extraction des instances engageant de 2 à 4 arguments sur 6 de la relation n-aire. Nous obtenons ainsi une quarantaine de motifs étendus d'arguments corrélés, avec un nombre d'arguments $n \geq 2$, dont un extrait est présenté dans le tableau 4.11.

Exemples de motifs étendus avec $n \geq 2$
<(packaging)(numvalthick um) thickness film-s->
<(packaging) film-s- thickness (numvalthick um)>
<thickness [prep] (packaging) film-s-(numvalthick um)>
<(numvalthick um) thickness (packaging) film-s->
<(numvalthick um) thickness [prep] film-s- (packaging)>
<(packaging)(numvaltemp °C) [conj/prep] (numvalrh%) RH>
<(numvaltemp °C) [conj/prep] (numvalrh%) RH (packaging)>
<(numvaltemp °C) [conj/prep] (numvalrh%) RH (packaging) film-s- thickness(numvalthick um)>
<(packaging) film-s- thickness(numvalthick um) (numvaltemp °C) [conj/prep] (numvalrh%) RH>
<thickness [prep] film-s- (numvalthick um) (packaging) (numvaltemp °C) [conj/prep] (numvalrh%) RH>
<(numvalthick um) thickness (packaging) film-s- (numvaltemp °C) [conj/prep] (numvalrh%) RH>

Tableau 4.11: Extrait de motifs étendus testés au cours des expérimentations.

Ces premiers motifs étendus ont été testés sur un échantillon de 11 articles dont 87 instances d'arguments ont été identifiées. Dans ce contexte d'évaluation de l'efficacité de la méthode hybride, nous privilégions la mesure de la précision pour l'extraction afin de limiter le bruit et faciliter l'analyse de la qualité de la méthode. Les premiers résultats d'évaluation sont restitués dans le tableau 4.12 selon les mesures de précision, de rappel et de F-mesure en fonction des différents types d'instances d'arguments recherchés et en fonction du type de motifs appliqués. Les motifs non enrichis correspondent aux motifs obtenus par l'approche de fouille de données et qui restituent les règles de corrélations. Les

Type d'évaluation	**Motifs non enrichis**			**Approche hybride**		
	Précision	*Rappel*	*F-mesure*	*Précision*	*Rappel*	*F-mesure*
Évaluation générale	0.5	0.8	0.6	0.7	0.8	0.6
Évaluation *packaging* et *thickness*	0.4	0.9	0.5	0.7	0.9	0.8
Évaluation *temperature* et *relative humidity*	0.3	0.9	0.4	0.8	0.7	0.7
Évaluation avec n > 2 arguments corrélés	0.6	0.6	0.6	0.7	0.6	0.6

Tableau 4.12: Évaluation de l'extraction d'arguments corrélés.

résultats obtenus avec ces premiers motifs, dont le niveau de restitution reste plutôt générique, montrent que leur capacité d'extraction est performante avec des rappels de 0.6 à 0.9 mais les résultats montrent également que les données extraites sont bruitées avec des taux de précisions faibles, en-dessous de 0.6. L'enrichissement des motifs en utilisant l'approche hybride permet d'améliorer les résultats d'extraction concernant essentiellement la précision (>0.7), ce qui facilite ainsi l'étape de validation et la qualité des instances extraites. Parmi ces instances correctement restituées, nous retrouvons dans 70% des cas les instances restituant l'emballage et son épaisseur extraites de manière correcte et dans 80% des cas, nous retrouvons toujours la température et l'humidité relative extraites de manière corrélée dans l'instance comme cela est suggéré par les règles restituées. Les instances sont extraites dans une fenêtre textuelle maximale $f_{\pm 1}$. Les résultats concernant les motifs étendus à 4 arguments corrélés sont plutôt encourageants sachant que pour ce processus d'extraction, nous travaillons avec un nombre très restreint de motifs étendus (environ une dizaine pour les 4 arguments corrélés). La RTO du domaine des emballages contient notamment trois relations n-aires comportant 6 arguments associés, les deux derniers arguments restant à extraire sont la pression partielle, *partial pressure*, et la perméabilité de l'emballage, *permeability*. À ce stade des travaux, nous n'avons pas mené d'expérimentations concernant ces deux derniers arguments. Néanmoins, nous avons déjà découvert, dans l'étape de fouille de données, que plusieurs MS montrent que ces arguments s'expriment fréquemment ensemble dans une fenêtre textuelle f_0. Nous avons également découvert d'autres motifs fréquents, e.g. *<(packaging)(quantity)(permeability)> (s=0.57), <(packaging)(permeability)> (s=0.64)*, dans une fenêtre textuelle maximale de $f_{\pm 1}$, qui expriment que l'expression de l'emballage est fréquemment suivie de l'expression de la perméabilité. Cette observation suggère que l'argument *packaging* pourrait jouer un rôle fondamental dans la mise en relation de tous les arguments des trois relations n-aires étudiées dans le domaine des emballages. En effet, il se retrouve à la fois dans le motif étendu aux 4 premiers arguments de la relation n-aire et dans ce dernier motif à étendre. Ces conclusions restent à confirmer en contrôlant la précision de l'extraction du dernier motif étendu et en contrôlant la cohérence de la relation n-aire complète. De

plus, la validation d'autres motifs obtenus au cours de l'approche de fouille de données, à partir de différentes fenêtres textuelles et à combiner selon l'approche hybride, permettra de produire plus de motifs étendus pertinents. Ces motifs étendus pertinents permettront de produire une meilleure couverture sur l'extraction, notamment pour les arguments corrélés avec n $>$ 2.

4.3.5 Conclusion

Dans cette section, nous avons proposé de nouvelles pistes de travail pour contribuer à l'extraction des instances d'arguments des relations n-aires dispersés dans les documents. Notre méthode hybride permet de découvrir des relations implicites entre arguments de la relation n-aire à partir des MS restitués. L'une des originalités de l'approche d'extraction des MS est de ne plus considérer l'expression des arguments de la relation n-aire de manière indépendante dans les textes mais bien comme partageant des relations implicites. L'analyse de relations syntaxiques permet de proposer une nouvelle fonction de rang et d'étendre les motifs séquentiels découverts pour faciliter l'extraction des instances dans les textes. La méthode a été évaluée sur un nombre restreint de motifs étendus mais les premiers résultats montrent des perspectives encourageantes pour l'extraction des instances.
L'architecture globale, présentée dans la figure 4.3, montre qu'elle reste générique pour s'appliquer à un autre domaine dont les résultats expérimentaux sont modélisés en relations n-aires au sein d'une RTO de domaine. Le principe consiste à découvrir, au cours de la première étape par fouille de données, à partir de nouvelles représentations des données textuelles guidées par la RTO du domaine, les motifs séquentiels fréquents restituant des régularités dans l'expressivité des arguments impliqués dans la relation n-aire, puis, enrichir/étendre ces motifs d'informations syntaxiques ciblées, obtenues en analysant les RS candidates pertinentes au domaine, pour l'obtention des motifs étendus de domaine.

Chapitre 5

Conclusion

Les travaux présentés dans ce mémoire s'inscrivent dans une démarche d'extraction et de capitalisation des connaissances dans le domaine des sciences du vivant et de l'environnement. La communauté fait face à un nombre croissant de données, en particulier de données publiées sur le web, sous forme d'articles scientifiques. Ces articles comportent de l'information pertinente qui doit être extraite et structurée afin, d'une part, d'être capitalisée dans une Ressource Termino-Ontologique (RTO) de domaine et, d'autre part, de pouvoir la ré-utiliser dans des outils d'aide à la décision. La RTO a été modélisée pour représenter les données pertinentes, dans un format structuré, sous forme d'arguments associés dans la relations n-aire. Les données pertinentes représentées correspondent à des données expérimentales, associant un objet étudié à ses paramètres de contrôle mesurés au cours des différentes expérimentations menées dans le domaine étudié. Les instances des relations n-aires représentées dans la RTO, présentes dans les textes, sont extraites afin de peupler la RTO de domaine.
Dans ce mémoire, nous proposons une contribution méthodologique à l'extraction des relations n-aires dans les textes en s'appuyant sur leur représentation dans la RTO de domaine. Les arguments de la relation n-aire comportent des arguments symboliques et des arguments quantitatifs associés à leurs unités de mesure. La tâche d'extraction des relations n-aires est complexe car elle se confronte à plusieurs verrous, notamment la forte variation des unités de mesure dans les textes et la forte variabilité d'expression des arguments dans les documents. Les arguments peuvent se trouver dans des phrases différentes et décrits tout au long des articles scientifiques.
Nous avons proposé une méthode en deux étapes pour l'extraction semi-automatique des variants d'unités de mesure dont les variations typographiques constituent un premier verrou à l'extraction des données quantitatives dans les textes. Notre méthode propose une première contribution qui s'appuie sur l'apprentissage supervisé afin de localiser efficacement les variants d'unités dans les textes en réduisant l'espace de recherche. Cette

première étape a permis de réduire de 86% l'espace de recherche dans le corpus des emballages et de 90% dans le corpus de bioraffinerie. Une des originalités de cette première étape est de ré-utiliser le même modèle appris sur les deux corpus, confirmant ainsi que l'unité de mesure correspond à un descripteur pertinent, comportant une information lexicale et sémantique fondamentales pour les données quantitatives de la relation n-aire. La méthode propose une deuxième contribution s'appuyant sur les particularités des unités de mesure afin de définir une nouvelle mesure de similarité, SM_{Db}, plus efficace pour identifier les variants localisés en les comparant à des unités de mesure déjà référencées dans la RTO. SM_{Db} propose des couples composés d'un variant d'unité et de son référent le plus pertinent dans la RTO. La validation de ces couples par les experts montre que SM_{Db} atteint un taux de précision allant jusqu'à 88%. Les variants validés enrichissent la partie terminologique de la RTO.
Dans la suite, nous avons proposé une méthode fondée sur la fouille de données, notamment les motifs et règles séquentiels afin de découvrir des régularités dans l'expressivité des instances d'arguments engagés dans la relation n-aire. Pour cela, nous avons proposé de nouvelles représentations des données textuelles à partir de nouvelles définitions fondées sur un descripteur pertinent, l'unité de mesure, et en s'appuyant sur la RTO de domaine afin de découvrir des cooccurrences entre plusieurs arguments de la relation n-aire.
Nous avons alors proposé, en s'appuyant sur la découverte de cooccurrences d'arguments dans les textes, un travail de recherche des relations syntaxiques pertinentes au domaine pour enrichir et étendre les motifs découverts. Les motifs étendus obtenus contribuent à l'extraction des instances d'arguments corrélés avec n arguments ($n \geq 2$). Les premières évaluations de la méthode hybride, testée avec un nombre restreint de motifs étendus sur le corpus du domaine des emballages, montrent des résultats encourageants avec des extractions d'instances comportant jusqu'à 4 arguments de la relation n-aire.

Perspectives.
Les travaux menés ont fait émerger plusieurs nouvelles questions à envisager comme perspectives à court et moyen terme :
La première concerne l'étape de fouille de données, notamment les approches à intégrer dans la méthode proposée afin de restreindre le nombre de motifs et règles séquentiels restitués à un ensemble strictement pertinent au domaine. Nous avons déjà réduit les motifs et règles restitués en proposant des représentations guidées par la RTO de domaine, mais la question se porte plus à présent sur l'existence éventuelle de contraintes de domaine sur les données, e.g. les valeurs numériques associées aux différents arguments quantitatifs de la relation n-aire, à discuter avec les experts. Nous pourrions également envisager d'apprendre ces contraintes de domaine. Toutefois, dans le cadre de cette perspective, il

serait nécessaire de produire un voire plusieurs corpus adaptés à ce type d'apprentissage. Le corpus n'intègrerait que les données d'intérêt à évaluer et prendrait en compte les conversions des unités de mesure fréquentes dans ce contexte, e.g. un changement de la valeur restituée si l'argument est exprimé dans une autre échelle de mesure, comme μm ou mm pour l'épaisseur.

Le deuxième concerne la contextualisation des motifs étendus à partir des relations syntaxiques spécifiques restituées par la fonction *n-ary ranking*. La notion de contextualisation des motifs séquentiels est abordée dans les travaux de (Rabatel and Bringay, 2011) qui proposent d'intégrer de nouvelles informations relatives au contexte de découverte des motifs séquentiels classiques. Cette notion peut être abordée de deux manières différentes dans le cadre de nos travaux. La première dans la poursuite de nos travaux actuels, en intégrant *a posteriori* dans les motifs, à partir de l'analyse des RS spécifiques qui intègrent notamment des verbes de type expérimental descriptif ou analytique ou encore des adjectifs liés à la mesure de certains arguments de la relation n-aire. Ces informations syntaxiques seraient intéressantes à intégrer dans l'extension des motifs afin de proposer des motifs étendus contextualisés. Ces motifs étendus contextualisés seraient utilisés pour l'extraction des instances d'arguments corrélés dans le contexte de certains paragraphes, e.g. motifs étendus de verbes de type descriptif utilisés dans le contexte du paragraphe de la description du matériel et des méthodes, les motifs étendus des verbes de type expérimental analytique utilisés dans le paragraphe des résultats. Les motifs étendus avec certains adjectifs comparatifs, e.g. *thinner, thicker*, pourraient être utilisés dans le contexte du paragraphe de discussion des résultats. La seconde proposerait d'intégrer dans les nouvelles représentations, la notion de paragraphe, i.e. *Abstract, Introduction, Materials and Methods, Results, Discussion, Conclusion*, comme information additionnelle. Dans ce cas, l'interprétation des motifs découverts concernant l'expressivité des instances d'arguments de la relation n-aire se ferait dans le contexte d'un paragraphe précis.

La troisième nous semble importante à aborder et est liée à une observation relevée au cours du processus d'extraction des instances d'arguments corrélés. Elle concerne la distinction à faire entre des instances de nouveaux résultats proposés dans le cadre de l'article, et les références à des résultats obtenus au cours de travaux précédents. Du moment qu'il s'agit d'une instance à extraire respectant la définition de la relation n-aire, la méthode permet l'extraction de tout type de résultat cohérent avec la définition dont les références. La question porte alors sur de nouvelles approches à envisager pour proposer de faire la distinction entre instances. Dans ce cas, là encore, les motifs étendus avec de l'information contextuelle seraient peut être une réponse à ce type de nouveau verrou.

Enfin, une dernière perspective viserait à proposer les corpus dans le cadre d'une campagne d'évaluation, comme cela est déjà largement utilisé dans le domaine biomédical et résumé par (Zhou et al., 2014) dans le tableau 5.1. Les challenges mis en place impliquent de proposer un ou plusieurs corpus dont l'annotation permet de tester et comparer différentes approches issues de différentes communautés dans un cadre d'évaluation défini. Dans ce cadre d'évaluation, les extractions sont maîtrisées puisque le nombre d'entités, le type d'entités et le nombre et le type de relations à extraire sont préalablement identifiés, facilitant les étapes d'évaluation et de validation des résultats par les experts.

Corpus name	General description	URL
GENIA	2,000 MEDLINE abstracts with more than 400,000 words and almost 100,000 annotations for biological terms.	http://www.nactem.ac.uk/genia/genia-corpus
LLL05	80 sentences in the training set including 106 examples of genic interactions without coreferences and 165 examples of interactions with coreferences.	http://genome.jouy.inra.fr/texte/LLLchallenge/
BioCreAtIvE II	Training data is derived from the content of the IntAct and MINT databases. The test set collection consists of a collection of PubMed article abstracts.	http://www.biocreative.org
AIMed	225 MEDLINE abstracts (200 abstracts describing interactions between human proteins and around 1000 tagged interactions).	ftp://ftp.cs.utexas.edu/pub/mooney/bio-data
BioInfer	1100 sentences annotated with protein names, their relationships, and PPI annotations.	http://mars.cs.utu.fi/BioInfer/
HPRD50	50 abstracts referenced by the Human Protein Reference Database including 266 relation instances.	http://www.hprd.org

FIGURE 5.1: Corpus annotés disponibles dans le domaine biomédical (Zhou et al., 2014).

Bibliographie

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Agrawal, R. and Srikant, R. (1995). Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, ICDE '95, pages 3–14, Washington, DC, USA. IEEE Computer Society.

Akhondi, S., Hettne, K., van der Horst, E., van Mulligen, E., and Kors, J. (2015). Recognition of chemical entities : combining dictionary-based and grammar-based approaches. *Journal of Cheminformatics*, 7(Suppl 1) :S10.

Arnaud, E., Cooper, L., Shrestha, R., Menda, N., Nelson, R. T., Matteis, L., Skofic, M., Bastow, R., Jaiswal, P., Mueller, L. A., and McLaren, G. (2012). Towards a Reference Plant Trait Ontology for Modeling Knowledge of Plant Traits and Phenotypes. In Filipe, J. and Dietz, J. L. G., editors, *KEOD*, pages 220–225. SciTePress.

Aussenac-Gilles, N., Condamines, A., and Sèdes, F., editors (2006). *Ressources termino-ontologiques*, volume Hors-série of *Information - Interaction - Intelligence*. Cépaduès Editions, http ://www.cepadues.com/.

Bach, N. and Badaskar, S. (2007). A Review of Relation Extraction.

Bachimont, B. (2000). Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. *Ingénierie des Connaissances : Evolutions récentes et nouveaux défis*, 1 :1–16.

Béchet, N., Cellier, P., Charnois, T., and Crémilleux, B. (2012). Discovering linguistic patterns using sequence mining. In *Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I*, CICLing'12, pages 154–165, Berlin, Heidelberg. Springer-Verlag.

Bendadouche, R., Roussey, C., Sousa, G. D., Chanet, J.-P., and Hou, K. M. (2012). Extension of the Semantic Sensor Network Ontology for Wireless Sensor Networks : The Stimulus-WSNnode-Communication Pattern. In Henson, C. A., Taylor, K., and Corcho, Ó., editors, *SSN*, volume 904 of *CEUR Workshop Proceedings*, pages 49–64. CEUR-WS.org.

Berrahou, L., Buche, P., Dibie, J., and Roche, M. (2013a). Unit of measure identification in unstructured scientific documents in microbial risk in food. In *ICPMF 2013 - International Conference on Predictive Modeling in Food*, pages 254–255.

Berrahou, S., Lebras, L., Buche, P., Dibie-Barthélemy, J., and Roche, M. (2014). Enrichissement d'une RTO par l'ajout de termes spécialisés. *Numéro spécial de RNTI (Revue des Nouvelles Technologies de l'Information) "Des Sources Ouvertes au Web de Données", Version étendue du résumé/poster d'EGC'13*, RNTI-W-2 :59–74.

Berrahou, S. L., Buche, P., Dibie, J., and Roche, M. (2015). Identification des unités de mesure dans les textes scientifiques. *TALN à paraitre*.

Berrahou, S. L., Buche, P., Dibie-Barthélemy, J., and Roche, M. (2013b). How to extract unit of measure in scientific documents ? In *KDIR/KMIS 2013 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval and the International Conference on Knowledge Management and Information Sharing, Vilamoura, Algarve, Portugal, 19 - 22 September, 2013*, pages 249–256.

Bossy, R., Jourde, J., Manine, A.-P., Veber, P., Alphonse, É., van de Guchte, M., Bessières, P., and Nedellec, C. (2012). Bionlp shared task - the bacteria track. *BMC Bioinformatics*, 13(S-11) :S3.

Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, ANLC '92, pages 152–155, Stroudsburg, PA, USA. Association for Computational Linguistics.

Buche, P., Dervaux, S., Dibie-Barthélemy, J., Ibanescu, L., Soler, L., and Touhami, R. (2013a). Vers la publication d'une RTO dédiée à l'annotation de relations n-aires.

Buche, P., Dervaux, S., Dibie-Barthélemy, J., Soler, L., Ibanescu, L., and Touhami, R. (2013b). Intégration de données hétérogènes et imprécises guidée par une ressource termino-ontologique. *Revue d'Intelligence Artificielle*, 27(4-5) :539–568.

Buche, P., Dibie-Barthélemy, J., Ibanescu, L., and Soler, L. (2013c). Fuzzy Web Data Tables Integration Guided by an Ontological and Terminological Resource. *IEEE Trans. Knowl. Data Eng.*, 25(4) :805–819.

Caracciolo, C., Stellato, A., Rajbhandari, S., Morshed, A., Johannsen, G., Keizer, J., and Jaques, Y. (2012). Thesaurus maintenance, alignment and publication as linked data : the AGROVOC use case. *IJMSO*, 7(1) :65–75.

Cellier, P., Charnois, T., Plantevit, M., Rigotti, C., Crémilleux, B., Gandrillon, O., Kléma, J., and Manguin, J. (2015). Sequential pattern mining for discovering gene interactions and their contextual information from biomedical texts. *J. Biomedical Semantics*, 6 :27.

Charnois, T., Plantevit, M., Rigotti, C., and Cremilleux, B. (2009). Fouille de données séquentielles pour l'extraction d'information dans les textes. *Revue Traitement Automatique des Langues (TAL)*, 50(3) :59–87.

Claveau, V. (2012). Vectorisation, Okapi et calcul de similarité pour le TAL : pour oublier enfin le TF-IDF. In *TALN - Traitement Automatique des Langues Naturelles*, pages –, Grenoble, France.

Cohen, W. W., Ravikumar, P., and Fienberg, S. E. (2003). A comparison of string distance metrics for name-matching tasks. In *Proc. of the Workshop on Information Integration on the Web (IIWeb-03)*, volume 47.

Compton, M., Barnaghi, P. M., Bermudez, L., Garcia-Castro, R., Corcho, Ó., Cox, S., Graybeal, J., Hauswirth, M., Henson, C. A., Herzog, A., Huang, V. A., Janowicz, K., Kelsey, W. D., Phuoc, D. L., Lefort, L., Leggieri, M., Neuhaus, H., Nikolov, A., Page, K. R., Passant, A., Sheth, A. P., and Taylor, K. (2012). The SSN ontology of the W3C semantic sensor network incubator group. *J. Web Sem.*, 17 :25–32.

Damerau, F. (1964). A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3) :171–176.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *IN PROC. INT'L CONF. ON LANGUAGE RESOURCES AND EVALUATION (LREC*, pages 449–454.

Fabrègue, M., Braud, A., Bringay, S., Ber, F. L., and Teisseire, M. (2012). Extraction de motifs spatio-temporels à différentes échelles avec gestion de relations spatiales qualitatives. In *Actes du XXXème Congrès INFORSID, Montpellier, France, 29 - 31 mai 2012*, pages 123–140.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors (1996). *Advances in Knowledge Discovery and Data Mining*. American Association for Artificial Intelligence, Menlo Park, CA, USA.

Fournier-Viger, P., Faghihi, U., Nkambou, R., and Nguifo, E. M. (2012). Cmrules : Mining sequential rules common to several sequences. *Knowl.-Based Syst.*, pages 63–76.

Fridman Noy, N. and Rector, A. (2006). Defining N-ary Relations on the Semantic Web.

Ghersedine, A., Buche, P., Dibie-Barthélemy, J., Hernandez, N., and Kamel, M. (2012). Extraction de relations n-aires interphrastiques guidée par une RTO. In *actes de la 9ième édition de la COnférence en Recherche d'Information et Applications, CORIA*, pages 179–190.

Grau, B., Ligozat, A.-L., and Minard, A.-L. (2009). Corpus study of kidney-related experimental data in scientific papers. In *Proceedings of the Workshop on Biomedical Information Extraction*, pages 21–26.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *KNOWLEDGE ACQUISITION*, 5 :199–220.

Guarino, N., Oberle, D., and Staab, S. (2009). What is an ontology ? In Staab, S. and Studer, R., editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 1–17. Springer Berlin Heidelberg.

Guillard, V., Buche, P., Destercke, S., Tamani, N., Croitoru, M., Menut, L., Guillaume, C., and Gontard, N. (2015). A Decision Support System to design modified atmosphere packaging for fresh produce based on a bipolar flexible querying approach. *CEA à paraitre*.

Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U., and Hsu, M.-C. (2000). Freespan : Frequent pattern-projected sequential pattern mining. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, pages 355–359, New York, NY, USA. ACM.

Hawizy, L., Jessop, D., Adams, N., and Murray-Rust, P. (2011). ChemicalTagger : a tool for semantic text-mining in chemistry. *Journal of cheminformatics*, 3(1) :17.

Jaillet, S., Laurent, A., and Teisseire, M. (2006). Sequential patterns for text categorization. *Intell. Data Anal.*, 10(3) :199–214.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406) :414–420.

Jessop, D. M., Adams, S. E., and Murray-Rust, P. (2011a). Mining chemical information from open patents. *Journal of cheminformatics*, 3(1) :40.

Jessop, D. M., Adams, S. E., Willighagen, E. L., Hawizy, L., and Murray-Rust, P. (2011b). OSCAR4 : a flexible architecture for chemical text-mining. *Journal of cheminformatics*, 3(1) :1–12.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proc. of the conf. on Uncertainty in artificial intelligence*, pages 338–345.

Jones, K. S., Walker, S., and Robertson, S. E. (2000). A probabilistic model of information retrieval : development and comparative experiments - part 1. *Inf. Process. Manage.*, 36(6) :779–808.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kohavi, R. and Quinlan, J. R. (2002). Data mining tasks and methods : Classification : decision-tree discovery. In *Handbook of data mining and knowledge discovery*, pages 267–276. Oxford University Press, Inc.

Levenshtein, V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10 :707.

Lin and Dekang (1998). Dependency-based evaluation of minipar. In *Proc. Workshop on the Evaluation of Parsing Systems*, Granada.

Ludovic, J.-L. (2011). *Approches supervisées et faiblement supervisées pour l'extraction d'événements et le peuplement de bases de connaissances*. PhD thesis, Université Paris Sud.

Maedche, A. and Staab, S. (2002). Measuring similarity between ontologies. In *Knowledge Engineering and Knowledge Management : Ontologies and the Semantic Web*, volume 2473 of *LNCS*, pages 251–263. Springer.

Masseglia, F., Cathala, F., and Poncelet, P. (1998). The psp approach for mining sequential patterns. In Zytkow, J. M. and Quafafou, M., editors, *PKDD*, volume 1510 of *Lecture Notes in Computer Science*, pages 176–184. Springer.

Mcdonald, R., Pereira, F., Kulick, S., Winters, S., Jin, Y., and White, P. (2005). Simple algorithms for complex relation extraction with applications to biomedical ie. In *In Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL-05*, pages 491–498.

Minard, A., Grau, B., Ligozat, A., and Thomas, S. R. (2013). Extraction de relations complexes. application à des résultats expérimentaux en physiologie rénale. *Technique et Science Informatiques*, 32(1) :77–111.

Monge, A. E. and Elkan, C. (1996). The field matching problem : Algorithms and applications. In *Proceedings of the second international Conference on Knowledge Discovery and Data Mining*, pages 267–270.

Nguyen, V. T., Gaio, M., and Sallaberry, C. (2010). Recherche de relations spatio-temporelles : une méthode basée sur l'analyse de corpus textuels. *CoRR*, abs/1002.0577.

Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., and Hsu, M. (2001). Prefixspan : Mining sequential patterns by prefix-projected growth. In *Proceedings of the 17th International Conference on Data Engineering*, pages 215–224, Washington, DC, USA. IEEE Computer Society.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods*, pages 185–208. MIT Press.

Quinlan, J. R. (1979). Discovering rules by induction from large collections of examples. In Michie, D., editor, *Expert Systems in the Micro-Electronic Age*, pages 168–201. Edinburgh University Press, Edinburgh.

Quinlan, J. R. (1993). *C4.5 : Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Rabatel, J. and Bringay, S. (2011). Extraction de motifs séquentiels contextuels. In *Extraction et gestion des connaissances (EGC'2011), Actes, 25 au 29 janvier 2011, Brest, France*, pages 11–22.

Reymonet, A., Thomas, J., and Aussenac-Gilles, N. (2007). Modelling ontological and terminological resources in owl dl. In Buitelaar, P., Choi, K.-S., Gangemi, A., Huang, C.-R., and Oltramari, A., editors, *From Text to Knowledge : The Lexicon/Ontology Interface - Workshop at ISWC07 [U+0096] 6th International Semantic Web Conference (OntoLex), Busan (South Korea), 11/11/2007*, page (on line), http ://www.pusan.ac.kr. University of Pusan.

Rijgersberg, H., van Assem, M., and Top, J. (2013). Ontology of units of measure and related concepts. *Semantic Web*.

Rijgersberg, H., Wigham, M., and Top, J. L. (2011). How semantics can improve engineering processes : A case of units of measure and quantities. *Advanced Engineering Informatics*, 25(2) :276–287.

Salton, G., Allan, J., and Buckley, C. (1993). Approaches to passage retrieval in full text information systems. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, pages 49–58, New York, NY, USA. ACM.

Sleator, D. D. and Temperley, D. (1995). Parsing english with a link grammar. *CoRR*, abs/cmp-lg/9508004.

Srikant, R. and Agrawal, R. (1996). Mining sequential patterns : Generalizations and performance improvements. In *Proceedings of the 5th International Conference on Extending Database Technology : Advances in Database Technology*, EDBT '96, pages 3–17, London, UK, UK. Springer-Verlag.

Su, J., Zhang, H., Ling, C. X., and Matwin, S. (2008). Discriminative parameter learning for bayesian networks. In *Proc. of the int. conf. on Machine learning*, pages 1016–1023.

Thompson, A. and Taylor, B. N. (2008). Guide for the use of the international system of units (SI).

Touhami, R., Buche, P., Dibie-Barthélemy, J., and Ibanescu, L. (2011). An ontological and terminological resource for n-ary relation annotation in web data tables. *On the Move to Meaningful Internet Systems : OTM 2011*, pages 662–679.

Van Assem, M., Rijgersberg, H., Wigham, M., and Top, J. (2010). Converting and annotating quantitative data tables. *The Semantic Web–ISWC 2010*, pages 16–31.

Willems, D. J., Rijgersberg, H., and Top, J. (2012). Identifying and extracting quantitative data in annotated text. *SWAIE*.

Winkler, W. E. (1999). The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Census Bureau.

Yan, X., Han, J., and Afshar, R. (2003). Clospan : Mining closed sequential patterns in large databases. In Barbará, D. and Kamath, C., editors, *SDM*. SIAM.

Zaki, M. J. (2001). Spade : An efficient algorithm for mining frequent sequences. *Mach. Learn.*, 42(1-2) :31–60.

Zhou, D., Zhong, D., and He, Y. (2014). Biomedical relation extraction : From binary to complex. *Comp. Math. Methods in Medicine*, 2014.

Zweigenbaum, P. (2004). L'umls entre langue et ontologie : une approche pragmatique dans le domaine médical. *Revue d'Intelligence Artificielle*, 18(1) :111–137.