Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2009

Annecy – 4 & 5 juin 2009

Publications précédentes

TOTh 2007

Actes de la première conférence TOTh - Annecy - 1er juin 2007

TOTh 2008

Actes de la deuxième conférence TOTh - Annecy – 5 et 6 juin 2008

Commandes à adresser à : toth@porphyre.org

Titre : TOTh 2009. *Actes de la troisième conférence TOTh - Annecy – 4 & 5 juin 2009*
Editeur : Institut Porphyre, *Savoir et Connaissance*
http://www.porphyre.org
Annecy, 2009
ISBN 978-2-9536168-0-4
EAN 9782953616804
© Institut Porphyre, *Savoir et Connaissance*

Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2009

Annecy – 4 & 5 juin 2009

avec le soutien de :

- Société française de terminologie
- Association Européene de Terminologie
- Ecole d'ingénieurs Polytech'Savoie – Université de Savoie
- Université de Sorbonne nouvelle
- Association EGC (Extraction et Gestion des Connaissances)
- ISKO (International Society for Knowledge Organization) France

Institut Porphyre
Savoir et Connaissance

http://www.porphyre.org

Comité scientifique

Président du Comité Scientifique : Christophe Roche

Comité de pilotage

Loïc Depecker	Professeur, Université de Sorbonne nouvelle
André Manificat	Directeur, GRETh
Christophe Roche	Professeur, Université de Savoie
Philippe Thoiron	Professeur émérite, Université de Lyon II

Comité de programme

Bruno de Bessé	Professeur, Université de Genève
Pierre Blanc	EDF SEPTEN
Danièle Bourcier	CNRS, CERSA Paris
Marc van Campenhoudt	Professeur, Termisti, ISTI, Bruxelles
Danielle Candel	CNRS, Université Paris Diderot
Stéphane Chaudiron	Professeur, Université de Lille III
Viviane Cohen	France Télécom, Paris
Rute Costa	Professeur, Université Nouvelle de Lisbonne
Luc Damas	MCF, Université de Savoie
Sylvie Desprès	MCF, Université Paris XIII
François Gaudin	Professeur, Université de Rouen
Anne-Marie Gendron	Chancellerie fédérale suisse, Section de terminologie
Jean-Yves Gresser	ancien Directeur à la Banque de France
Ollivier Haemmerlé	Professeur, Université de Toulouse
Jean-Paul Haton	Professeur, Université de Nancy 1
Michèle Hudon	Professeur, Université de Montréal
John Humbley	Professeur, Université Paris 7
Michel Ida	Directeur MINATEC, CEA
Hendrik Kockaert	Professeur, Lessius Hogeschool (Anvers)
Michel Léonard	Professeur, Université de Genève
Pierre Lerat	Professeur honoraire, Université Paris XIII
Widad Mustafa	Professeur, Université de Lille III
Henrik Nilsson	Terminologicentrum TNC, Suède
Jean Quirion	Professeur, Université du Québec en Outaouais
Renato Reinau	Suva, Lucerne
François Rousselot	MCF, Université de Strasbourg
Gérard Sabah	CNRS, Orsay
Michel Simonet	CNRS Grenoble
Marcus Spies	Professeur, Université de Munich
Dardo de Vecchi	Professeur associé, Euro-Med Management

Comité d'organisation :

Responsable : Luc Damas
Samia Chouder, Joëlle Pellet

Avant propos

Dès la troisième édition, les conférences TOTh ont trouvé une structuration qui traduit bien à la fois le caractère scientifique et pluridisciplinaire de la terminologie et l'intérêt de notre communauté pour d'autres domaines partageant des préoccupations communes.

Ainsi, la conférence d'ouverture a été donnée par une personnalité invitée issue d'une discipline différente de la nôtre – ici la phylogénèse – mais pour laquelle le langage et la pensée jouent également un rôle primordial.

Les contributions se sont réparties naturellement, et par le jeu des évaluations de façon équitable, en trois groupes ayant donné lieu à trois sessions.

Le premier groupe a rassemblé les articles portant principalement sur la dimension linguistique de la terminologie. Ont été abordés l'extraction terminologique à partir de dictionnaire, la place accordée aux corpus dans la construction de terminologies, l'acquisition de connaissances à partir de textes et l'apport des ressources linguistiques issues du web.

La deuxième session s'est donc logiquement intéressée à la dimension conceptuelle de la terminologie. Les notions de concept, de relation, d'ontologie ont été au cœur des présentations portant sur les cartes conceptuelles pour les bibliothèques numériques, les relations dynamiques et les graphes conceptuels, l'alignement d'ontologies et l'accès multilingue aux ontologies.

Enfin, la troisième session a été consacrée à la présentation de plusieurs applications terminologiques pour des secteurs aussi différents que l'ingénierie nucléaire, l'informatique, le domaine bancaire ou l'agriculture biologique. Il est à souligner que ces applications ont permis d'aborder différents points théoriques tels que la variation terminologique, la diachronie ou la structure des dictionnaires.

La richesse des débats qui ont animé ces deux jours de conférence – chaque présentation, questions comprises, s'est vue allouer plus de quarante cinq minutes de temps de parole – a été certainement une des plus belles récompenses pour les participants de TOTh 2009.

Christophe Roche

Président du Comité Scientifique

Table des matières

CONFERENCE INVITEE

La nomenclature biologique aujourd'hui : que reste-t-il de Linné ? 1
Michel Laurin

SESSION 1

Approche lexico-sémantique de l'extraction terminologique : utilisation de ressources lexicographiques et validation sur corpus 19
Bertrand Gaiffe, Evelyne Jacquey, Laurence Kister

Quelle place accorder aux corpus dans la construction d'une terminologie ? 33
Marie Calberg-Challot, Pierre Lerat, Christophe Roche

Extraction de connaissances orientées évolution dans les textes techniques 53
Kata Gabor, François Rousselot, François De Bertrand de Beuvron

Corpus et Web : deux alliés pour la construction de l'enrichissement automatique de classes conceptuelles 73
Nicolas Béchet, Mathieu Roche, Jacques Chauché

SESSION 2

Following the path between conceptual maps and visual thesauri 93
Olga Bessa Mendes

Dynamic concept relations: a definition and representation proposal 107
Chiara Messina

Construction et alignement d'ontologies pour évaluer le risque alimentaire 127
Liliana Ibanescu, Patrice Buche, Juliette Dibie-Barthélemy

Accès multilingue à une ontologie par des correspondences avec un lexique pivot 143
David Rouquet, Hong-Thai Nguyen

La reformulation : processus dynamique d'acquisition des connaissances. Le cas du discours technique arabe d'Internet 161

Andrée Affeich

SESSION 3

Structuration d'un dictionnaire de spécialité pour sa publication sur internet. Bénéfices du langage XML 181
Jacques Joseph

Mémoire du Club informatique des grandes entreprises françaises (CIGREF) : nouveau plan de classement 197
Jean-Yves Gresser, M.P. Lacroix

Les secteurs d'activité à l'épreuve du discours 217
Frédéric Erlos

De l'agriculture biologique aux espaces naturels : une étude des syntagmes terminologiques à l'intérieur des textes de spécialité 235
Elisa Lavagnino

Pages blanches 253

CONFERENCE INVITEE

La nomenclature biologique aujourd'hui : que reste-t-il de Linné ?

Michel Laurin

Résumé : La nomenclature biologique est une activité essentielle étant donné que nous connaissons plus d'un million et demi d'espèces d'êtres vivants. La nomenclature Linnéenne-Stricklandienne, utilisée par la plupart des systématiciens depuis deux siècles, repose sur l'utilisation d'un seul type et d'un rang absolu (catégorie Linnéenne). Or, si le type a bien une existence objective, les catégories Linnéennes en sont dépourvues. De plus l'utilisation d'un seul type et d'un rang ne peut aucunement fournir une délimitation unique et stable du contenu des taxons. Cette absence de délimitation, apparemment recherchée par les commissions ayant rédigé les codes formalisant cette nomenclature, semble inappropriée puisque généralement, les scientifiques tentent de définir les termes techniques aussi précisément que possible afin de communiquer efficacement. Ceci est démontré par des comparaisons avec la géochronologie, la géopolitique et la chimie. Tout ceci explique le développement, à partir des années 1980, d'une nomenclature phylogénétique reposant sur l'utilisation d'au moins deux déterminants (types) et permettant de délimiter précisément les taxons à l'aide d'une phylogénie (arbre évolutif).

Mots-clés : Terminologie, taxonomie, systématique, codes de nomenclature

1. Introduction

La nomenclature biologique est une activité essentielle étant donné que nous connaissons plus d'un million et demi d'espèces d'êtres vivants. Il n'y a des spécialistes de ce domaine que depuis le 18ème siècle environ, Linné ayant été un des premiers auteurs à consacrer la plus grande partie de son activité à ce domaine. Cependant, divers auteurs du anciens s'y sont adonnés, dont Aristote est sans doute le plus connu et l'un des plus anciens (Voultsiadou & Vafidis 2007) et on retrouve des traces de cette activité dès la Genèse, qui fut écrite il y a presque 3000 ans. Par exemple, on peut y lire (Genèse 2 : 19-20) : "Le Seigneur Dieu modela du sol toute bête des champs et tout oiseau du ciel qu'il amena à l'homme pour voir comment il les désignerait". Il est donc clair que le besoin de classer la diversité naturelle s'est fait sentir très tôt dans l'histoire de l'humanité.

La nomenclature biologique comporte diverses particularités liées à la nature des êtres vivants et surtout à l'évolution biologique. En effet, la plupart des objets que nous classons forment justement des classes (d'où le verbe "classer", soit "ranger dans des classes"). Les entités formant des classes sont identiques, ils peuvent exister n'importe où dans l'univers, et ne sont pas temporellement délimités. Par exemple, un atome d'hydrogène ou d'hélium d'un isotope donné ne diffère aucunement des autres atomes du même élément et du même isotope. De tels atomes peuvent exister n'importe où dans l'univers. Ils peuvent avoir existé depuis environ 300 000 ans après le Big Bang (temps nécessaire pour que la matière refroidisse suffisamment pour que les atomes capturent leurs électrons) et peuvent exister indéfiniment, si l'univers est ouvert. Les êtres vivants, au contraire, sont spatio-temporellement délimités. Ainsi, les chats *(Felis domesticus)* et les humains *(Homo sapiens)* n'existent que sur terre, n'ont jamais existé où que ce soit ailleurs, et n'existeront jamais sur une autre planète sauf si nous les y transportons dans un futur lointain. Ils n'existent que depuis moins d'un million d'années et n'existeront pas éternellement. De plus, il n'y a pas deux chats (ou d'êtres humains) identiques. Même les jumeaux dits identiques ne le sont que de par leur génome, mais des différences apparaissent dans leur personnalité, leur système immunologique, etc. De tels êtres ne forment pas des classes, mais bien des individus, selon de nombreux auteurs (Ereshefsky 2007), même si certains préfèrent y voir un type d'entité intermédiaire entre des universaux formant des classes et des individus (Rieppel 2005). Depuis quelques décennies, on considère souvent que les taxons ne forment pas des classes, mais plutôt des individus, ce qui explique la difficulté de les définir d'après leurs propriétés intrinsèques.

On comprendra donc qu'en nomenclature biologique, l'étymologie s'explique, mais n'explique pas. En effet, les taxons sont souvent nommés d'après une propriété partagée par la plupart de ses membres mais cette propriété n'est ni nécessaire, ni suffisante pour déterminer l'appartenance d'un organisme à un taxon. Ainsi, le taxon Tetrapoda inclut des animaux ayant quatre membres, mais certains animaux dépourvus de membres, comme les serpents et les gymnophiones, en font aussi partie. Ils descendent d'ancêtres qui possédaient quatre membres et sont ainsi unanimement classés parmi les tétrapodes. De plus, si des animaux acquéraient quatre membres de façon convergente, ils ne seraient pas des tétrapodes. On voit ainsi qu'on ne peut généralement pas utiliser les attributs intrinsèques des organismes directement pour les attribuer à des taxons ; cette attribution dépend en fait plutôt de la position des organismes dans l'arbre évolutif du vivant. Dans ce fonctionnement même, on voit que la classification des êtres vivants opère de façon plus similaire à la généalogie, qui ne concerne que des individus et repose sur l'histoire, qu'à la classification d'universels, qui repose sur leurs propriétés intrinsèques.

L'arbre évolutif du vivant s'impose donc comme la base de toute classification du vivant. Mais alors, comment subdiviser cet arbre, qui forme un tout de sa racine au sommet de chaque branche ? Pour les systématiciens qui continuent à conceptualiser les taxons comme des classes, comment peuvent-ils définir et délimiter ces classes ? Si les taxons sont bien des individus, comment les définir et les délimiter ? Ces questions sont au cœur des débats actuels en nomenclature biologique, et nous verrons que leurs réponses diffèrent profondément entre la nomenclature Linnéenne-Stricklandienne et la nomenclature phylogénétique.

2. Caractéristiques principales de la nomenclature Linnéenne-Stricklandienne

En nomenclature Linnéenne-Stricklandienne, on attribue à chaque taxon un rang absolu ou catégorie Linnéenne. Les catégories principales vont, des plus grandes aux plus petites, du règne à l'espèce. Ainsi, la classification de notre espèce peut être représentée ainsi :

Règne : Metazoa
Embranchement : Chordata
Classe : Mammalia
Ordre : Primates
Famille : Hominidae
Genre : *Homo*
Espèce : Homo sapiens

On utilise en plus des préfixes pour augmenter le nombre de rangs disponibles. On peut ainsi former les noms de la superfamille Hominoidea et de la sous-famille Homininae. Notez que pour la série-famille, les terminaisons changent également ; pour les animaux, ces noms se terminent par "-oidea" pour les superfamilles, "-idae" pour les familles, et "-inae" pour les sous-familles.

Chaque taxon est ancré dans la réalité par un type, qui est soit un organisme préservé (squelette, spécimen préservé dans l'alcool, etc.), pour les espèces, soit une espèce, pour les genres, soit un genre, pour les familles. Ainsi, seules les espèces sont directement ancrées à la réalité par les spécimens. Au-dessus de ce rang, le lien est indirect, mais ultimement, même une famille est ancrée dans la réalité car elle est définie par un genre-type, qui est lui-même défini par une espèce-type, qui est elle-même définie par un spécimen-type. En zoologie, les taxons d'un rang supérieur à celui de la série-famille n'ont pas de types ; leur sens est donc un peu moins bien déterminé. Linné n'utilisait pas de types ; ils ont été introduits en nomenclature biologique par le code de Strickland (Strickland *et al.* 1842, 1843), d'où l'expression de nomenclature "Linnéenne-Stricklandienne" proposée par Dubois (2006).

Le principe de priorité, reconnu à divers degrés par tous les codes de nomenclature, stipule que généralement, le premier nom proposé pour un taxon est valide (ainsi que sa définition). Les autres sont soit des synonymes (d'autres noms pour le même taxon), soit des homonymes (le même nom défini autrement).

La définition des noms de taxons dans ce système consiste uniquement en un type et un rang (Laurin 2008a). On caractérise également les taxons par des diagnoses, qui sont des listes de caractères qui sont censées permettre de déterminer si un organisme appartient à un taxon. Ainsi, une diagnose des tétrapodes mentionnera la présence de quatre membres pourvus de doigts, l'absence de branchies, la présence de poumons, etc. Cependant, les diagnoses ne font pas partie des définitions car elles peuvent être modifiées et deux taxons ne peuvent pas être déclarés synonymes parce qu'ils partagent la même diagnose.

La nomenclature Linnéenne-Stricklandienne est régulée par trois codes principaux : le code zoologique pour les animaux, le code botanique pour les plantes et le code bactériologique pour les eubactéries et archées (Laurin 2005). Les limites entre les juridictions de ces codes sont parfois floues et artificielles. Ainsi, certaines bactéries (les cyanobactéries) sont photosynthétiques et sont souvent considérées comme des plantes ; leurs noms sont généralement régulées par le code botanique, même si elles sont plus étroitement apparentées aux autres bactéries qu'aux plantes vertes. De nombreux organismes eucaryotes ont acquis la capacité de photosynthèse par endosymbiose ; ils ont en quelque sorte incorporé des cyanobactéries dans leurs cellules, et ces bactéries sont devenues les chloroplastes. Certaines espèces de ces taxons ont ensuite perdu les chloroplastes. Traditionnellement, les noms de tous les organismes photosynthétiques et des champignons sont régis par le code botanique, alors que les noms de tous les autres eucaryotes sont régis par le code zoologique. Ceci signifie que dans de nombreux groupes d'organismes, les noms de certaines espèces sont régulés par le code zoologique, alors que d'autres espèces étroitement apparentées sont régulées par le code botanique. Pire, dans certains cas, une espèce donnée possède un nom régi par le code botanique, et un autre par le code zoologique !

3. Problèmes causés par la nomenclature Linnéenne-Stricklandienne

Les catégories attribuées aux taxons sont subjectives, car il n'y a ni famille, ni embranchement dans la nature. Ceci signifie que les systématiciens peuvent altérer ces rangs, ce qui résulte souvent en un changement de la délimitation du taxon. En effet, l'utilisation d'un seul type et d'un rang subjectif ne peut pas délimiter un taxon (Figure 1).

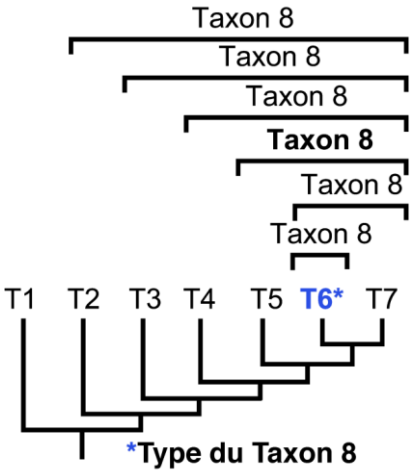


Figure 1. Délimitation floue en nomenclature Linnéenne-Stricklandienne

Par exemple, dans la Figure 1, si T1–T7 sont des genres et que T8 est une famille, la désignation de T6 comme type ne permet pas de délimiter T8. Plusieurs délimitations possibles sont indiquées sur la figure. Même si l'auteur ayant érigé T8 a spécifié qu'il souhaitait y inclure T5–T7 (délimitation en caractères gras), cette délimitation n'est pas contraignante ; les autres systématiciens ne sont pas tenus de la respecter, même s'ils n'ont aucune raison objective de redélimiter T8.

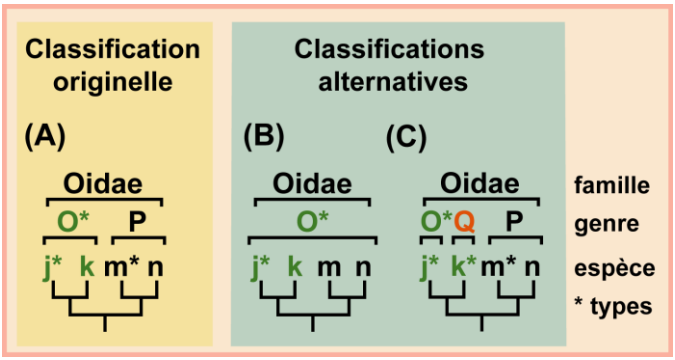


Figure 2. Synonymie et instabilité en nomenclature Linnéenne-Stricklandienne. Adapté de Laurin (2008b)

Cette délimitation floue est en partie liée à des synonymies potentielles de taxons étroitement apparentés, synonymie liée à la nature subjective des catégories Linnéennes (Lee et Skinner 2007). Ceci peut être illustré par un

exemple (Figure 2). Supposons qu'on découvre quatre nouvelles espèces (j, k, m et n) étroitement apparentées et qu'une phylogénie indique que ces espèces forment deux clades (Figure 2A ; le terme "clade" désigne un groupe d'organismes biologiques comprenant un ancêtre et tous ses descendants). Le systématicien qui décrira ces espèces pourrait alors choisir de reconnaître deux genres (O et P) et une nouvelle famille (Oidea). Les espèces j et m pourraient être désignées comme types des genres O et P et le genre O, comme type de Oidae. Sans aucune raison objective, un autre taxonomiste pourrait décider de mettre le genre P en synonymie avec O, étendant ainsi le contenu de O (Figure 2B). Un troisième taxonomiste pourrait décider, au contraire, de reconnaître un genre additionnel (Q) en choisissant l'espèce k comme type (Figure 2C), ce qui a pour résultat de réduire le contenu du genre O à sa seule espèce-type (j). D'autres considérations mènent également à considérer les catégories Linnéennes comme de dangereux reliquats d'un système de nomenclature périmé (Minelli 2000).

Puisque les codes Linnéens-Stricklandiens visent à ne pas délimiter les taxons précisément, toutes ces alternatives (et bien d'autres encore) sont simultanément et indéfiniment valables, ce qui signifie que le sens des noms est ambigu. Par exemple, le genre O peut contenir une seule espèce (j), deux (j, k), ou quatre (j, k, m et n), pour ne mentionner que ces trois possibilités. Ceci n'est pas qu'un problème purement théorique. Il affecte de nombreux taxons (peut-être la plupart) depuis avant même l'entrée en vigueur des codes de nomenclature qui ont justement été inaugurés pour réduire le chaos nomenclatural qui prévalait déjà dans les années 1840 (Strickland *et al.* 1842). Il semble malheureusement que les codes Linnéens-Stricklandiens n'ont eu, de ce point de vue, qu'un succès mitigé (Laurin 2008b). Ainsi, Rowe et Gauthier (1992) ont inventorié pas moins de 10 sens du nom Mammalia (taxon qui inclut les mammifères) dans la littérature scientifique publiée dans des années 1960 à 1990. On pourrait argumenter que la classe Mammalia n'a pas de type et que le code zoologique ne peut donc pas stabiliser son sens, mais de tels problèmes affectent tout autant les taxons des séries-famille, genre et espèce, qui ont pourtant des types. Ainsi, T. M. Keesey a récemment répertorié six sens du nom Hominidae (Laurin & Bryant 2009), et un débat a actuellement lieu concernant la classification des très nombreuses espèces (plus de 1000) du genre *Rana* (Frost *et al.* 2006 ; Hillis 2007).

4. Comparaisons avec d'autres domaines

La nomenclature Linnéenne-Stricklandienne semble relativement isolée des autres domaines de la connaissance humaine en tentant de ne pas définir précisément ses termes techniques. En fait, on pourrait argumenter que les noms de taxons tels qu'ils sont définis dans ce système ne sont pas véritablement des termes car ils ne correspondent pas à une délimitation stable de la réalité (Calberg-Challot *et al.* ce tome). Dans les autres domaines, on cherche généralement à préciser autant que possible les termes techniques.

Ainsi, la géochronologie utilisait autrefois des sections-types analogues aux types de la nomenclature Linnéenne-Stricklandienne. Ces sections-types définissaient des périodes plus ou moins longues, mais elles ne les délimitaient pas. Ainsi, la section-type du Lutétien ne représente qu'une petite proportion du temps représenté par ce dernier (Figure 3). Pendant les vingt dernières années, les géologues ont remplacé ces sections-types par des limites-types, appelées GSSPs, pour "Global Stratotype Section and Point" (Gradstein *et al.* 2004 : 20–21). Ces GSSPs se présentent sous forme d'un marqueur précis (par exemple, un clou enfoncé dans une section) identifiant précisément la limite inférieure d'une division temporelle (étage, période ou ère). La limite inférieure de la division suivante (plus récente) sert simultanément de limite supérieure à la division précédente. Ainsi, on est passé d'un système de nomenclature flou à un système très précis.

Les chimistes et les physiciens utilisent depuis longtemps une classification très précise de la matière fondée sur le nombre de protons dans le noyau atomique. Les éléments sont ensuite regroupés par familles chimiques en fonction du nombre d'électrons de valence, ce qui donne le fameux tableau des éléments initialement proposé par le Chimiste Russe Mendeleev en 1869. Ainsi, le lithium (Li) comporte trois protons et un électron de valence. Le sodium (Na) et le potassium (K) possèdent 11 et 19 protons et un seul électron de valence, ce qui explique qu'on les réunisse dans la même famille que le lithium. Les adeptes de la nomenclature Linnéenne-Stricklandienne devraient peut-être tenter de convaincre les chimistes et les physiciens d'un système de nomenclature dans lequel "lithium" désigne tantôt les atomes comportant 3 protons, tantôt ceux comportant 3 ou 11 protons, et tantôt ceux comportant 3, 11, ou 19 protons et dans lequel tous ces sens sont simultanément et indéfiniment valides. Un consensus sur le sens de ce terme devrait alors émerger (ou pas ?) par un accord spontané (et peut-être temporaire).

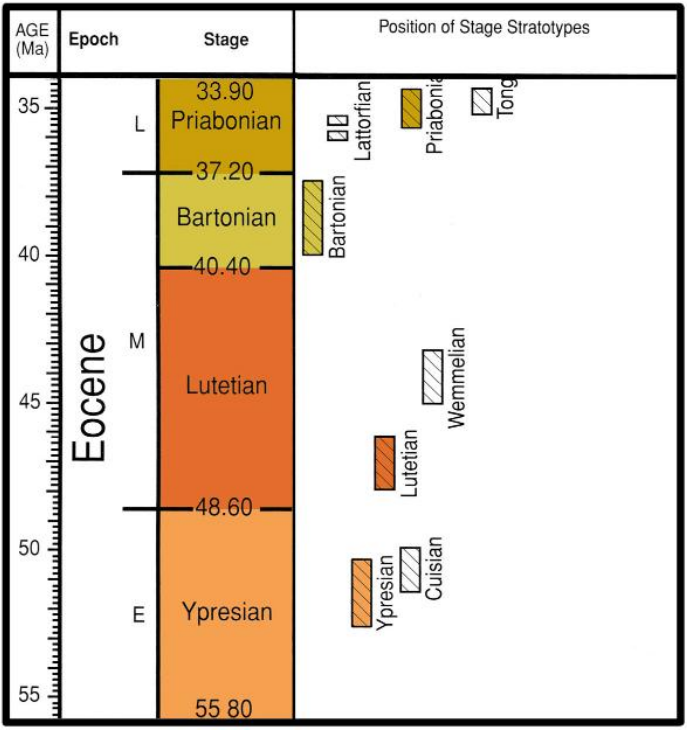


Figure 3. Nomenclature géochronologique. Adapté de Gradstein *et al.* (2004)

Les entités géographiques, comme les taxons, possèdent des rangs (continent, pays, province ou état, ville, quartier, etc.). On peut même considérer que les entités territoriales ont des types (chefs-lieux, capitales). Mais quel géographe ou politicien se contenterait de définir les entités territoriales par un rang et un chef-lieu sans aussi spécifier des frontières ou limites ? Imaginez la carte de l'Europe si la France était définie comme l'entité territoriale de rang Pays ayant Paris pour Capitale, et si tous les autres pays étaient définis de manière similaire. Évidemment, les géographes et les politiciens délimitent les entités territoriales depuis longtemps de façon précise en utilisant souvent soit des frontières naturelles (chaînes de montagnes, fleuves, côtes, etc.), soit des bornes placées aux frontières, comme c'était le cas dans la Mésopotamie antique, région plate se prêtant peu à l'utilisation de frontières naturelles. Par analogie avec les taxons, on peut considérer que les pays peuvent être mis en synonymie (conquis) ou redélimités (lors de conflits territoriaux). Cependant, contrairement aux taxons, les frontières des pays sont âprement défendues depuis des millénaires, comme en témoigne la stèle des Vautours, conservée au Louvre et datant du 2450 AC (Bottéro 1994).

Ceci est logique, puisque le territoire est précieux. Devrait-on également conclure, toujours par analogie, que les taxons, qui représentent le produit du travail des taxonomistes, ne sont pas précieux ? Les taxonomistes n'attachent-ils donc aucune valeur à leur propre travail ?

5. Pourquoi cette nomenclature est-elle toujours utilisée ?

Cette démonstration par l'absurde de l'isolement de la nomenclature Linnéenne-Stricklandienne conduit naturellement à se questionner sur la persistence d'un tel système dans la biologie du 21ème siècle. Bien qu'il soit difficile d'apporter une réponse définitive, deux facteurs ont probablement joué un rôle important. Le premier est de nature historique. Lorsque le premier code de nomenclature fut formulé (Strickland *et al.* 1842), la phylogénétique commençait à peine et la commission qui formula le code pensa sans doute qu'il serait utopique de fonder un code de nomenclature sur une phylogénie qui était alors largement inconnue et donc l'existence même n'était pas acceptée par tous les naturalistes. Même Darwin, qui faisait partie de cette commission, fut sans doute de cet avis car il fut contraint d'accepter des groupes paraphylétiques (n'incluant pas tous les descendants d'un ancêtre) dans ses fameuses monographies sur les cirripèdes, faute de pouvoir proposer une phylogénie détaillée (Padian 1999). Déjà à cette époque, la nomenclature biologique avait une longue histoire que le code de Strickland respecta. Ainsi, les codes de nomenclature Linnéens-Stricklandiens représentent l'aboutissement d'une tradition de plus de deux siècles et demi. Il n'est donc pas surprenant que de nombreux systématiciens ne pensent même pas à remettre en cause un système que des générations de taxonomistes ont utilisé, et qu'ils soient même choqués par l'initiative de certains de remplacer ce système.

La seconde raison de la persistance de la nomenclature Linnéenne-Stricklandienne est simplement l'absence, jusqu'à un passé très récent, d'une alternative viable. En effet, la seule alternative ayant réussi à obtenir le soutien d'une communauté internationale de taxonomistes est la nomenclature phylogénétique, bien que quelques autres aient été proposées (e.g. Dubois 2006 ; Béthoux 2007). Celle-ci fut initialement utilisée il y a un peu plus de 20 ans dans un article sur l'origine des oiseaux (Gauthier 1986). Les principes de base de cette nomenclature furent explicités quelques années plus tard (de Queiroz et Gauthier 1990). Divers autres systématiciens commencèrent alors à utiliser ce système pendant les années 1990 (e.g. Laurin 1991 ; Wolsan 1993 ; Bryant 1994). Une première version du PhyloCode, un code de nomenclature pylogénétique, fut

publiée sur l'Internet en 1998 par un comité de 30 systématiciens, mais une société internationale pour encadrer son développement (ISPN : International Society for Phylogenetic Nomenclature) ne fut fondée qu'en 2004, lors d'un congrès réunissant 70 systématiciens de 11 pays (Laurin et Cantino 2004). Ce code devrait officiellement entrer en vigueur d'ici quelques années, mais il est déjà utilisé par plusieurs de dizaines (si ce n'est de centaines) de systématiciens dans de nombreux pays.

6. Nomenclature phylogénétique

Comment définit-on les noms de taxons en nomenclature phylogénétique ? Il y a trois façons possibles : par nœud (nodale), par branche, ou par apomorphie (Figure 4). Une définition nodale utilise au moins deux types (appelés déterminants, pour les différentier des types Linnéens-Stricklandiens) internes. Elles peuvent prendre la forme : "Le plus petit clade comprenant A et B", ou (ce qui est identique) : "le dernier ancêtre commun de A et B, et tout les descendants de cet ancêtre", où A et B désignent des espèces (ou bien des individus). Le groupe ainsi défini est identifié en rouge sur la figure.

Une définition par apomorphie utilise une apomorphie (nouveauté évolutive ou caractère récent) et au moins une espèce (ou un individu) comme déterminants. Elle peut prendre la forme : "Le clade délimité par l'apomorphie M synapomorphique avec A. Le terme "synapomorphique" signifie que l'apomorphie M n'a pas disparue avant que n'apparaisse l'espèce A, ou que cette dernière n'a pas acquis cette apomorphie de façon convergente. Ainsi, la définition "Aves est le clade délimité par l'apparition de l'aile synapomorphique avec *Passer domesticus* (le moineau domestique)" désigne l'ensemble des oiseaux, mais pas les chauve-souris, ptérosaures (reptiles volants du Mésozoïque), ou insectes, qui ont tous acquis des ailes par convergence (le dernier ancêtre commun de tous ces animaux n'avait pas d'ailes). Le groupe ainsi défini est identifié en bleu sur la figure.

Une définition par branche utilise au moins une espèce (ou individu) comme déterminant interne et au moins une espèce comme déterminant externe. Elle peut prendre la forme : "Le plus grand clade incluant A mais pas Z". Le groupe ainsi défini est identifié en vert sur la figure.

Notez que si seules les espèces A, B et Z étaient initialement connues (C et D n'étant découvertes que plus tard), les taxons correspondant à ces trois définitions auraient même contenu, mais comme les définitions sont de trois types, ces taxons ne sont pas synonymes. D'ailleurs, l'éventuelle

découverte d'espèces supplémentaires (C, D) démontrerait bien que ces taxons sont bien distincts. Même dans ce cas, aucune décision subjective ne doit être prise. L'espèce C appartient seulement au clade défini par une branche (en vert), alors que l'espèce D appartient à ce taxon ainsi qu'au taxon fondé sur une apomorphie.

Ce système n'utilise que des définitions comportant des déterminants et des phylogénies de référence pour délimiter les taxons. Les catégories Linnéennes, qui sont artificielles et dépourvues d'existence objective, ne sont pas utilisées. Deux taxons ne sont synonymes que s'ils réfèrent au même clade. Par exemple, si après que les trois définitions mentionnées ci-dessus soient publiées, un taxonomiste définissait un autre taxon comme "le plus grand clade incluant B mais pas Z", ce dernier serait synonyme du taxon fondé sur une branche (en vert) ; le premier taxon nommé et défini pour ce clade aurait priorité.

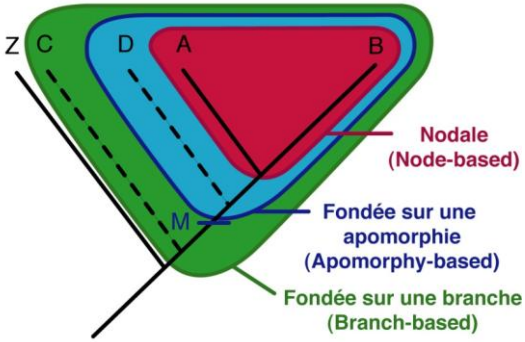


Figure 4. Définitions en nomenclature phylogénétique. Les lettres A–D et Z représentent des espèces ; M représente une apomorphie (caractère récent) Adapté de Laurin (2008a)

Le PhyloCode ne régira pas (au moins initialement) les espèces (Laurin & Cantino 2007), entre autres parce que de nombreux concepts d'espèces sont actuellement utilisés et que sous plusieurs d'entre eux, les espèces ne forment pas forcément des clades (de Queiroz 2007). Ces noms continueront donc à être régis par les codes Linnéens-Stricklandiens.

7. Conclusions et perspectives

Les opposants à la nomenclature phylogénétique prétendent que le passage vers un nouveau type de nomenclature génèrerait trop de confusion pour être profitable (Nixon *et al.* 2003). Cet argument est peu crédible car le système actuel permet un nombre très élevé de sens pour

chaque nom de coexister et permet à chaque clade de posséder de nombreux noms (Figures 1, 2). Peut-on imaginer un système plus confus ? Le PhyloCode permettra tout simplement de stabiliser le sens des noms de taxons en choisissant une des acceptions les plus fréquemment utilisées pour chaque nom, ce qui peut difficilement créer de la confusion.

Une comparaison avec l'informatique peut illustrer la relation entre confusion et progrès. Au début des années 1980, DOS (Disk Operating System) dominait largement le marché des micro-ordinateurs. Puis, Apple inventa le MacOS (en 1984), muni d'un interface graphique. IBM et Microsoft lancèrent peu après leurs propres systèmes d'exploitation munis d'interfaces graphiques (OS/2 et Windows). Cette révolution de l'interface graphique généra plus de confusion chez les millions d'utilisateurs d'ordinateurs du monde entier et coûta bien plus que les sommes mobilisables pour n'importe quelle révolution en nomenclature biologique. Pourtant, aujourd'hui, qui voudrait retourner à DOS ? Cette comparaison montre également que seul un système alternatif comportant des avantages importants et remportant un assez grand succès (comme Windows et le MacOS) peut générer de la confusion chez les utilisateurs des autres systèmes. Les autres systèmes (comme OS/2) n'ont pas assez d'impact pour créer une telle confusion.

Ainsi, seul un assez grand succès du PhyloCode lui permettrait de créer de la confusion chez certains adeptes de la nomenclature Linnéenne-Stricklandienne qui ne se seraient pas familiarisés avec les principes de base de la nomenclature phylogénétique. Or, pour qu'un tel nombre de systématiciens adopte un nouveau système, il faudrait qu'il comporte des avantages réels, car l'inertie favorise toujours le système déjà en place. L'avenir nous dira bientôt si le PhyloCode révolutionnera (ou non) la nomenclature biologique.

Finalement, que reste-t-il de Linné dans la nomenclature d'aujourd'hui ? Les rangs existent toujours, mais pour combien de temps ? Si la nomenclature phylogénétique est adoptée, ils joueront un rôle de moins en moins important. La nomenclature binominale semble devoir persister bien plus longtemps. Même si des propositions alternatives ont été faites (Cantino *et al.* 1999 ; Dayrat *et al.* 2004), elles n'ont pour l'instant pas obtenu le soutien d'une proportion suffisamment grande de la communauté scientifique pour avoir un impact. Ce système de noms binominaux est sans doute la contribution la plus importante et la plus durable de Linné à la nomenclature biologique.

Remerciements

Je remercie Christophe Roche et le comité d'organisation du congrès TOTh 2009 pour l'invitation et le financement qui me permirent de présenter la conférence inaugurale que cette contribution résume et David Marjanovic pour une relecture critique.

Bibliographie

Béthoux O. (2007) : Propositions for a character-state-based biological taxonomy. Zoologica Scripta, vol. 36, N° 4, pp. 409–416

Bottéro J. (1994) : Babylone: à l'aube de notre culture, Gallimard, Paris

Bryant H. N. (1994) : Comments on the phylogenetic definition of taxon names and conventions regarding the naming of crown clades. Systematic Biology, vol. 43, N° 1, pp. 124-130.

Calberg-Challot M., Lerat P., Roche C. (à paraître) : "Quelle place accorder aux corpus dans la construction d'une terminologie ? ", Actes de la troisième conférence TOTh 2009, Terminologie & Ontologie : Théories et applications, Christophe Roche éd., Annecy, Institut Porphyre

Cantino P. D., Bryant H. N., de Queiroz K., Donoghue M. J., Eriksson T., Hillis D. M. et Lee M. S. Y. (1999) : Species names in phylogenetic nomenclature. Systematic Biology, vol. 48, N° 4, pp. 790-807

de Queiroz K. (2007) : Species concepts and species delimitation. Systematic Biology, vol. 56, N° 6, pp. 879-886

de Queiroz, K. et Gauthier J. (1990) : Phylogeny as a central principle in taxonomy: Phylogenetic definitions of taxon names. Systematic Zoology, vol. 39, N° 4, pp. 307-322

Dayrat B., Schander C. et Angielczyk K. D. (2004) : Suggestions for a new species nomenclature. Taxon, vol. 53, N° 2, pp. 485-491

Dubois A. (2006) : Incorporation of nomina of higher-ranked taxa into the International Code of Zoological Nomenclature: some basic questions. Zootaxa, vol. 1337, pp. 1-37

Dubois A. (2007) : Naming taxa from cladograms: a cautionary tale. Molecular Phylogenetics and Evolution, vol. 42, N° 2, pp. 317-330

Ereshefsky M. (2007) : Foundational issues concerning taxa and taxon names. Systematic Biology, 56, (2) pp. 259–301

Frost D. R., Grant T., Faivovich J., Bain R. H., Haas A., Haddad C. F. B., de Sá R. O., Channing A., Wilkinson M., Donnellan S. C., Raxworthy C. J., Campbell J. A., Blotto B., Moler P., Drewes R. C., Nussbaum R. A., Lynch J. D., Green D. M. et Wheeler W. C. (2006) : The amphibian tree of life. Bulletin of the American Museum of Natural History, vol. 297, N° pp. 1–370

Gauthier J. (1986) : Saurischian monophyly and the origin of birds, in Padian, K. (eds.), The Origin of Birds and the Evolution of Flight, California Academy of Sciences, San Francisco, pp. 1–55

Gradstein F. M., Ogg J. G. et Smith A. G. (2004) : A Geologic Time Scale 2004, Cambridge University Press, Cambridge

Hillis D. M. (2007) : Constraints in naming parts of the Tree of Life. Molecular Phylogenetics and Evolution, vol. 42, N° 2, pp. 331-338

Laurin M. (1991) : The osteology of a Lower Permian eosuchian from Texas and a review of diapsid phylogeny. Zoological Journal of the Linnean Society, vol. 101, N° 1, pp. 59-95

Laurin M. (2005) : The advantages of phylogenetic nomenclature over Linnean nomenclature, in Minelli, A. et Ortalli, G. et Sanga, G. (eds.), Animal Names, Instituto Veneto di Scienze, Lettere ed Arti, Venice, pp. 67-97

Laurin M. (2008a) : Le PhyloCode, in Prat, D. et Roguenant, A. et Raynal-Roques, A. (eds.), Peut-on classer le vivant ? Linné et la systématique aujourd'hui, Belin, Paris, pp. 411–420

Laurin M. (2008b) : The splendid isolation of biological nomenclature. Zoologica Scripta, vol. 37, N° 2, pp. 223–233

Laurin M. et Cantino P. D. (2004) : First international phylogenetic nomenclature meeting: a report. Zoologica Scripta, vol. 33, N° 5, pp. 475-479

Laurin M. et Cantino P. D. (2007) : Second meeting of the International Society for Phylogenetic Nomenclature: a report. Zoologica Scripta, vol. 36, N° 1, pp. 109-117

Laurin M. et Bryant H. N. (2009) : Third Meeting of the International Society for Phylogenetic Nomenclature: a Report. Zoologica Scripta, vol. 38, N° 3 pp. 333–337

Lee M. S. Y. et Skinner A. (2007) : Stability, ranks, and the PhyloCode. Acta Palaeontologica Polonica, vol. 52, N° 3, pp. 643–650

Minelli A. (2000) : The ranks and the names of species and higher taxa, or a dangerous inertia of the language of natural history, in Ghiselin, M. T. et Leviton, A. E. (eds.), Cultures and Institutions of Natural History : Essays in the History and Philosophy of Sciences, California Academy of Sciences, San Francisco, pp. 339-351

Nixon K. C., Carpenter J. M. et Stevenson D. W. (2003) : The PhyloCode is fatally flawed, and the "Linnean" System can easily be fixed. The Botanical Review, vol. 69, N° 1, pp. 111-120

Padian K. (1999) : Charles Darwin's views of classification in theory and practice. Systematic Biology, vol. 48, N° 2, pp. 352–364

Rieppel O. (2005) : Monophyly, paraphyly, and natural kinds. Biology and Philosophy, vol. 20, pp. 465-487

Rowe T. et Gauthier J. (1992) : Ancestry, paleontology, and definition of the name Mammalia. Systematic Biology, vol. 41, N° 3, pp. 372-378

Strickland H. E., Henslow J. S., Phillips J., Shuckard W. E., Richardson J. B., Waterhouse G. R., Owen R., Yarrell W., Jenyns L., Darwin C., Broderip W. J. et Westwood J. O. (1842) : Report of a committee appointed "to consider of the rules by which the Nomenclature of Zoology may be established on a uniform and permanent basis". Annals and Magazine of Natural History, 11, pp. 1-17

Strickland H. E., Henslow J. S., Phillips J., Shuckard W. E., Richardson J. B., Waterhouse G. R., Owen R., Yarrell W., Jenyns L., Darwin C., Broderip W. J. et Westwood J. O. (1843) : Series of propositions for rendering the nomenclature of zoology uniform and permanent, being the Report of a Committee for the consideration of the subject appointed by the British Association for the Advancement of Science. Annals and Magazine of Natural History, vol. 11, N° pp. 259-275

Voultsiadou E. et Vafidis D. (2007) : Marine invertebrate diversity in Aristotle's zoology. Contributions to Zoology, vol. 76, N° 2, pp. 103–120

Wolsan M. (1993) : Phylogeny and classification of early European Mustelida (Mammalia : Carnivora). Acta Theriologica, vol. 38, N° 4, pp. 345-384

A propos des auteurs

Michel Laurin
UMR 7207 – Centre de Recherches sur la Paléobiodiversité et les Paléoenvironnements
Muséum National d'Histoire Naturelle, 43 rue Buffon, CP 48
75005 Paris
michel.laurin@upmc.fr
http://tolweb.org/notes/?note_id=3669

SESSION 1

Approche lexico-sémantique de l'extraction terminologique : utilisation de ressources lexicographiques et validation sur corpus

Bertrand Gaiffe, Evelyne Jacquey, Laurence Kister

Résumé : Cet article présente une expérience d'extraction de terminologie à partir d'un dictionnaire en vue d'annoter des textes de spécialité par l'intermédiaire de leurs termes. Il décrit la méthode d'extraction de la terminologie et la méthode d'annotation des textes. Les difficultés liées à l'ambiguïté de forme de certains termes ("aspect" dans le domaine linguistique, par exemple) sont abordés ainsi que quelques solutions destinées y faire face : utilisation d'un dictionnaire (extraction de collocations) et de techniques endogènes habituelles pour l'extraction de candidats termes (patrons syntaxiques) avec l'utilisation de modifieurs recensés comme relevant du domaine considéré par le dictionnaire.

Mots-clés : Terminologie, Acquisition de ressources, Etiquetage de textes

1. Introduction

L'augmentation et la diversification des échanges, la généralisation de l'Internet, conduisent à une explosion de la quantité d'informations textuelles à laquelle il est possible d'accéder très facilement. Chaque domaine scientifique est en constante évolution et on constate de plus en plus d'interpénétrations des domaines de spécialités du fait des nombreuses questions abordées de manières multi- ou interdisciplinaires. La difficulté relative à cette évolution réside dans la faible précision de l'information obtenue et son manque d'exhaustivité. Comme l'ont souligné (Bourigault & Aussenac-Gilles 2003), les ressources terminologiques et ontologiques, leur extraction et leur structuration constituent une contribution majeure pour les travaux touchant l'information et la documentation (Roche 2004), l'édition (El Mekki & Nazarenko 2002), la recherche d'informations, la classification de documents (Sanjuan & Ibekwe-Sanjuan 2002), ou encore la détection de documents à caractère raciste sur la Toile (Valette & Grabar 2004). Comme le soulignent la plupart des travaux sur le sujet, notamment (Bourigault *et al.* 2001) et (L'Homme 2004), l'extraction automatique et la structuration de ressources terminologiques (Nazarenko & Hamon 2002) ont déjà donné lieu à la création de nombreux outils s'appuyant sur différentes méthodes (symbolique/numérique), prenant en compte différents types d'informations pour la structuration (regroupements sur la base de relations hiérarchiques d'ordre conceptuel, de relations de sémantique lexicale, de convergences et de divergences entre termes, de l'analyse distributionnelle des contextes d'apparition des termes dans les textes, etc). Les travaux évoqués ci-dessus partagent un point méthodologique important : les ressources terminologiques produites sont essentiellement extraites à partir de textes liés à un domaine de spécialité, qu'il s'agisse d'une extraction automatique, manuelle ou assistée. L'utilisation de ressources externes, comme par exemple l'utilisation de dictionnaires, est seconde (L'Homme 2004).

Les travaux présentés ici proposent de partir d'une ressource lexicographique de référence fortement domanialisée, le Trésor de la Langue Française informatisé (TLFi) (Dendien & Pierrel 2002). Cette ressource comporte de nombreux sens explicitement associés à des domaines de spécialité (97330 sens domanialisés sur 271165, soit près de

36 %[1]). A partir d'une terminologie extraite automatiquement pour le domaine des sciences du langage, nous procédons à une validation sur un corpus spécialisé grâce à la détection automatique des termes dans les textes. Cette approche permet de contribuer à l'avancement des recherches dans trois domaines : l'information et la documentation, la linguistique textuelle et l'analyse des relations sémantiques et discursives.

Dans le domaine de l'information et de la documentation, plusieurs auteurs ont montré qu'il est possible d'indexer des textes à partir d'une terminologie ou d'un thésaurus (Bourigault *et al.* 2004) et (Aussenac-Gilles *et al.* 2000). Sur ce point particulier, nous avons montré dans des travaux antérieurs que Thésaulangue, le thésaurus constitué au laboratoire et intégré au portail terminologique de l'INIST, n'est pas totalement satisfaisant pour l'indexation et la classification des documents par les documentalistes (Kister & Jacquey 2007a et b ; Kister, Jacquey & Gaiffe 2008). Disposer d'une terminologie validée sur un corpus scientifique de spécialité constitue une plus-value pour le travail d'annotation automatique ou semi-automatique que nous envisageons.

Dans le domaine de la linguistique textuelle, la poursuite de nos travaux consacrés à la confrontation de la structure hiérarchique d'une terminologie et de la structure thématique des textes spécialisés correspondants est nécessaire. La structure thématique est appréhendée par repérage et étiquetage des termes dans les textes en tenant compte des différentes variations qu'ils subissent comme, par exemple, la reprise anaphorique. Ces travaux s'inscrivent dans la mise en regard de la terminologie et de la linguistique textuelle (Poibeau 2005) et dans le développement d'une terminologie textuelle (Bourrigault & Slodzian 1999).

L'importance du repérage et de l'étiquetage des termes dans les textes nous permet d'étudier 'in vivo' les relations sémantiques et discursives (L'Homme 2004) qui pourraient amender et améliorer la terminologie. L'analyse de ces relations doit permettre de mieux structurer la terminologie et par la même le thésaurus.

2. Extraction de la terminologie à partir du TLFi

La définition de "terme" que nous adoptons est fondée sur celle proposée par (L'Homme 2004) qui montre et synthétise l'évolution sémantique et conceptuelle de celui-ci depuis (Wüster 1981). A la suite de

1 Pour déterminer cette proportion, nous faisons l'hypothèse que chaque définition dans chaque bloc cohérent d'information du point de vue lexicographique représente un sens du lemme défini ou de l'élément de composition.

L'Homme, nous considérons un terme comme "une unité lexicale associée à un domaine de spécialité", pouvant être réalisé sous forme simple (lexème) ou complexe (syntagmes). Comme mentionné dans l'introduction, la plupart des travaux en extraction de terminologie utilisent des critères syntaxiques et des critères statistiques pour identifier des candidats termes. En ce qui nous concerne, les termes sont extraits à partir d'une ressource lexicographique. Nous nous sommes limités dans un premier temps à la catégorie du nom - les autres catégories grammaticales ne sont impliquées que dans la mesure où elles apparaissent dans des termes complexes (consonne fricative sonore, changement de sens) globalement considérés comme des noms complexes.

La ressource lexicographique que nous utilisons est le Trésor de la Langue Française informatisé (TLFi). Ce dictionnaire jouit d'une couverture remarquable pour le français des 19ème et 20ème siècles. Il comporte ~92 000 entrées (principales et secondaires) parmi lesquelles ~90 000 concernent des unités lexicales et grammaticales. Du point de vue lexical, il comporte :

- ~9 000 entrées de verbes
- ~16 500 entrées d'adjectifs
- ~48 300 entrées de noms
- ~11 700 entrées de noms et d'adjectifs confondus

ce qui correspond à ~85 500 entrées qui produisent ~278 500 définitions dont ~94 800 dépendent d'un domaine technique explicite, soit une proportion de ~34%. Dans le cadre d'expériences faisant intervenir une approche TAL, le dictionnaire est utilisé sous son format XML et étiqueté en catégories grammaticales pour les définitions et les exemples. Dans cette version XML, les techniques classiques de transformation XSLT permettent d'atteindre les objets lexicographiques balisés. Ceux qui nous intéressent ici, sont :

- la vedette <ved>
- le code grammatical <cod>
- les blocs d'information <B>
- les indicateurs d'emploi <ind>
- l'indication d'un domaine <dom>
- le texte de la définition <def>
- les conditions d'usage <cro> – délimitées par des crochets
- les synonymes et antonymes <syno>
- les syntagmes illustratifs <syntita n=i> qui sont les constructions courantes du lexème

- l'organisation hiérarchique <H>

Il faut encore préciser un point particulier concernant les collocations définies, c'est-à-dire les objets lexicographiques proches de ce qu'on appelle des expressions figées dans la littérature linguistique. Dans le dictionnaire, ces éléments sont considérés comme des entrées à part entière et sont repérables automatiquement car balisés par <syntita n=d>, directement suivi d'une définition <def>.

2.1. Procédure d'extraction

Dans le TLFi, nous avons sélectionné le domaine Sciences du Langage qui compte plusieurs sous domaines parmi lesquels nous avons sélectionné les sous domaines suivants[2] : grammaire, lexicographie, lexicologie, linguistique, philologie, phonétique, phonologie, rhétorique, sémiologie, sémiotique, stylistique, toponymie.

La procédure d'extraction consiste, dans un premier temps, en une feuille XSLT qui extrait du TLFi les informations pertinentes relatives à chaque occurrence d'un des domaines mentionnés. Si la vedette est de catégorie « substantif », la feuille extrait cette vedette, son code grammatical, les définitions relatives au domaine et les éventuels synonymes ou antonymes. Si, au contraire, la mention de domaine est dominée hiérarchiquement par un syntagme défini, on extrait sa définition et ses éventuels synonymes et antonymes. Dans un second temps, la terminologie extraite est revue manuellement pour calculer des variantes formelles. L'extraction de ce type de variantes fait l'objet de procédures automatisées sur l'italien (Dell'Ortella *et al.* 2008) pour des corpus juridiques et environnementaux. Outre la réduction de formes de vedettes telles "aberrant, -ante" ramenée à "aberrant", cette étape permet de ramener des termes complexes tels "Grammaire comparée ou linguistique comparée" aux variantes "Grammaire comparée" et "linguistique comparée"[3]. La dernière tâche réalisée à cette étape est l'étiquetage en parties du discours de chacune des variantes isolées.

2.2. Terminologie extraite pour les sciences du langage

La terminologie extraite à partir du TLFi comporte 2 402 entrées. En comparaison, le thésaurus initialement constitué dans le domaine des

2 Les sous domaines utilisés dans le TLFi sont accessibles dans la recherche assistée.

3 Dans le TLFi, parce qu'il a d'abord été édité sous la forme papier et pour limiter la place occupée par les informations, les lexicographes ont établi plusieurs procédés typographiques leur permettant de factoriser l'information. Dans la mesure où la terminologie devait être utilisée pour étiqueter des corpus, il a fallu expanser les informations factorisées.

sciences du langage au laboratoire comporte 872 entrées. Le nombre d'entrées est multiplié par 2,75. Cependant, l'augmentation en quantité ne se fait pas forcément à qualité constante, c'est pourquoi nous procédons ensuite à un étiquetage sur corpus et nous analysons quantitativement et qualitativement les résultats.

Du point de vue qualitatif, la terminologie, contrairement au thésaurus, n'est pas structurée mais elle est plus riche et plus précise car chaque entrée de la terminologie dispose de l'ensemble des informations lexicographiques extraites dans le bloc d'information correspondant à l'un des sous-domaines des sciences du langage. Ces informations lexicographiques concernent les conditions d'emploi, les liens de synonymie éventuels, les syntagmes illustratifs du sens domanialisé et un ou plusieurs exemples, le cas échéant. Dans l'exemple ci-dessous, le terme "voyelle" n'est pas seulement atteint, mais il est aussi associé aux informations présentes dans la ressource lexicographique qui fournit l'indication du domaine, la définition de "voyelle", un renvoi synonymique et une locution courante dans laquelle le terme apparaît.

```
<terme>
  <ved xml:id="e1876">voyelle</ved>
  <categorie>voyelle:n</categorie>
  <dom>PHONÉT.</dom>
  <def>Phonème constituant à lui seul un son ...</def>
  <syno>Synon. vx voix (v. ce mot I A 1)</syno>
  <syntagme_illustratif>Système des voyelles françaises.</syntagme_illustratif>
</terme>
```

En comparaison avec les approches d'extraction de termes à partir de textes, l'avantage est le fait de réduire la vérification manuelle :

- le dictionnaire - la ressource initiale - a été vérifié par des experts linguistes au moment de sa rédaction, il n'est pas nécessaire, à cette étape, de faire appel à des experts du domaine
- le nombre très restreint de termes extraits réduit grandement le coût des traitements manuels
- les variantes associées aux termes - au nombre de 472 - ont été vérifiées et munies des informations lexicographiques du bloc d'information extrait en fonction de l'étiquette de domaine

  A titre de comparaison, (Aussenac-Gilles & Bourigault 2000) extraient 21068 candidats de fréquence supérieure ou égale à 1 à partir de deux corpus :

- le corpus AFIA – riche de 31 212 occurrences - qui regroupe des descriptions de laboratoire
- le corpus LIVRIC – riche de 178 336 occurrences - qui contient des publications scientifiques

L'ensemble des termes candidats repérés, après avoir été filtrés en fonction du nombre de documents dans lesquels ils apparaissent et/ou leurs fréquences, sont ensuite évalués manuellement par les auteurs des articles. Chaque évaluateur évalue en moyenne 81 candidats extraits du corpus LIVRIC et 48 candidats repérés dans le corpus AFIA.

Pour la terminologie que nous avons constituée à partie du TLFi, le traitement manuel a consisté à vérifier les codes grammaticaux, l'expansion des alternatives, des énumérations et des optionalités ainsi que la qualité de l'extraction opérée automatiquement.

3. Repérage et étiquetage des termes en corpus

3.1. Le corpus

Le corpus de spécialité constitué pour l'expérience que nous présentons ici est d'une taille raisonnable selon (Bourigault & Aussenac-Gilles 2003) : il comporte 149 772 occurrences. Il est relativement homogène et cohérent avec la ressource lexicographique qui a permis d'extraire la terminologie que nous désirons valider sur ce corpus. Par la suite, nous envisageons d'appliquer la méthode à d'autres documents - le corpus des journaux du CNRS[4] - du domaine des sciences du langage ou d'autres domaines apparaissant dans le TLFi afin de partir d'une terminologie initiale de même nature que celle utilisée pour cette expérience.

Le corpus actuel compte trois œuvres fondatrices en linguistique datant du 20ème siècle : Cours de linguistique générale de Ferdinand de Saussure [1916], Le langage et la vie de Charles Bally (1952) et La linguistique de Jean Perrot (édition de 1989).

3.2. Procédure de détection et d'annotation

Pour repérer les termes dans les textes, nous avons étiqueté les textes en parties du discours en utilisant TreeTagger, nous avons ensuite identifié

4 Des négociations, à l'initiative du service de communication et de valorisation de la recherche du laboratoire sont en cours afin que nous puissions utiliser ces documents sous droit.

les termes grâce aux variantes également étiquetées en parties du discours. L'étiquetage morpho-syntaxique est évidemment indispensable pour s'affranchir de confusions de formes telles : "son" adjectif possessif versus "son" nom commun et terme dans le domaine qui nous intéresse. Outre cet exemple anecdotique, l'étiquetage en morpho-syntaxe permet d'éviter un grand nombre de confusions nom commun/adjectif.

Nous avons étiquetés tous les termes, même si nous ne retenons in fine que ceux d'extension maximale :

```
<terme ref='e630'><terme
ref='e637'>grammaire</terme>comparée</terme>
```

Notons enfin que chaque occurrence de terme annotée fait référence (via l'attribut <ref>) à la terminologie extraite, ce qui permet, si nécessaire, de se reporter aux informations lexicographiques associées au terme.

3.3. Résultats quantitatifs

Comme le montrent les décomptes ci-dessous (Tableau 1 - Densité en termes par rapport aux noms présents dans le corpus), le nombre d'occurrences de termes reconnus parmi les noms présents dans le corpus est multiplié par 10 par rapport au nombre d'entités dans le thésaurus : 3% des noms sont étiquetés comme des termes à partir de la nomenclature du thésaurus quand 32% des noms sont étiquetés comme des termes à partir de la nomenclature complète de la terminologie, termes et variantes. Parallèlement, le nombre de termes différents reconnus passe de 132 avec le thésaurus à 451 avec la terminologie.

	Corpus	Nb Noms	Thésaurus	Terminologie
Occurrences	149 772	54 119	1716 - 3%	17374 - 32%
Candidats Termes	5 662	2 629	132	451

Tableau 1. Densité en termes par rapport aux noms présents dans le corpus

L'exemple ci-dessous montre un résultat d'annotation (termes reconnus en gras et expressions référentielles en fonction de l'étiquetage[5], soulignées).

La numération intéresse tous les **aspects** du langage : de la **phonologie** (**nombre** et **fréquence** des **phonèmes** dans une **langue** donnée) à la **syntaxe** (par exemple , **fréquence relative** des différentes dispositions possibles dans la **phrase** pour les éléments constituants) , au **lexique** (des dénombrements en montrent l'**extension** , liée aux besoins auxquels le **vocabulaire** doit répondre) et à la **stylistique** , qui tend de plus en plus à prendre pour **base** , dans l'appréciation des faits individuels, des statistiques de **fréquence** des différentes **réalisations** possibles dans la **langue** .

4. Analyse des résultats obtenus

Un premier regard sur les résultats d'étiquetage permet de constater que la procédure donne de bons résultats tant que les formes des termes ne sont pas ambiguës et qu'il s'agit, lorsqu'ils sont réalisés de différentes manières, de leur réalisation maximale. Cependant, les termes peuvent apparaître sous une forme ambiguë, soit par nature, soit du fait de l'élision d'éléments par rapport à sa forme maximale. Ainsi « objet », par exemple, a une définition terminologique très claire, il s'agit du complément d'objet direct, mais ce nom a aussi un sens très général qui est celui que l'on trouve dans "cette étude a pour objet". De la même manière, "aspect" est reconnu à tort comme terme dans l'exemple ci-dessus dans "tous les aspects du langage" alors qu'il a un sens non équivoque en sciences du langage quand on s'intéresse à la conjugaison.

Pour aller au delà de cette première intuition, nous comparons les fréquences relatives des termes présents dans le corpus de spécialité ~150 000 occurrences - données à 1 pour 1 000 - avec celles que l'on observe dans un corpus ne relevant pas de cette spécialité constitué de deux journées complètes de l'Est Républicain toutes éditions locales confondues soit ~682 000 occurrences. Bien entendu, les deux corpus reçoivent les mêmes prétraitements (normalisation, étiquetage morpho-syntaxique, lemmatisation). La comparaison des fréquences est illustrée sur le graphique suivant pour les formes les plus fréquemment rencontrées dans le corpus de spécialité.

5 Les étiquettes utilisées sont du type NP pour nom propre, PR pour pronom, etc.

	Langue	forme	son	cas	rapport	temps	analogie	sujet	objet	lieu
F1[6]	0,76	0,25	0,16	0,14	0,13	0,10	0,06	0,05	0,07	0,04
F2[7]	0,04	0,10	0,04	0,21	0,10	0,60	0	0,08	0,15	0,75
D	0,72	0,15	0,12	0,07	0,03	0,50	0,06	0,03	0,08	0,71

Tableau 2. **Fréquences relatives sur le corpus de spécialité et le corpus tout venant**

Au vu de ce tableau, deux cas de figures apparaissent. De façon très claire, un terme comme "langue" est peu ambigu dans le corpus de spécialité bien que très fréquent dans la mesure où il est au contraire peu fréquent dans le corpus tout venant. Les formes "lieu" et "temps" ont un comportement exactement inverse. Il s'agit probablement d'éléments qui différencient les deux types de corpus et on peut donc penser que leurs emplois dans le corpus de spécialité sont essentiellement terminologiques. Le second cas de figure est illustré par des formes telles "son", "cas", "objet" ou "sujet" pour lesquelles on peut soupçonner que l'étiquetage en termes fondé sur leur seule forme demande à être vérifié. Deux pistes sont actuellement explorées :

- éliminer de l'analyse des emplois relevant de collocations : "au sujet de", "dans tous les cas", etc.
- enrichir la terminologie de termes complexes construits à partir de ces termes de forme ambiguë : "cas nominatif", "sujet profond", etc.

4.1. Détection des collocations et filtrage

Deux sources fournissent des collocations. La première consiste à reprendre l'ensemble des collocations référencées dans le TLFi si elles ne sont pas associées au domaine considéré. Les collocations suivantes ont été trouvées pour "sujet" : "sujet (battu et) rebattu, à ce sujet, au sujet de, sujet psychologique, sujet-contact, sujet de la connaissance, sujet transcendantal, sujet secondaire". La seconde source est fournie par tout corpus tout venant dont, une fois la liste de formes à vérifier connue, il est aisé d'extraire les collocations les plus fréquentes.

En pratique, il faut croiser les deux sources : la notion de collocations "les plus fréquentes" dans un corpus suppose le choix d'un seuil difficile à

6 Fréquence 1 = fréquence relative dans le corpus de spécialité
7 Fréquence 2 = fréquence relative dans le corpus 'tout venant'

fixer. Par ailleurs, les collocations du TLFi ne peuvent pas être exploitées directement, comme pour la terminologie extraite, elles demandent quelques corrections manuelles destinées à expanser et vérifier l'information transcrite de manière économique dans le dictionnaire.

4.2. Enrichissement en termes complexes

Pour extraire la terminologie, nous avons pris le parti de ne considérer que les substantifs étant donné le rôle central joué par cette catégorie comme le souligne la littérature sur le sujet. Comme le mentionne (L'Homme 2004), les adjectifs et les verbes participent à la définition d'une terminologie exhaustive. C'est également la stratégie à l'œuvre dans LEXTER qui considère des patrons dont l'expression maximale est ADJ? NOM [NOM | ADJ | de]*. Pour détecter des candidats termes supplémentaires (qui permettront de désambiguïser certains des emplois des termes de forme ambiguë), nous avons extrait les adjectifs relatifs au domaine des Sciences du Langage et nous avons récolté les "chunk" auxquels ces adjectifs participent. On récolte de cette façon 213 candidats termes dont 56 sont des termes nouveaux. Parmi eux, certains permettent de désambiguïser, par exemple, "rapport" dans "rapports syntagmatiques".

L'inconvénient, en revanche, de ces méthodes est que les candidats termes ainsi récoltés n'ont pas de définitions.

5. Conclusions et perspectives

Dans cet article, nous avons explicité une procédure permettant d'extraire une terminologie des Sciences du Langage à partir de la ressource lexicographique que constitue le TLFi. Dans sa forme XML, le balisage des objets lexicographiques tels la vedette, le domaine d'emploi et la définition permettent l'extraction, minimalement des triplets <ved>, <dom>, <def>. Un traitement manuel a ensuite été appliqué afin d'expanser les informations factorisées, notamment les différentes variantes possibles de collocations définies comme, par exemple, "complément d'objet direct", "complément d'objet", etc. Le second aspect du traitement manuel de la terminologie a consisté à étiqueter grammaticalement les termes pour lesquels les codes grammaticaux ne pouvaient pas être inférés à partir de l'objet lexicographique <cod>.

Afin de procéder à la validation de la terminologie extraite, et parce que nous nous inscrivons dans la perspective de deux types de recherches en cours :

- l'étude de l'évolution thématique des termes dans les textes, dans la philosophie de travaux du même ordre (Ferret & Grau 2006) et à la suite de travaux personnels antérieurs (Kister & Jacquey 2007a et b ; Kister, Jacquey & Gaiffe, 2008)
- l'étude des relations lexicales sémantiques entre termes

nous avons effectué l'étiquetage d'un corpus homogène, de taille raisonnable dans le domaine des Sciences du Langage.

L'analyse des résultats montre une nette amélioration quantitative par rapport au thésaurus utilisé initialement dans les perspectives de recherche évoquées ci-dessus. Sur le plan qualitatif, nous avons avancé plusieurs pistes de solution afin de traiter notamment la question de l'ambiguïté des formes des termes, même lorsqu'elles apparaissent dans un corpus de spécialité et qu'elles font référence à des termes non ambigus de la terminologie.

Pour la suite, nous envisageons de structurer la terminologie en prenant appui sur les travaux déjà menés dans cette perspective (Nazarenko et Hamon 2002). Deux axes seront mis en œuvre :

- un axe en domaines, autrement dit, structurer les termes selon l'organisation hiérarchique du thésaurus
- un axe spécifique/générique en exploitant les définitions associées aux termes

Pour ce qui est de l'étiquetage dans les textes, les solutions esquissées pour le traitement de l'ambiguïté des formes des termes seront intégrées dans la procédure automatique d'ensemble.

Bibliographie

Aussenac-Gilles Nathalie & Bourigault Didier (2000) : The Th[IC]2 Initiative : Corpus-Based Thesaurus Construction for Indexing WWW Documents, in Proceedings of the EKAW'2000 workshop "Ontologies and texts", Juan-Les-Pins, Université Paul Sabatier, Toulouse, pp. 71-78, octobre 2000

Bourigault Didier et Slodzian Monique (1999) : Pour une terminologie textuelle, in Terminologies nouvelles, 19, pp. 29-32

Bourigault Didier, Jacquemin Christian et L'Homme Marie-Claude (2001) : Recent Advances in Computational Terminology, Amsterdam/Philadelphie : John Benjamins

Bourigault Didier et Aussenac-Gilles Nathalie (2003) : Construction d'ontologies à partir de textes, Dans : Actes de la conférence TALN 2003, Batz-sur-Mer

Bourigault Didier, Aussenac-Gilles Nathalie & Charlet Jean (2004) : Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas, Dans : Revue d'Intelligence Artificielle (RIA), Numéro spécial sur les techniques informatiques de structuration de terminologies, M. Slodzian (Ed.), Hermès, Paris, Vol. 18, N. 1/2004, pp. 87-110

Dell'Ortella F., Lenci A., Marchi S., Montemagni S., Pirelli V. & Venturi G. (2008) : *Dal testo alla conoscenza e ritorno : estrazione terminologica e annotazione semantica di basi documentali di dominio*, Analisi Testuale e Documentazione nella città digitale, Convegno nazionale dell'Associazione Italiana per la Terminologia, I-TerAnDo, Università di Calabria, Rende, 5-6-7 juin, AIDAinformazioni, 26, 1-2, pp. 185-206

Dendien Jacques et Pierrel Jean-Marie (2002) : Le trésor de la langue informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence, in TAL

El Mekki Touria et Nazarenko Adeline (2002) : Comment aider un auteur à construire l'index d'ouvrage, in Actes de la conférence CIFT 2002. pp. 141 – 157, Tunisie

Kister Laurence & Jacquey Evelyne (2007) : Comparaison des structures thématiques de textes spécialisés et de thésaurus ou de terminologies, Terminologia e mediazione linguistica : approcci e metodi a confronto, ASS.I.term et Università di Bologna, sede di Forli, Bertinoro, 8 juin, Realiter, en ligne, (http://realiter.net/spip.php?article951)

Kister Laurence & Jacquey Evelyne (2007) : Acquisition sémantique à partir de données lexicographiques au service de la comparaison entre des structures thématiques de textes spécialisés et de thésaurus, Terminologie : approches transdisciplinaires, Gatineau (Québec), 2-4 mai, an ligne, (http://www.uqo.ca/terminologie2007/documents/kister_Jacquey.pdf)

Kister Laurence, Jacquey Evelyne & Gaiffe Bertrand (2008) : *Repérage de la référence à partir du thesaurus, de la terminologie et de la sémantique lexicale*, Analisi Testuale e Documentazione nella città digitale, Convegno nazionale dell'Associazione Italiana per la Terminologia, I-TerAnDo, Università di Calabria, Rende, 5-6-7 juin, AIDAinformazioni, 26, 1-2, pp. 25-36

L'Homme Marie-Claude (2004) : La terminologie : Principes et Techniques. Presses Universitaires de Montréal

Nazarenko Adeline & Hamon Thierry (2002) : Structuration de terminologie, Editeurs de ce numéro dans TAL, vol 43, n°1, 174 p

Poibeau Thierry (2005) : Parcours interprétatifs et terminologie, in Actes de la conférence TIA 2005, Rouen

Roche Mathieu (2004) : Intégration de la construction de la terminologie de domaines spécialisés dans un processus global de fouille de textes, PhD Thesis, Université Paris 11

Sanjuan Eric & Ibekwe-Sanjuan Fidelia (2002) : Terminologie et classification automatique des textes, in Actes de la conférence JADT 2002, pp. 677-688.

Valette Mathieu & Grabar Natalia (2004) : Caractérisation de textes à contenu idéologique : statistique textuelle ou extraction de syntagme ? L'exemple du projet PRINCIP, in Actes de la conférence JADT 2004

Wüster Eugen (1981) : L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses, in Textes choisis de terminologie 1, Fondements théoriques de la terminologie, Presses de l'université de Laval, pp. 55-114

A propos des auteurs

Bertrand Gaiffe

Lexique ATILF UMR 7118
44, avenue de la Libération BP 30687
54063 Nancy Cedex
www.atilf.fr
Bertand.Gaiffe@atilf.fr

Evelyne Jacquey

Lexique ATILF UMR 7118
44, avenue de la Libération BP 30687
54063 Nancy Cedex
www.atilf.fr
Evelyne.Jacquey@atilf.fr

Laurence Kister

Lexique ATILF UMR 7118
44, avenue de la Libération BP 30687
54063 Nancy Cedex
www.atilf.fr
Laurence.Kister@atilf.fr

Quelle place accorder aux corpus dans la construction d'une terminologie ?

Marie Calberg-Challot, Pierre Lerat, Christophe Roche[1]

Résumé : Au travers d'une démarche scientifique pluridisciplinaire, nous nous interrogeons sur la place à accorder aux corpus dans la construction d'une terminologie. Pour répondre à cette question, il nous faudra dans un premier temps définir les notions de "terme" et de "concept" pour ensuite préciser ce qu'est une terminologie et quel est son rôle. Ceci nous conduira à nous interroger sur le statut des mots qui composent tout texte spécialisé. En effet, ces mots relèvent-ils du lexique d'une langue de spécialité ou appartiennent-ils à la terminologie du domaine ? Dans ce dernier cas, ils deviendraient des termes. De même, le concept appartient-il au texte ou la connaissance du domaine est-elle hors du texte ? Nous insisterons sur l'importance de séparer les deux dimensions de la terminologie, à savoir la dimension conceptuelle d'une part, et la dimension linguistique d'autre part. Ceci nous conduira à différencier définitions formelles et définitions en langue comportant chacune, comme toute définition scientifique, des informations sur ce qu'est la chose - sa nature - , la perception que l'on en a - sa description - et l'usage que l'on peut en faire - sa fonction. L'ensemble de ces données nous permettra alors d'introduire la notion d'ontoterminologie. Nous présenterons enfin notre méthode de travail illustrée au travers d'exemples. Les terminologies réalisées et validées par des experts dans diverses communautés de travail et domaines mettront en lumière la place que nous accordons aux corpus dans notre méthode.

Mots-clés : Corpus, terme, concept, connaissance, terminologie, ontologie, ontoterminologie.

1 Cet article réunit trois auteurs aux profils complémentaires, avec pour certains une spécialisation plus forte en linguistique, terminologie et/ou logique, qui se sont rencontrés dans un souci de répondre à des préoccupations communes.

1. Introduction

Convient-il encore de s'interroger sur la place à accorder aux corpus dans la construction d'une terminologie ?

Dans un contexte où l'utilité des corpus semble acquise, force est cependant de constater que les terminologies ainsi construites restent prisonnières des contingences du discours, loin des propriétés attendues de la terminologie classique telles que le consensus, le partage ou la cohérence. Faut-il pour autant rejeter ou nier des propriétés qu'on ne peut démontrer ? Ne faut-il pas davantage s'interroger sur la méthode suivie et, ici, se pencher à nouveau sur le rôle exact des corpus pour la terminologie ?

Pour répondre à ces questions, nous suivrons une démarche scientifique pluridisciplinaire puisant à la linguistique, la terminologie et la logique. Après un détour historique et un survol des acceptions dans la littérature contemporaine, nous préciserons les définitions que nous adopterons pour les notions de "terme" et de "concept" pour ensuite formuler ce qu'est une terminologie et quel est son rôle.

Ceci nous conduira à nous interroger sur le statut des mots qui composent tout texte spécialisé. En effet, ces mots relèvent-ils du lexique d'une langue de spécialité ou appartiennent-ils à la terminologie du domaine [2] ? Comment un mot accède-t-il au statut de terme ? Comment distinguer un mot d'un terme ? De même, le concept appartient-il au texte ou la connaissance du domaine se trouve-t-elle hors du texte ?

Nous soulignerons l'importance de séparer les deux dimensions de la terminologie, à savoir la dimension conceptuelle d'une part, et la dimension linguistique d'autre part. Ceci nous conduira à différencier définitions formelles et définitions en langue comportant chacune, comme toute définition scientifique, des informations sur ce qu'est la chose - sa nature - , la perception que l'on en a - sa description - et l'usage que l'on peut en faire - sa fonction. L'ensemble de ces données nous permettra alors d'introduire la notion d'ontoterminologie.

La dernière partie sera consacrée à notre méthode de travail illustrée au travers d'exemples pour la construction de terminologies. Les terminologies réalisées et validées par des experts dans diverses communautés de travail et domaines mettront alors en lumière la place que nous accordons aux corpus dans notre méthode.

---

2 Ce qui sous-entend que le lexique de la langue de spécialité et l'ensemble des termes ne se confondent pas.

2. A propos des notions de "terme" et de "concept"

2.1. Détour historique

Avant d'aller plus avant dans cet exposé, traçons brièvement un historique des notions de "terme" et de "concept" qui nous permettra de comprendre les diverses acceptions qui se rencontrent aujourd'hui. Rappelons auparavant que la terminologie est liée à une communauté de travail[3] qui partage un domaine de connaissances et une même langue de spécialité. Elle traduit l'idée que "le meilleur moyen pour éviter la confusion des mots qui se rencontrent dans les langues ordinaires, est de faire une nouvelle langue et de nouveaux mots, qui ne soient attachés qu'aux idées que nous voulons qu'ils représentent" (Arnault & Nicole 1992 : 78).

La notion de "terme", du latin "terminus" dans le vocabulaire de l'Organon scolastique, désigne ce qui dé-limite une proposition, comme le point limite de la ligne (Cassin 2004 : 1284). La notion de "terme" a ensuite évolué pour désigner "ce qui limite le sens" d'un mot et le *terme* devient l'équivalent d'un "mot spécifique" (Pruvost 2008 : 10).

En effet, "étant des noms de notions, les termes suscitent des attentes doubles : il faut que ce soient des unités terminologiques intégrables dans des énoncés et pouvant y remplir des fonctions syntaxiques, même si leur morphologie n'est pas conforme aux règles de bonne formation lexicale, et il faut en même temps que ce soient des unités de connaissance à contenu stable, donc plus indépendantes du contexte que les mots ordinaires. La première exigence engage la cohérence de l'analyse linguistique, la seconde applique le principe scientifique de réflexivité, c'est-à-dire de l'identité constante des unités prises en compte. […] Les dénominations techniques sont dans la langue puisqu'elles sont susceptibles d'être traduites en langue étrangère, mais ce sont des dénominations de connaissances spécialisées, et c'est ce qui les rend pertinentes terminologiquement" (Lerat 1995 : 45).

La notion de concept, quant à elle, du latin "conceptus", prend une place significative avec la terminologie philosophique occidentale dans la seconde moitié du XIIIe siècle. Elle est utilisée dans une acception dérivée au sens de "représentation intellectuelle se développant dans l'esprit" et prend une grande importance avec les théoriciens de la connaissance. Cette notion prend son essor pour se démarquer de la notion d' "intellectus" désignant à la fois la "faculté intellectuelle et ses unités de

3 Communauté qui relève d'une même pratique professionnelle et qui partage une conceptualisation du monde et une langue communes.

représentation – et parfois même le sens des mots". La notion de "conceptus" dénote au sens littéral une représentation mentale et par son étymologie "le rassemblement d'une pluralité d'éléments dans une appréhension unique" (Cassin 2004 : 248). Production intérieure de la pensée d'une part, et généralité de l'autre, telles sont bien les deux composantes clés du "conceptus". L'usage ultérieur de *concept* varie entre l'évocation d'un objet abstrait entièrement dépsychologisé (Frege 1971) et celle d'une représentation mentale.

Le concept est donc tantôt défini comme "la représentation mentale abstraite d'un objet, distincte des idées particulières", tantôt comme "une représentation symbolique associée à un signe linguistique" (Neveu 2004 : 76).

Le concept correspondrait à une "représentation intellectuelle permettant de viser le réel suivant des déterminations abstraite et générale et non dans sa singularité concrète" (Dictionnaire de philosophie)

Pour la terminologie, le concept est un "contenu de connaissances normé et nommé de façon consensuelle" (Lerat 2009b : 76).

On comprend alors les deux types d'activités qui découlent naturellement de la terminologie. La première activité centre son travail sur les mots et le vocabulaire de spécialité et opterait pour une démarche plutôt sémasiologique tandis que la seconde se concentre davantage, au travers d'une démarche onomasiologique, sur les concepts et les termes les dénotant. Ces deux activités complémentaires et indissociables gagneraient à se distinguer pour mieux restituer l'ensemble des travaux du domaine de la terminologie.

2.2. Qu'en est-il aujourd'hui ?

Un terme est une "unité définie dans les textes de spécialité" (Kocourek 1991 : 180). La définition de Kocourek fait du terme un mot "presque" comme les autres. Partant de cette vue uniquement linguistique de la terminologie est apparue une terminologie qui se confond avec une lexicographie de spécialité.

Combien d'unités terminologiques découpent de façon stable la réalité – car c'est bien là que se trouve le terme – dans des résultats d'extraction de termes candidats quand ne seront retenus in fine que quelques centaines de termes sur plusieurs milliers produits (Calberg-Challot *et al.* 2008b) ? Quel est le sens des termes retenus ? Beaucoup d'expressions dénominatives (mettant en jeu la polysémie, la synonymie, la reformulation) sont une aide dans la construction d'une terminologie mais ne sont pas des termes. C'est

le cas, par exemple, d'un définisseur lexicographique de verbe tel que *le fait de*.

Tout ce qui désigne n'est pas forcément un terme ; à plus forte raison, tout ce qui ne désigne pas mais qui peut faire partie d'une langue spécialisée. Par exemple, dans les jugements des tribunaux, les expressions telles que *vu* ou *considérant* au début des *attendus*.

La notion de "désignation" renvoie à l'énonciation, qui n'est pas forcément textuelle. Le mot en texte (ou vocable) est un mot situé, parmi d'autres ; ce qui en fait un terme, c'est un savoir spécialisé rarement défini dans le texte, toujours présupposé dans la culture partagée par les professionnels du domaine. Les mots en situation textuelle ou discursive ne peuvent avoir que de façon aléatoire un statut de termes au sens de l'ISO 1087-1[4]. Souvent le concept est confondu avec le mot qui en parle[5].

Soyons plus précis : le terme est "le nom donné dans une langue à une entité conceptualisée par une communauté de travail" (Lerat 2009a : 217) où l'on a besoin de dénommer cette conceptualisation.

Ou encore, le terme a pour "fonction de désigner des concepts clairement identifiés à l'intérieur d'un domaine donné" (Sager 2000 : 55).

"Il reste que les points de vue divers posent des problèmes de compatibilité. Ainsi, des syntagmes descriptifs comme *véhicules non motorisés* ou *moyens de transports dépourvus de moteur*, qui rendent service en documentation, ne sont pas des unités lexicales comparables à *véhicule*, *moteur*, *bras* ou *transport*. Chaque option a ses avantages et ses inconvénients. L'option lexicale impose de désambiguïser tout ce qui est susceptible de polysémie (le bras du robot ne va avec le bras du fauteuil que dans les dictionnaires de langues). L'option en faveur des expressions descriptives impose l'acceptation prévue d'une liste de paraphrases reconnues comme équivalentes (*véhicule sans moteur* devra être reconnu comme synonyme de *véhicule non motorisé*)" (Lerat 1995 : 153).

Le problème est de savoir dans quelle mesure une unité de discours est un terme ou non. Il ne suffit pas qu'elle ait une sémantique référentielle. Il faut qu'elle dénote un découpage stable du réel.

Les unités du discours relèvent t-elles du lexique de la langue spécialisée ou de la terminologie ? On constate que les corpus sont riches en

4 Voir la norme ISO-1 1087 où le terme est défini comme "désignation verbale d'un concept général dans un domaine spécifique".

5 On n'insistera jamais assez sur le fait qu'il ne faut pas confondre la conceptualisation d'un domaine avec les discours auxquels elle peut donner lieu (Roche 2009 : 55) Aujourd'hui être c'est être dit et non plus être pensé (Roche 2005).

informations variées et qu'elles sont mélangées. Les corpus spécialisés sont composés d'expressions qui ne sont pas toutes des termes mais qui peuvent donner des indices quant à la structure du modèle conceptuel. C'est le cas, en immunologie, des formulations équivalentes en termes "d'information" "*appears*" et "*is found*" dans les énoncés "Antibody appears in plasma cells" et "Antiboby is found in plasma cells" (Harris 1988 : 39) alors que "to appear" et "to be found" ne sont pas synonymes. Ceci sous-entend que ces expressions linguistiques dépendent du modèle épistémologique. Par exemple, dans la description de l'usinage (fraisage, alésage, tournage …), l'expression "enlèvement de matière" n'est pas un terme mais nous permet de structurer le modèle conceptuel. Qu'est-ce qui relève alors de la terminologie ? Le sens de "matière". *Enlèvement* est une nominalisation d'un verbe de sens très général, mais le couple *enlèvement de matière* est porteur d'une information précise en tant qu'interprétation experte de l'énoncé descriptif.

La terminologie doit être un vocabulaire normalisé le moins ambigu possible et pour cela doit s'ancrer dans la conceptualisation du domaine.

3. Qu'est ce que la terminologie ?

"Le terme *terminologie* est aujourd'hui une forme banalisée tant par le manque de précision dans son emploi que par la confusion sur ses limites d'utilisation. Il convient d'emblée de préciser que le mot *terminologie* est polysémique" (Cabré 1994 : 590).

Le besoin d'une description normée de la signification des termes scientifiques et techniques se développe au XVIIIe siècle dans le contexte des projets encyclopédiques et de la pensée qu'une science est avant tout une langue bien faite (Condillac ; Roche 2005 : 51). L'acception moderne de la terminologie naît avec les travaux d'Eugen Wüster, ingénieur et chef d'entreprise, reconnu comme le père fondateur de la terminologie moderne en posant les principes méthodologiques du travail terminologique.

La terminologie, science des termes, est une pratique et une discipline scientifique et autonome dont le principal objet est de comprendre le monde et de trouver les mots justes pour en parler. Elle requiert pour son étude de puiser à l'épistémologie, la logique et la linguistique.

La terminologie est la dénomination en langue naturelle d'objets scientifiques et techniques et doit être comprise au sens d' "ensembles cohérents de termes" qui reposent sur une "conceptualisation des objets

du monde que partage une communauté de pratiques" (Depecker & Roche 2007 : 112).

3.1. Différences entre terminologie et ontologie

La terminologie et l'ontologie ne relèvent pas des mêmes activités même si leur visée est commune (Roche 2005). "Il faut d'abord distinguer clairement concepts et termes, ce à quoi invite la terminologie classique" (Lerat 2009b : 74).

Alors que la terminologie, qui s'intéresse à la langue de spécialité, donnera des "explications *linguistiques*" (Kocourek 1991 : 180) du terme (au sens de Lerat 2009a : 217 précédemment cité), l'ontologie s'intéressera, dans des langages *formels*, à la "définition des concepts et de leurs relations (spécifications *logiques*)" (Roche 2008 : 64-65).

Dans ces conditions, "l'élaboration de réseaux notionnels suppose donc idéalement la collaboration de linguistes, d'informaticiens et d'experts des domaines de connaissances considérés. Les premiers sont indispensables pour rappeler constamment les contraintes résultant de la combinatoire et de la polyvalence des mots. Les deuxièmes peuvent résoudre des problèmes comme celui des polyhiérarchies ; ainsi, une base de données terminologiques 'intelligente' permet de sélectionner à la demande des sous-ensembles d'informations […]. Les troisièmes sont seuls en mesure de récupérer les attributs cruciaux ; ces derniers sont en petit nombre dans un univers très restreint […]" (Lerat 1995 : 152-153).

3.2. Les tâches de la terminologie

Le rôle du concept est primordial en terminologie. Le concept exprime ce qu'est la chose, c'est "un ensemble unique de caractères" (ISO 1087-1), constitutif de la connaissance des choses.

Ainsi, "les questions que peut soulever la réception des termes sont intimement liées aux concepts et relèvent plus de la démarche scientifique que de la compréhension strictement linguistique" (Mortureux 1995 : 22-23).

La terminologie n'est pas une lexicologie des discours plus ou moins spécialisés telle qu'elle peut être réalisée dans des études lexicométriques à l'aide de "patrons linguistiques".

Elle n'est pas non plus purement textuelle car on a besoin de "ressources externes" à cause du "caractère elliptique de la formulation en langues

naturelles" (Daladier 1990 : 59) pour comprendre le texte et parce que "les corpus textuels de spécialité vieillissent rapidement" (Lino 2006 : 510).

Il est nécessaire de valider des définitions par le genre et la différence en les reliant à des définitions encyclopédiques qui donnent des descriptions du terme (Lerat 2009a) et de valider les dénominations en délimitant ce qui relève du lexique et ce qui relève de la terminologie

Il conviendra enfin de donner des définitions scientifiques. On notera que divers points de vue peuvent porter sur l'objet : un point de vue fonctionnel et un point de vue structurel. Dans les deux cas, il faudra dire ce qu'est la chose – sa nature – et la perception que l'on en a – sa description. Dans la partie formelle de la terminologie, la nature de la chose se traduit en genre prochain et la perception que l'on en a est rendue à travers les propriétés et les attributs valués.

3.3. Vers l'ontoterminologie

L'ontoterminologie, "terminologie dont le système notionnel est une ontologie formelle, insiste sur l'importance des principes épistémologiques qui président à la conceptualisation du modèle – c'est l'ontologie dans sa définition première. Elle insiste également sur la nécessité d'une approche scientifique de la terminologie où l'expert joue un rôle fondamental – c'est l'ontologie dans ses définitions plus récentes où la logique et les langages de représentation des connaissances tiennent une place prépondérante. Enfin, elle met en relation le modèle conceptuel et les termes (d'usage ou normés) qui en parlent, tout en distinguant les définitions formelles des concepts (spécifications logiques) des définitions en langue naturelle des termes (explications linguistiques)" (Roche 2008 : 70).

Il s'agira de définir par le générique immédiatement supérieur "*Nächstoberbegriff*" (Wüster, 1985 : 30) et éventuellement d'élaborer des "controlled vocabularies […] which consist of well-defined or standard concepts corresponding to words and phrases in the domain" (Friedman *et al.* 2002 : 226) en leur associant dans la mesure du possible base de données terminologiques, base de données textuelles et base de données iconiques.

4. Dire n'est pas concevoir

On a pu penser, au vu des résultats et des premiers succès de l'informatique linguistique, qu'il serait possible de construire des ontologies à partir des textes. Certains ont cru pouvoir affirmer que "[…] le travail

scientifique est considéré comme en grande partie constitué par du langage, plus spécialement par des **textes** et la connaissance scientifique est elle-même considérée comme une information conceptuelle obtenue à partir de textes" (Slodzian 1995 : 14), réduisant par là même, lorsqu'ils ne les confondent pas, les concepts à des mots. Or "on ne voit jamais personne devenir médecin par la simple étude des recueils d'ordonnances" (Aristote 1997 : X, 10, 1181b). En effet, aucun corpus ne comporte en lui-même toutes les connaissances nécessaires à sa compréhension. Et comme le précise Marie-Claude l'Homme (2004 : 223 ; 2008), le traitement des textes spécialisés requiert des ressources externes. Non seulement "aucun corpus textuel n'explicite toutes les connaissances que sa lecture présuppose" (Lerat 2009b : 80) mais tout texte de spécialité nécessite une connaissance minimum du domaine. Même si la terminologie est mobilisée au sein des discours et des pratiques langagières, ces démarches s'intéressent en premier lieu aux unités linguistiques qui dénotent les choses plus qu'aux choses elles-mêmes.

Il a été montré à plusieurs reprises que la structure lexicale ne se superpose pas avec la structure conceptuelle (Rastier 1995, 2004 ; Roche 2007 ; Calberg *et al.* 2008a ; Desprès et Szulman 2009). Parler de la chose et définir ce qu'elle est ne relèvent pas des mêmes activités (Figure 1). Il faut donc décrire les deux activités distinctes et complémentaires dans la construction d'une terminologie

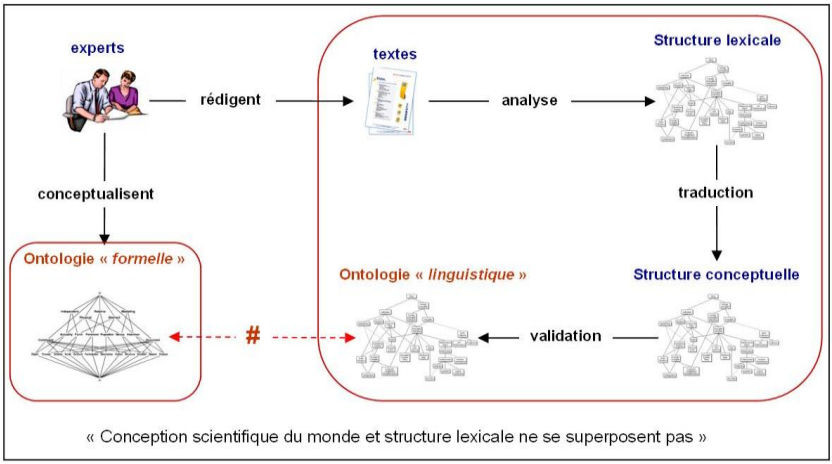


Figure 1. Dire n'est pas concevoir (Roche 2007)

Par exemple, "c'est […] bien à partir de l'expertise et de la réalité de la parole quotidienne de l'expert et non à partir des textes que les listes sur lesquelles nous avons travaillé ont pu être obtenues" (de Vecchi & Estachy

2009 : 42). Ces auteurs précisent dans le même article que "Ces deux disciplines (ndlr terminologie et ontologie) ne s'excluent pas l'une de l'autre, elles se complémentent dans deux volets distincts du traitement de ce que les acteurs activent dans l'univers de la pensée : le concept" (de Vecchi & Estachy 2009 : 42).

Ainsi le concept, élément de pensée, est une connaissance portant sur une pluralité de choses distinctes répondant à une même loi. Il reste à le définir de manière formelle, ce qui pourra effectuer en deux étapes. La modélisation semi-formelle d'un domaine a pour objectif d'identifier, avec les experts, les concepts du domaine et les différentes relations qui les lient pour obtenir un réseau conceptuel. Les concepts ayant été identifiés, nous pouvons les organiser sous la forme d'une ontologie formelle.

Il reste aux experts à identifier ensuite les termes d'usage et les termes normés qui seront associés dans une dernière étape aux concepts de l'ontologie correspondants.

5. Vers une nouvelle construction des terminologies ?

5.1. Objectif et principes

"Il convient de se rappeler que tout travail terminologique devrait être fondé sur les notions et non sur les termes" (Felber 1984).

Notre objectif est de réaliser des terminologies qui soient consensuelles, partageables, réutilisables, cohérentes et computationnelles (exploitables par des programmes informatiques). Ces propriétés dépendent directement des propriétés du système notionnel, c'est-à-dire de la modélisation du domaine sur laquelle repose la signification des termes. Une représentation formelle de cette modélisation permet de garantir ces propriétés. La notion d'ontoterminologie (Roche 2008, 2009), introduite à de telles fins, apporte un regard nouveau sur la terminologie en insistant sur la nécessité de distinguer, pour mieux en clarifier les rapports, les deux dimensions de la terminologie, à savoir la dimension conceptuelle et la dimension linguistique. Une "adéquation est nécessaire entre la connaissance et la langue, l'une ne devant jamais précéder l'autre, et inversement" (Cottez 1994 : 688). Les termes (dénominations du concept) doivent être motivés lexicalement et sémantiquement pour permettre aux experts du domaine d'identifier ou de classer les objets de leur domaine de connaissance, voire d'identifier un nouvel objet sur lequel travailler. L'ontologie permet de comprendre immédiatement la place du concept, et les choses ont alors un rapport "à nous et entre elles" (Condillac 1780). Lavoisier avait déjà écrit

qu'une "nomenclature nouvelle, pourvu qu'elle ait été entreprise sur de bons principes ; pourvu que ce soit une méthode de nommer plutôt qu'une nomenclature, elle s'adaptera naturellement aux travaux qui seront faits dans la suite ; elle marquera d'avance la place et le nom des nouvelles substances qui pourront être découvertes et elle n'exigera que quelques réformes locales et particulières" (Lavoisier III, 17 et I, 186).

5.2. Démarche

La modélisation d'un domaine et la rédaction de discours à laquelle elle peut donner lieu sont deux activités différentes. La modélisation construite directement par les experts à l'aide d'un langage formel et la modélisation construite à partir de textes ne sont pas isomorphes (Roche 2007). De plus, une modélisation construite à partir de documents est dépendante du corpus et ne vérifie pas l'ensemble des propriétés recherchées. C'est pourquoi nous mettons l'accent sur la conceptualisation du domaine (démarche onomasiologique) et sur les principes épistémologiques qui la guide. Dans ce cadre, la présence des experts est indispensable.

La modélisation semi-formelle d'un domaine a pour objectif d'identifier, avec l'aide des experts, les concepts du domaine et les différentes relations qui les lient : relation de généralisation-spécialisation, relation partitive, relation fonctionnelle, etc. Le résultat est un réseau conceptuel. Les concepts ayant été identifiés, nous pouvons les organiser sous la forme d'une ontologie formelle, c'est-à-dire sous la forme d'un arbre de concepts liés par la relation de généralisation-spécialisation.

Afin d'aider les experts dans le choix des termes d'usage, c'est-à-dire des termes métier utilisés en discours oraux ou dans la rédaction de documents scientifiques et techniques, des termes candidats peuvent être générés automatiquement à partir d'un corpus de référence. Les experts identifient alors les termes d'usage (pour cela, on pourra se servir des termes candidats générés lors de la phase précédente) et les termes normés. Les termes d'usage sont les termes utilisés en discours (oraux ou écrits). Ils permettent de prendre en compte la diversité langagière. Ils peuvent être polysémiques. Les termes normés sont des termes univoques qui sont soit en usage, soit prescriptifs et qui sont retenus comme dénomination aux fins de l'ontologie.

La dernière étape consiste à mettre en regard, c'est-à-dire à associer, les termes d'usage et les termes normés avec les concepts de l'ontologie correspondants (Figure 2).

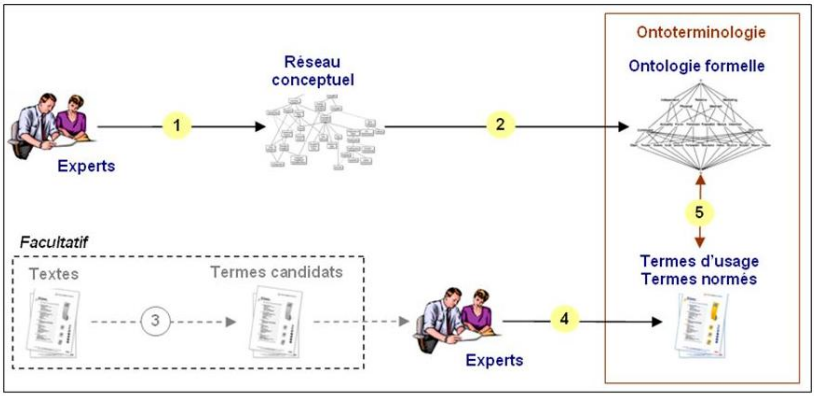


Figure 2. Les étapes de notre méthode

Entre l'ontologie et la terminologie, une étape nécessaire est le choix de termes normés qui doivent être justifiés tant d'un point de vue lexicale, syntaxique et sémantique (c'est-à-dire au regard du système conceptuel). Par exemple, un "amortisseur arrière de moto" se dit dans telle usine italienne aussi bien "ammortizzatore posteriore" que "monoammortizzatore" (Bertaccini *et al.* 2006 : 324), mais c'est le second qui a le statut de "terme technique spécifique-standardisé". De façon générale, un terme est nécessairement validé par son usage dans une "communauté de pratiques" (Depecker & Roche, 2007 : 112). Cela vaut pour un "jargon d'entreprise" comme pour l'ISO.

6. Une démarche éprouvée[6]

6.1. Le rôle des experts

On ne peut pas se passer des connaissances et du savoir des experts dans la construction d'une terminologie. Leur rôle est fondamental dans la conceptualisation et la représentation des objets de leur domaine. Pour ce faire, un échange en langue est nécessaire pour partager les connaissances et le savoir des experts. Mais il est important de ne pas tomber dans les pièges de la langue. En effet, les variations d'usage et de pratiques entre les différents acteurs, qu'ils soient ingénieurs, exploitants ou techniciens étant inhérentes à tout discours, il faudra arriver à ce que les experts témoignent de la conceptualisation et de la représentation des objets de leur domaine

6 Cette dernière partie s'appuyant sur des exemples, les concepts seront signalés entre chevrons ouvrants et fermants <Concept> et les termes, qu'ils soient d'usage ou normés entre guillemets "terme".

et non de la dimension linguistique de leur activité pour tenter de trouver un consensus entre les experts.

"La compréhension de figures de rhétorique, telles que l'ellipse ou la métonymie fréquentes dans les documents scientifiques et techniques, nécessite que les locuteurs s'accordent sur ce même extralinguistique qui par définition n'appartient pas à la langue" (Roche 2008 : 3) et "se référer à la conceptualisation du domaine peut être une autre manière d'apporter des éléments de réponse" (Calberg *et al.* 2008a : 133).

De façon générale, "l'interrogation des spécialistes du domaine peut remplacer l'introspection du lexicographe" (Thoiron *et al.* 1996 : 513).

Pour illustrer ces propos, prenons en exemple dans le domaine de l'hydraulique, le cas des jantes où la langue peut induire en erreur. De prime abord, les experts semblent distinguer à travers leurs discours trois types de jantes désignées par les expressions "jantes feuilletées", "jantes massives" et "jantes soudées". En s'appuyant sur des relations linguistiques comme l'hyperonymie ou l'hyponymie entre ces expressions, on peut être amené à structurer les différents concepts de <Jante> sous la forme du réseau suivant (Figure 3) :

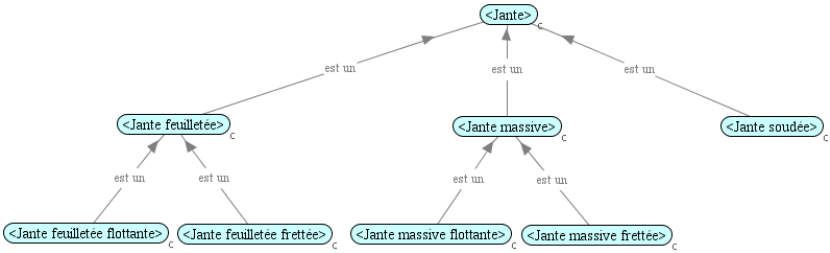


Figure 3. Réseau conceptuel[7] d'une jante

Ce réseau conceptuel présenté aux experts n'a pas été validé dans la mesure où il ne traduit pas le fait qu'une <Jante soudée> est une sorte de <Jante massive>. L'expression "jante massive soudée" n'existe pas, ils parlent, par économie de la langue (ellipse), de "jante soudée" dans la mesure où une jante soudée est nécessairement massive. Ceci a donné lieu à la construction d'une nouvelle conceptualisation du domaine (Figure 4) qui, tout en entrainant le consensus des experts[8], a permis d'introduire un terme normé calqué sur l'identifiant du concept pour parler de <Jante massive soudée>. "Ces termes normés, s'ils n'ont pas à être imposés, sont

7 Réalisé à l'aide de iMap, éditeur de réseaux conceptuels (Condillac).

8 Cette ontologie a été validée par les acteurs de différentes communautés de travail (ingénieurs, exploitants, techniciens).

indispensables à la désignation du système notionnel. Ils participent également à l'identification et à la définition des termes d'usage" (Roche 2008 : 18).

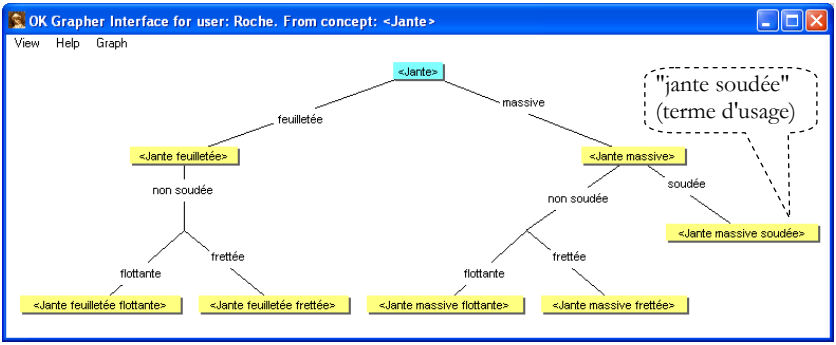


Figure 4. Représentation conceptuelle[9] d'une jante

Cet exemple illustre parfaitement les deux dimensions de la terminologie et insiste une fois encore sur l'importance de conceptualiser le domaine de travail avant de relever les divers pratiques langagières.

Prenons une autre illustration (Figure 5). Le terme "dispositif d'accrochage par clés d'aronde", normé parce que par sa simple lecture on comprend la position du concept dénoté dans le système notionnel, ne sera pas utilisé pas les experts car trop long. Lorsqu'ils parlent d'un <Dispositif d'accrochage par clés>, on a soit un <Dispositif d'accrochage par clés en T> soit un <Dispositif d'accrochage par clés d'aronde>. Dans ce dernier cas, ils parleront de "queue d'aronde", de "clé d'ironde" ou de "queue d'ironde" (sachant que ce terme résulte d'une analogie avec la queue d'une hirondelle).

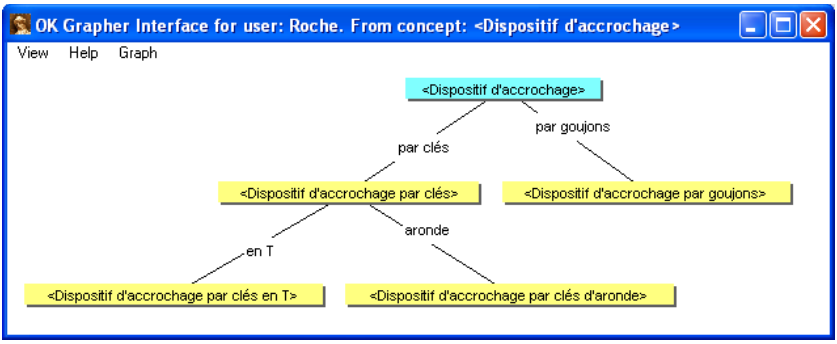


Figure 5. Représentation conceptuelle d'un dispositif d'accrochage

9 Réalisée à l'aide de l'environnement iConcept, éditeur d'ontologies par différenciation spécifique (Roche 2001).

Dans un "système conceptuel structuré, il s'agit surtout d'observer si la forme du terme est différente de celle des autres termes du système, si elle indique des oppositions pertinentes, et rien qu'elles, si elle reflète le degré de différence entre les concepts désignés" (Kocourek 1991 : 226).

6.2. L'intérêt d'un recours aux schémas

"Il y a […] quelque chose d'éternel dans un schème technique … et c'est cette qualité qui reste toujours présente et peut être conservée dans une chose" (Simondon 1989).

Lorsque les experts sont invités à conceptualiser leur domaine, ils ne viennent pas avec des textes. Les documents servant de supports sont avant tout des schémas ou des figures. Et s'ils ont besoin d'expliciter et de clarifier leurs connaissances, ils auront de nouveau recours à des schémas sur le tableau.

Dans le cas de la modélisation d'un groupe hydraulique (Figure 6), on a proposé le terme normé "Organe de protection amont", terme normé et motivé lexicalement et syntaxiquement, car le terme "Organe de protection aval" existe déjà et il manquait un terme pour désigner l'ensemble des organes de protection amont des différents types de turbines hydrauliques. Dans ce cas présent, il s'avère que ce terme normé dénomme également le concept mais ce n'est pas toujours le cas.

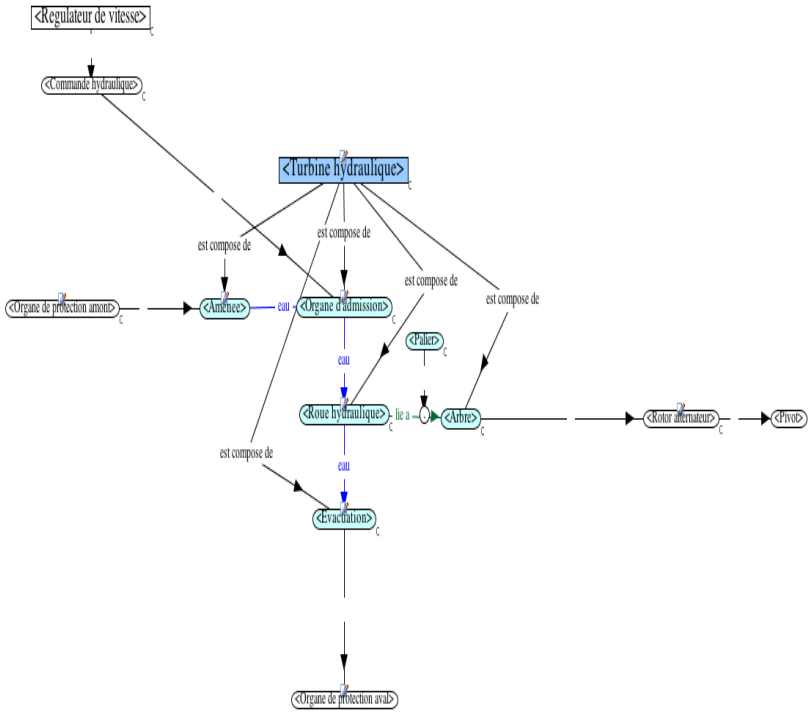


Figure 6. Réseau conceptuel d'une Turbine hydraulique

Comme l'a écrit Pierre Lerat, "la représentation graphique d'un objet est souvent irremplaçable. Pour faire comprendre ce qu'est un outil […], un dessin fera gagner du temps. A plus forte raison, là où une description en langue naturelle aura bien du mal à rendre compte de ce qu'est un vérin, un dessin industriel correct, assisté ou non, fera voir de quoi il s'agit" (Lerat 1995 : 149).

7. Conclusion et perspectives

Le recours à des corpus est nécessaire, mais seulement dans la seconde étape de la construction d'une terminologie ou ontoterminologie.

Nous avons montré l'importance d'une terminologie formelle, base d'un système notionnel ou ontologique, et distincte de la terminologie.

En effet, de notre point de vue, la démarche ontoterminologique que nous avons présentée naît de la recherche pluridisciplinaire d'une représentation

commune d'une réalité face aux variations d'usage des différentes communautés de travail en interaction avec cette réalité.

Il n'est pas question ici d'imposer un langage contrôlé mais de travailler en équipe avec un partage des méthodes.

Il importe de prendre appui sur une conceptualisation commune et consensuelle pour introduire les variations d'usage. Cette conceptualisation permet une meilleure compréhension entre les communautés de travail tout en préservant et valorisant la diversité linguistique.

Bibliographie

Aristote (1997) : *Ethique à Nicomaque*, Traductions et notes par J. Tricot, Vrin, Bibliothèque des Textes Philosophiques – Poche, 578 p

Arnauld Antoine & Nicole Pierre (1992) : *La logique ou l'art de penser*, Notes et postface de Charles Jourdain, Gallimard, collection tel

Bertaccini Franco & Matteucci Alessandra (2006) : "La terminologie d'entreprise et ses contextes d'usage", *Mots, termes et contextes*, Blampain D., Ph. Thoiron et M. Van Campenhoudt éd., Paris, AUF, pp. 317-326

Cabré Maria Teresa (1994) : "Dictionnaires et terminologie", *Meta : journal des traducteurs*, vol. 39 n°4, pp. 589-597

Calberg-Challot Marie, Candel Danielle & Roche Christophe (2008a) : "De la variation des usages au consensus terminologique : vers un dictionnaire de l'ingénierie nucléaire", *Actes de la première conférence TOTh 2007, Terminologie & Ontologie : Théories et applications*, Christophe Roche éd., Annecy, Institut Porphyre, pp. 199-141

Calberg-Challot Marie, Candel Danielle , Bourigault Didier , Dumont Xavier , Humbley John & Joseph Jacques (2008b) : "Une analyse méthodique pour l'extraction terminologique dans le domaine du nucléaire", *Terminology 14:2*, pp. 183–203

Cassin Barbara (2004) : *Vocabulaire européen des philosophies, Dictionnaire des intraduisibles*, Paris, Le Seuil/Le Robert, 1 532 p

(de) Condillac Etienne Bonnot (1780) : *La logique, ou les premiers développements de l'art de penser*, Paris

Cottez Henri (1994) : "Les bases épistémologiques et linguistiques de la nomenclature chimique de 1787", *Meta : journal des traducteurs*, vol. 39 n°4, pp. 676-691

Daladier Anne (1990) : "Aspects constructifs des grammaires de Zellig Harris", Langages 99, pp. 57-84

Depecker Loïc & Roche Christophe, 2007 : "Entre idée et concept : vers l'ontologie", *Langages 168*, La terminologie : nature et enjeux, pp. 106-114

Despres Sylvie & Szulman Sylvie (2009) : "Réseau terminologique versus Ontologie", *Actes de la deuxième conférence TOTh 2008, Terminologie & Ontologie : Théories et applications*, Christophe Roche éd., Annecy, Institut Porphyre, pp. 17-34

Dictionnaire de Philosophie (1995) : Baraquin Noëlla, Baudart Anne, Dugué Jean, Laffitte Jacqueline, Ribes François et Wilfert Joël, Paris, Armand Colin, Collection Dictionnaire, 384 p

Felber Helmut (1984) : *Manuel de terminologie*, Unesco, Paris

Frege Gottlob (1971) : *Ecrits logiques et philosophiques* (Traduction et introduction de Claude Imbert), Paris, Seuil

Friedman Carol, Kra Pauline & Rzhetsky Andrey (2002) : "Two biomedical sublanguages : a description based on the theories of Zellig Harris", *Journal of Biomedical Informatics*, 35-4, pp. 222-235, *www.sciencedirect.com*

Harris Zellig Sabbetai (1988) : *Language and Information*, New York, Columbia University Press (trad. fr. La langue et l'information, A.H. Ibrahim et C. Martinot éd., Paris, Cellule de recherche en Linguistique, http://crl.exen.fr)

ISO 1087-1, 2001. Vocabulaire, ISSN 0335-3931

Kocourek Rostislav (1991) : *La langue française de la technique et de la science : vers une linguistique de la langue savante* (1982), 2e éd., Wiesbaden, Oscar Brandstetter

Lavoisier Antoine (1789) : *Traité élémentaire de Chimie, Mémoires, tableaux et Synonymie de 1787*, réunis dans le 3e tome, Nouvelle édition, Cuchet

Lerat Pierre (1995) : *Les langues spécialisées*, Paris, PUF

Lerat Pierre (2009a) : "La combinatoire des termes. Exemple : Nectar de fruits", *Hermes Journal of Langage and Communication Studies* 42, pp. 211-232

Lerat Pierre (2009b) : "Propositions pour un réseau conceptuel des instruments de mesure oenologiques", *Actes de la deuxième conférence TOTh 2008, Terminologie & Ontologie : Théories et applications*, Christophe Roche éd., Annecy, Institut Porphyre, pp. 73-90

L'Homme Marie-Claude (2004) : *La terminologie : principes et techniques*, Montréal, Les presses de l'Université de Montréal

L'Homme Marie-Claude (2008) : "Ressources lexicales, terminologiques et ontologiques : une analyse comparative dans le domaine de l'informatique", *Revue française de linguistique appliquée*, XIII-1, pp. 97-118

Lino Teresa (2006) : "Contextes et néologie terminologique dans le domaine médical", *Mots, termes et contextes*, Blampain D., Ph. Thoiron et M. Van Campenhoudt éd., Paris AUF, pp. 509-514

Mortureux Marie-Françoise (1995) : "Les vocabulaires scientifiques et techniques", *Les carnets du Cediscor* 3, Les enjeux des discours spécialisés, Presses de la Sorbonne Nouvelle, pp. 13-26

Neveu Franck (2004) : *Dictionnaire des sciences du langage*, Armand Colin, Collection "Dictionnaires", Paris, 320 p

Pruvost Jean (2008) : "Préface", *Néologie et terminologie dans les dictionnaires*, éd. Honoré Champion, Paris

Rastier François (1995) : "Le terme : entre ontologie et linguistique", *La banque des mots*, numéro spécial 7-1995, pp. 35-65

Rastier François (2004) : "Ontologie(s)", *Revue d'Intelligence Artificielle*, vol. 18 n°1, pp. 15-40

Roche Christophe (2001) : "The 'Specific-Difference' Principle : a Methodology for Building Consensual and Coherent Ontologies", Actes de la conference IC-AI'2001, Las Vegas , USA

Roche Christophe (2005) : "Terminologie et ontologie", *Langages 157*, La terminologie : nature et enjeux, pp. 48-62

Roche Christophe (2007) : "Dire n'est pas concevoir", IC 2007 : 18e Journées Francophones d'Ingénierie des Connaissances, Grenoble 2-6 juillet 2007

Roche Christophe (2008) : "Le terme et le concept : fondements d'une ontoterminologie", *Actes de la première conférence TOTh 2007, Terminologie & Ontologie : Théories et applications*, Christophe Roche éd., Annecy, Institut Porphyre, pp. 1-22, 2007

Roche Christophe (2009) : "Faut-il revisiter les principes terminologiques ? ", *Actes de la deuxième conférence TOTh 2008, Terminologie & Ontologie : Théories et applications*, Christophe Roche éd., Annecy, Institut Porphyre, pp. 53-72

Sager Juan Carlos (2000) : "Pour une approche fonctionnelle de la terminologie", *Le sens en terminologie*, Presses universitaires de Lyon, pp. 40-60

Simondon Gilbert (1989) : *Du mode d'existence des objets techniques*, Aubier philosophie, (1958, 1969), 3e édition

Slodzian Monique (1995) : "Comment revisiter la doctrine terminologique aujourd'hui ?, *La banque des mots*, numéro spécial 7-1995, pp. 11-18

Thoiron Philippe, Arnaud Pierre, Henri Béjoint et Boisson Claude Pierre (1996) : "Notion 'd'archi-concept' et dénomination", *Meta : journal des traducteurs*, vol. 41, n°4, pp. 512-524

(de) Vecchi Dardo & Estachy Laurent (2009) : "Pragmaterminologie : les verbes et les actions dans les métiers", *Actes de la deuxième conférence TOTh 2008, Terminologie & Ontologie : Théories et applications*, Christophe Roche éd., Annecy, Institut Porphyre, pp. 35-52

Wüster Eugen (1985) : *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie* (1979), Copenhague, Ecole des Hautes Etudes Commerciales

A propos des auteurs

Marie Calberg-Challot
(1) 88 ter, boulevard de Belgique
78 110 Le Vésinet
marie.calberg-challot@orange.fr

(2) Equipe Condillac – Laboratoire Listic
Campus Scientifique
73 376 Le Bourget du Lac cedex
http://ontology.univ-savoie.fr

Pierre Lerat

(1) 34 rue Notre-Dame de Recouvrance
45 000 Orléans
pierre.lerat@wanadoo.fr

(2) Equipe Condillac – Laboratoire Listic
Campus Scientifique
73 376 Le Bourget du Lac cedex
http://ontology.univ-savoie.fr

Christophe Roche

Equipe Condillac – Laboratoire Listic
Campus Scientifique
73 376 Le Bourget du Lac cedex
christophe.roche@univ-savoie.fr
http://ontology.univ-savoie.fr

Extraction des connaissances orientées évolution dans les textes de brevets

Kata Gábor, François Rousselot, François de Bertrand de Beuvron, Denis Cavallucci, Dildar Wu

Résumé : Nous présentons une nouvelle approche pour l'extraction de connaissances à partir des brevets à l'usage des ingénieurs experts d'un domaine. Celle-ci est basée sur la Conception Inventive (CI dans la suite). Une ontologie de la CI aide à préciser la tâche et à définir les informations à extraire. Les résultats obtenus peuvent être directement utilisés par un concepteur, en lui donnant une vision élargie des inventions déposées dans son domaine. Ils seront également exploités comme données d'entrée d'un logiciel développé au sein du LGECO, nommé TrizAcquisition qui permet de mener une étude de CI détaillée sur un artefact donné.

Mots-clés : Extraction d'information, extraction de connaissances, brevet, invention, Conception Inventive, TRIZ

1. Introduction

1.1. Analyse des brevets

Le travail des ingénieurs en Conception Inventive lors de la conception d'une invention consiste souvent à améliorer un artefact existant. Pour ce faire, les ingénieurs ont besoin de connaître l'état de l'art, et en particulier les caractéristiques des produits proches qui existent dans le domaine. Ils ont donc besoin d'accéder à des bases de données de brevets. Celles-ci inventorient la quasi-totalité des brevets existants. Elles disposent de fonctions de recherche limitées basées principalement sur des mots clés. Il est urgent de développer des outils de Traitement Automatique des Langues pour permettre de lancer des requêtes plus fines et obtenir des résultats pertinents et sémantiquement structurés.

1.2. Le modèle de connaissances : la Conception Inventive

Nous travaillons avec des experts praticiens de la "Conception Inventive", issue de la TRIZ[1]. La TRIZ, créée dans les années 50 en ex-URSS et affinée jusqu'à la disparition d'Altshuller (Altshuller 1999) en 1988, voit le processus d'invention comme la résolution d'une contradiction. Nous décrivons ci-dessous les éléments de cette théorie pertinents pour l'extraction de connaissances, et renvoyons le lecteur à (Zanni *et al.* 2009) pour une présentation plus générale.

Nous avons élaboré et formalisé un modèle des connaissances de la Conception Inventive qui rend compte de notre point de vue sur cette théorie et qui en propose une extension. La TRIZ est une théorie de l'évolution des artefacts, elle part du postulat que tout objet conçu par l'homme est le résultat d'une évolution guidée par des lois objectives et la création d'une invention comme résultant d'une impossibilité pour l'objet d'évoluer en cohérence avec ces lois. Une de ses caractéristiques qu'on veut faire évoluer est bloquée par un conflit d'origine technique ou physique. La théorie propose alors une méthode pour formuler précisément ce conflit en termes d'une contradiction qui peut s'énoncer de la manière suivante :

Soit trois paramètres P1, P2, et P3 de l'objet à modéliser. P1 peut prendre les valeurs opposées A ou . or, si P1 prend la valeur A, P2 est améliorée,

1 TRIZ propose une démarche très dirigée pour concevoir de façon inventive en exploitant les analogies entre solutions issues de domaines différents et induit une vue très spécifique sur le processus de l'invention.

mais P3 est dégradée, inversement, si P1 prend la valeur , P2 est dégradée, alors que P3 est améliorée.[2]

Dans une application, une fois qu'un ensemble réduit de contradictions importantes est clairement identifié, TRIZ fournit à l'inventeur des techniques et des bases de connaissances qui lui permettent de générer des Concepts de Solutions. La solution sera qualifiée d'*inventive* si elle est nouvelle et ne constitue pas un compromis[3].

La TRIZ dans sa version actuelle ne concerne que des problèmes ne comportant que peu de contradictions. Nous avons étendu la théorie à des problèmes complexes pouvant comporter jusqu'à plusieurs centaines de contradictions. Dans le but de développer un maximum l'aide informatique apportée pendant la résolution, nous avons construit une ontologie des concepts de la théorie étendue[4] (Cavallucci *et al.* 2009) qui sera appelée Conception Inventive ou CI dans la suite. Cette ontologie va permettre ici, d'une part de représenter un ensemble de brevets traitant du même artefact avec ses évolutions successives comme un ensemble structuré de brevets reliés par des liens avec une sémantique précise, d'autre part d'accéder dans chaque brevet aux connaissances exprimant des concepts pertinents en CI.

Un expert TRIZ, lorsqu'il analyse un problème va construire un modèle de ce problème. La structure de ce modèle dépend de la TRIZ, et contiendra des données qui dépendent du domaine d'application. Ces données seront, en général, fournies par un expert du domaine. Le but de notre outil d'extraction dans les brevets est d'aider à la collecte de ces données dépendant du domaine, et de leur mise en forme dans le modèle du problème. Associé au logiciel TrizAcquisition conçu par le LGECO pour assister l'application de la méthode de CI, cet outil permet donc d'automatiser ou d'aider le travail de deux experts : de CI et du domaine.

Après un bref aperçu de l'état de l'art en section 2, nous donnons en section 3 une explication des principaux concepts de l'ontologie TRIZ nous nous positionnons parmi les principales approches d'extraction de connaissances à partir des brevets. Nous présentons ensuite les tâches à effectuer que nous avons identifiées et les difficultés de chacune d'elles. En

2 Exemple de l'éolienne: P1 = taille des pales ; A = grand; = petit, P2 = puissance générée, P3 = résistance aux vents violents

3 Exemple : une éolienne de taille moyenne serait un compromis; mais une éolienne avec un axe vertical est une évolution inventive, car elle résout la contradiction.

4 La TRIZ n'a jamais été formalisée et l'ontologie que nous proposons est la seule existant à notre connaissance qui précise le sens des termes principaux utilisés par les experts TRIZ et qui permet de formaliser les modèles proposés par cette théorie.

conclusion, nous donnerons des pistes pour réaliser une chaîne de traitement des brevets en vraie grandeur.

1.3. Spécificités de la tâche et difficultés rencontrées

Nous avons choisi de travailler sur des brevets rédigés en anglais plus faciles d'accès. Mais les difficultés rencontrées ne sont pas propres à l'anglais et les méthodes que nous avons choisies sont adaptables à d'autres langues.

Nous avons constitué un corpus initial représentatif de la langue des brevets, qui comprend 100 brevets électroniques publiés entre 2000 et 2009, téléchargeables sur le web (sources *www.googlepatents.com* et *www.patents.com*) appartenant à divers domaines. Nous avons collecté ces brevets en faisant des requêtes avec des mots clés génériques non liés à un domaine particulier. La collection de textes a été convertie en XML, avec une annotation structurelle.

Les principales difficultés de l'accès aux informations contenues dans les brevets viennent de la spécificité de ces textes. Il s'agit de documents relativement courts généralement structurés en paragraphes avec des titres de section standard. Ils sont caractérisés par la présence de phrases longues comportant de nombreuses répétitions et des énumérations dans un jargon très particulier, ni proche de la langue générale, ni de la langue scientifique. Une autre source de difficulté a pour origine le fait que le but visé par le dépôt d'un brevet est plus d'ordre juridique (protection des droits intellectuels) qu'explicatif (explications souvent volontairement confuses). Pour toutes ces raisons, la plupart des outils linguistiques développés pour des textes généraux ou descriptifs, appliqués aux textes de brevets ont de piètres performances.

Par rapport à l'objectif visé ici, un autre problème survient du fait que le modèle abstrait fournit par la CI ne correspond généralement pas à la logique discursive habituellement suivie dans les brevets.

2. État de l’art

S’il existe de nombreuses applications en extraction d'informations et de connaissances, relativement peu sont spécialisées dans les brevets. Un important besoin existe cependant dans ce domaine, car un ingénieur expert doit lire des dizaines, voire des centaines de brevets pour se faire une idée de l'état de l'art dans son domaine d’activité spécifique.

Généralement, les applications spécialisées dans le traitement des brevets utilisent des méthodes hybrides (statistiques–linguistiques). La plupart opèrent un prétraitement linguistique qui comprend en général tokenisation, étiquetage, délimitation des phrases, et généralement reconnaissance des entités nommées. Elles contiennent généralement un module de règles écrites à la main par des experts du domaine soit par des linguistes soit conjointement. Elles utilisent également des méthodes statistiques pour assurer une certaine robustesse. Soit les connaissances humaines peuvent sont exploitées de manière automatique grâce à un module de règles à l'intérieur du système, soit elles servent de base d'apprentissage pour des algorithmes statistiques. Certains systèmes ne visent pas à automatiser complètement le processus : ils font appel à l'intervention extérieure d'un expert qui évalue et trie les résultats fournis par le logiciel.

Une caractéristique importante qui différencie les approches est la nature des informations cherchées: informations de domaine, termes du domaine, informations sur l'évolution de l'artefact. Quant à celles qui s'appuient sur des ontologies, celles-ci ont des rôles divers : ontologies domaines, génériques, structurelles ou non, statiques ou décrivant des connaissances dynamiques. Certains systèmes utilisent une ontologie domaine fournie extérieurement, malheureusement, celle-ci peut alors ne pas être adaptée à la tâche d'extraction. De plus, l'adaptation de la méthode à un nouveau domaine sera parfois problématique, car elle dépendra de l'existence d'une ontologie du nouveau domaine. À partir d'une ontologie, il reste à construire semi automatiquement une Res-source Termino-Ontologique à partir d'un corpus de brevets, afin de relier les concepts aux différentes manières de les exprimer. C'est une tâche concevable, mais qui nécessite du temps et des efforts humains importants . C'est pourquoi nous avons choisi une solution alternative : celle d'automatiser le plus possible la démarche complète, tout en gardant la possibilité de compléter le système par une ressource ontologique.

Nous donnons ici un aperçu sur l'état de l'art dans l'analyse des brevets ainsi que des méthodes existant en extraction de connaissances. Certaines recherches, plutôt linguistiques, par exemple (Guyot *et al.* 2004) se sont intéressées aux caractéristiques qui définissaient le genre des textes des brevets, mais la plupart relèvent du domaine du Traitement Automatique des Langues. (Feldman *et al.* 1998 présentent une approche d'extraction de connaissances à partir de textes non structurés applicable aux brevets. Après un prétraitement linguistique de base, ils produisent d'une liste de mots candidats à être des mots clés. Un filtrage morpho-syntaxique et un filtrage par pertinence statistique sont appliqués à la liste, qui est ensuite

proposée à l'utilisateur pour l'aider à construire la taxonomie du domaine de l'artefact.

Parmi les traitements spécialisés dans les brevets qui s'appuient sur la structure et des propriétés spécifiques du document, il faut mentionner (Ghoula *et al.* 2007) qui présentent une chaîne de traitements réalisant une annotation sémantique automatique des brevets, grâce à une ontologie structurelle et à une ontologie du domaine, dans leur cas la biologie. (Agatonovic *et al.* 2008) utilisent la plate-forme GATE (Cunningham *et al.* 2002) pour l'annotation. L'objectif est de créer un outil robuste et efficace pour pouvoir traiter de grandes quantités de brevets. L'analyse produit également des annotations structurelles et des annotations internes : éléments, unités de mesure et autres types d'entités nommées.

Ces deux travaux définissent une chaîne de traitement robuste, mais aucune n'accède à des connaissances propres au processus d'invention, car elles ne prennent pas en compte les évolutions. Patexpert (Mille *et al.* 2008) est un système commercial qui utilise un système de règles et qui produit un résumé automatique en langue naturelle de la partie "Revendications" (Claims) qui est seule prise en compte. (Sheremetyeva 2003) propose un système hybride (statistique-symbolique). Son système effectue une analyse linguistique par une grammaire de dépendances, basée sur un lexique très riche construit à partir de 1 000 brevets annotés manuellement qui contient : informations morphologiques, structures argumentales des verbes et des rôles thématiques des compléments, fréquence des structures argumentales, classification sémantique. L'idée d'exploiter les propriétés linguistiques des brevets et de se baser sur leur spécificité est intéressante, mais le travail est centré sur les connaissances descriptives.

La plupart des approches apparentées à l'extraction des connaissances, considérées jusqu'ici, sont des approches qui sont centrées sur le contenu du brevet et qui tendent soit à interpréter le contenu par rapport à une ontologie du domaine, soit à accéder aux termes du brevet et à la description statique qu'il contient. Elles visent à faciliter la consultation des brevets par des experts connaissant le domaine de l'artefact. Les approches que nous allons voir ne sont pas centrées sur les connaissances statiques, mais opèrent une sélection d'entités orientées évolution dans les textes de brevets. (Goujon 1999) décrit un système de veille technologique qui utilise la méthode dite de l'exploration contextuelle pour extraire des expressions correspondant à quelques notions qu'elle considère intuitivement comme pertinentes : par exemple *changement*, *utilisation* ou *amélioration*. Elle ne s'appuie sur aucun modèle de connaissances. (Cascini *et al.* 2007) qui s'inspirent de la TRIZ est le travail le plus proche du nôtre

pour ses objectifs et pour son utilisation d'outils de TAL. Leur approche est toutefois distincte de la nôtre. D'une part, elle n'est pas basée sur une ontologie formalisée définissant les concepts utiles, d'autre part, elle est centrée sur le repérage des liaisons (qu'ils appellent "fonctions") entre les éléments de l'artefact. Le résultat du traitement est une représentation du brevet sous forme de triplets (élément, fonction, élément) sélectionnés dans la liste de tous les triplets (Sujet, Verbe, Objet) du texte. Cette représentation, quoique consistante en elle-même, ne satisfait généralement pas les experts. L'extraction des contradictions est à peine abordée dans l'article, la notion de contradiction n'y est pas définie de façon formelle et ne fait pas intervenir d'autres concepts tels : paramètre et valeur (voir section suivante).

3. L'extraction des connaissances de la CI

3.1. L'ontologie

L'ontologie de la Conception Inventive est générique. Elle permet parmi les éléments et les sous-éléments d'un artefact qui sont en général nombreux et agencés de façon complexe, de sélectionner ceux qui entrent en jeu dans une évolution possible. Le résultat visé par le système est donc la population dans un domaine donné de cette ontologie générique pour construire un modèle du domaine, très différent des ontologies domaines statiques habituelles, comportant uniquement des informations concernant des évolutions de paramètres et des conséquences d'évolution de ceux-ci.

Notre approche essaie de suivre la démarche de l'expert en CI lorsqu'il étudie les brevets dans la phase initiale et qu'il repère dans les textes de ceux-ci les **paramètres** qui vont lui servir à mieux cerner son **problème**. Il doit savoir quel est le problème à l'origine de l'invention et quelle **solution partielle** ce brevet propose. Chaque problème est à l'origine d'une ou plusieurs **contradictions** que résout le brevet. La rhétorique sous-jacente aux textes des brevets sert donc à exprimer des informations telles : "Considérant tel artefact, tel défaut s'est révélé, le présent brevet apporte une amélioration qui supprime ce problème." Les concepts de notre ontologie que nous précisons rapidement ci-dessous vont permettre de prendre en compte cette rhétorique.

Plus précisément, un problème exprime les caractéristiques insatisfaisantes d'un système : il est généralement décrit par des expressions de jugement négatif. Un brevet propose une solution partielle à celui-ci, qui est exprimée par l'expression d'un progrès, d'une amélioration. Les éléments sont des composants du système. Les éléments qui nous intéressent sont

ceux qui possèdent des paramètres dont les **valeurs** changent au cours de l'amélioration apportée par le brevet. Il y a deux types de paramètres : les **paramètres d'action** sur lesquels on peut agir et les **paramètres d'évaluation** dont on peut constater le changement de valeur.

En fonction des éléments de l'information à retrouver, notre démarche comporte les étapes suivantes : 1) trouver le problème concernant l'artefact auquel le brevet propose une solution, 2) trouver la solution (partielle) ou l'amélioration apportée par l'invention.

Pour ce faire, il a fallu se faire une idée de la façon dont ces textes expriment les informations en question, c'est-à-dire trouver des régularités entre la structure informationnelle et la structure (morpho)syntaxique du texte. Nous avons ensuite testé des algorithmes d'extraction d'information différents et les avons adaptés. Nous avons finalement conçu une méthode hybride : statistique (filtrage et de l'extraction statistique) combinée avec un module de règles basées sur une analyse linguistique. Dans la suite, nous présentons les tâches à effectuer.

3.2. Filtrage des paragraphes

La contradiction résolue par le produit breveté se formule comme une amélioration par rapport à un état antérieur. Ainsi, nous sommes amenés à chercher des parties de texte qui contiennent une référence au déroulement temporaire d'un changement, de préférence complétée par une ou plusieurs expressions de jugement de valeur (jugement négatif sur les caractéristiques qui constituaient le problème, positif sur les paramètres après l'amélioration apportée, (voir les exemples plus loin). L'opposition entre les propriétés des autres systèmes et celles du nouveau système peut amener à découvrir des contradictions. L'identification des parties de texte susceptibles de contenir des valeurs opposées permettra de filtrer les textes à traiter et ainsi de réduit l'espace de recherche où s'appliqueront des méthodes d'extraction plus ciblées.

Structure du document

Certaines généralisations peuvent être formulées sur la structure des parties pertinentes. L'étude des textes a montré que de nombreuses comparaisons sont explicitées dans les parties appelées *Background* ou *Summary of the Invention* qui constituent souvent des sections séparées, mais malheureusement pas toujours. On y trouve généralement des phrases ou des paragraphes entiers qui détaillent les inconvénients des autres systèmes. Les expressions utilisées dans cette partie du texte sont reprises plus tard lors de la description de l'invention. Les paragraphes pertinents

ne correspondant pas toujours à des sections titrées nous avons conçu une méthode qui filtre ces paragraphes par leur contenu.

Les paragraphes des parties pertinentes s'organisent typiquement autour des thématiques suivantes :

- -a) description du fonctionnement des autres systèmes,
- -b) difficultés, inconvénients ou risques encourus lors de l'usage des autres systèmes,
- -c) objectifs de l'invention, notamment ceux qui visent à éliminer les problèmes mentionnés auparavant,
- -d) description partielle du système breveté, qui explique comment il arrive à résoudre ces problèmes.

Un paragraphe parle, en général, soit des autres systèmes, soit de l'invention et il est très rare de voir des paragraphes dans lesquels il y ait à la fois des phrases qui parlent de l'état de l'art et des phrases qui parlent de l'invention. Nous classifions les paragraphes en trois catégories : 1) ceux qui parlent de l'état de l'art (a-b), 2) ceux qui décrivent la solution apportée par l'invention (c-d), et 3) les autres, c'est-à-dire ceux qui ne sont pas intéressants pour nos objectifs.

Comme les paragraphes de type a) - d) présentent tous des traits linguistiques spécifiques, ils peuvent être repérés automatiquement sur la base de ces spécificités (p. ex. les paragraphes qui parlent de l'état de l'art utilisent le pluriel beaucoup plus fréquemment que les autres parties du texte : "*Most systems are…* "). Ces marqueurs linguistiques caractéristiques ouvrent la voie à plusieurs méthodes pour classifier les paragraphes automatiquement dans une des trois catégories suivantes : État de l'art/Invention/Autre. L'objectif est de filtrer les paragraphes et ainsi réduire la quantité de texte à analyser, tout en préservant le maximum d'informations pertinentes, mais aussi de définir la référence des propriétés énumérées dans le document, de créer le lien entre un élément et un de ses paramètres et ainsi à savoir entre quels sous-systèmes il faut chercher les oppositions qui décrivent la contradiction.

Marqueurs linguistiques

Les régularités cherchées correspondent à des concepts génériques, il n'est donc pas utile de disposer d'une ontologie du domaine. En effet, nous avons observé qu'un nombre relativement restreint de structures linguistiques est utilisé de manière répétitive pour décrire l'état de l'art et

l'apport de l'invention dans le domaine. Voici quelques exemples pour montrer les régularités :

État de l'Art – inconvénients des systèmes "typiques" :

Injection molding is typically done in molds which operate at high temperatures...
Conventionally, the fluid cover stock material enters the mold cavity...
Most injection molds comprise halves that mate to define an internal cavity...
Typical molds include means to heat the molds at numerous point...

However, the known molds of this type still require substantial changeover time…
This presents disadvantages both in cost and in the downtime required to change over a molding machine from one part to another...
Although a two level stack mold can produce product at roughly twice the rate possible with a non-stacked mold, mold costs are considerably higher because of…

Résumé de l'invention – solutions :

A new retractable pin mold for golf balls has now been discovered which alleviates a number of the problems of conventional golf ball retractable pin molds.
An object of this invention is to provide improved quick-changeover cavity inserts…
The present invention also provides increased reliability in the feedback control loop as it enables the user to eliminate numerous junctions which can introduce errors into the control system.

Vu le petit nombre de textes annotés dont nous disposions, nous avons voulu vérifier la généralité de notre algorithme. Nous avons effectué un test sur 12 textes annotés à la main, en associant à chaque paragraphe une des catégories *État de l'art/Invention/Autre*. Nous avons construit des listes de marqueurs établies autour de deux notions : la temporalité (précédent ou typique vs nouveau) et l'amélioration (problème vs solution) :

État de l'art : *typical, conventional, generally, usually, most, often, known*
Invention : *invention, object, relates, provides*
Problème : *disable, damage, disadvantage, loss, error, risk, undesirable, fail, difficulty*
Amélioration : *capable, advantage, can, allow, able, possible, advantageous*

On pourrait se contenter de chercher ces marqueurs dans les paragraphes et appliquer une recherche booléenne donnant ainsi un poids élevé aux marqueurs collectés. Cependant, si la précision des ressources construites à la main est généralement élevée, c'est au détriment du rappel. Or, ces listes sont probablement loin d'être exhaustives, car le corpus qui a servi à la collecte des marqueurs est petit. Dans ce cas, cette méthode donnerait des résultats nettement moins probants sur de nouveaux textes, différents structurellement ou thématiquement du corpus initial. Pour plus de robustesse, nous avons donc choisi d'opérer une classification probabiliste,

où les marqueurs trouvés lors de l'analyse, et chaque mot du paragraphe joueront un rôle dans la classification, mais avec un poids différent.

Classification par apprentissage supervisé

La classification probabiliste a l'avantage de s'appuyer sur un ensemble de propriétés significativement plus grand que celui d'une liste construite à la main. Cependant, pour fonctionner de manière à la fois robuste et précise, une telle méthode a besoin d'un grand corpus d'apprentissage (composé de textes annotés à la main). Comme nous ne disposions que d'un petit corpus de 12 textes annotés selon les trois catégories, nous avons décidé de combiner les deux approches pour obtenir une valeur optimale de rappel et de précision. Nous nous sommes servis des listes de marqueurs extraits à partir du corpus de 12 textes pour construire semi automatiquement un corpus d'entraînement qui fournisse davantage d'exemples positifs (c'est-à-dire de paragraphes classés de manière correcte dans l'un des deux deux groupes pertinents : *Prior Art* ou *Invention*). Une partie du corpus de 100 brevets, contenant 1361 paragraphes, a été annotée à l'aide des marqueurs, et l'annotation a été corrigée à la main. Les marqueurs ont été rangés dans deux catégories suivant leur fiabilité, un poids deux fois plus important a été associé aux marqueurs les plus sûrs qu'aux marqueurs moins fiables.

Par la suite, nous avons défini expérimentalement les paramètres de l'algorithme qui donnent une précision maximale par rapport aux paragraphes annotés semi-automatiquement. Puis, nous avons établi les seuils qui définissent l'appartenance à chaque catégorie en fonction du poids et de la quantité des marqueurs présents dans le paragraphe. Nous avons ensuite annoté le corpus entier en utilisant les paramètres qui ont donné les résultats avec la meilleure précision sur le corpus de test de 1 361 paragraphes.

Le corpus de 100 brevets, malgré sa taille limitée, est suffisant pour servir de corpus d'apprentissage pour la classification probabiliste "bayésienne naïve", que nous avons retenue. Cette méthode utilise comme modèle de probabilité sous-jacent un *modèle de traits indépendants*. Elle peut être appliquée de manière efficace à des tâches de classification supervisée, où l'estimation des valeurs de traits se fait par une estimation de probabilité maximale. Ici, l'espace des traits est construit à partir du vocabulaire des paragraphes pertinents du corpus d'apprentissage et la probabilité de chaque mot du vocabulaire d'appartenir à chacune des catégories est fournie par ce corpus.

Une évaluation de la classification sur les 12 textes annotés manuellement, non inclus dans le corpus d'apprentissage a montré que, comme prévu, les nouveaux paragraphes des textes sont correctement classifiés. Les résultats sont prometteurs malgré la taille du corpus d'apprentissage et le fait qu'il soit annoté semi-automatiquement. Nous avons ainsi opéré une réduction de 70 à 80 % en taille des documents. La précision de la classification est de 68 %, tandis que le rappel moyen est 87 %. Ces résultats sont améliorables avec relativement peu d'effort, en utilisant davantage de textes, en complétant la liste de marqueurs et en réappliquant les règles d'annotation sur le corpus d'apprentissage.

3.3. Extraction des contradictions

Une fois les paragraphes identifiés, il s'agit de rechercher des contradictions dans ceux-ci. Les expressions régulières et les mesures statistiques de pertinence aident à identifier le contexte syntaxique et ainsi à restreindre la liste des candidats. Le plus important est de détecter des oppositions. Or celles-ci peuvent s'exprimer au niveau grammatical ou lexical. Certaines répétitions sont également intéressantes dans la mesure où elles peuvent indiquer un segment de texte pertinent.

Les concepts clés que nous cherchons sont les éléments (les composants du système technique), les paramètres des éléments et les valeurs correspondantes. Dans un système technique, parmi les nombreux éléments composant l'artefact, seuls nous intéressent, ceux subissant un changement. Ce changement l'évolution de l'artefact. Les paramètres ont des valeurs qui peuvent avoir des influences soit positives soit négatives. On remarque très souvent que les trois items du triplet paramètres, valeurs, et éléments sont présents ensemble. Cependant, la contradiction est rarement exprimée dans sa totalité : la plupart des documents n'expriment qu'un changement de valeur du paramètre à la fois (amélioration ou détérioration). Dans ces conditions, notre système ne pourra que signaler à l'utilisateur des contradictions qu'il devra valider et auxquelles il rajoutera éventuellement à la main la partie non explicitée.

Il nous faut donc maintenant, d'une part identifier les trois entités, éléments paramètres valeurs, lier paramètre et élément, paramètre et valeur, d'autre part identifier les deux sens de variations des paramètres pertinents marqués par des oppositions trouvées généralement dans des paragraphes différents.

Repérage des éléments

Pour identifier les éléments pertinents de l'artefact, nous cherchons les entités spécifiques du domaine et plus encore les entités spécifiques du brevet en question. Les mots qui les désignent seront donc significativement plus fréquents dans ce brevet que dans le corpus entier.

Nous constituons des listes de candidats à être des éléments en nous basant sur une analyse de surface et en éliminant, grâce à certaines heuristiques et à des listes d'exclusions, les mots composés qui ne peuvent en être. Nous nous sommes intéressés à la fréquence relative des mots désignant des éléments dans le texte lemmatisé et avons finalement utilisé la mesure **tf-idf,** mesure de pertinence fréquemment utilisée en fouille de textes et extraction d'information. [5]

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

$$tf-idf_{i,j} = tf_{i,j} \times idf_i$$

où ni,j est le nombre d'occurrences du terme i dans le document j. où le dénominateur est le nombre d'occurrences de tous les termes dans le document j, où |D| dénote le nombre total de documents du corpus.

Cette mesure permet de sélectionner environ une dizaine d'éléments candidats par texte avec les réglages que nous avons choisis (une valeur minimale significativement plus basse pour les noms que pour les autres catégories grammaticales). Nous avons cherché les mots avec une valeur tf-idf élevée qui sont en position de tête d'un groupe nominal[6], ainsi que les adjectifs attributs. Les éléments sont repérés avec un bon rappel, mais la densité de leurs occurrences dans le texte demande davantage de filtrage. Aussi, envisageons-nous dans le futur de compiler le corpus d'une façon différente, en sélectionnant d'abord un ensemble de textes appartenant au domaine de l'artefact. La valeur tf-idf rendra alors davantage d'éléments spécifiques du document et moins d'éléments commun au domaine. Mais,

5 Le tf-idf (de l'anglais term frequency – inverse document frequency) est une méthode de pondération qui permet de quantifier l'importance informationnelle d'un mot dans un ensemble de document, un mot présent partout n'apportant aucune information, un mot présent seulement dans un sous-ensemble de documents permet de caractériser ce sous-ensemble.

6 L'analyse des syntagmes nominaux est effectuée grâce au CRFchunker disponible à http://crfchunker.sourceforge.net/

il restera à voir si cette meilleure sélection n'introduira pas de silence dans le remplissage de notre modèle CI.

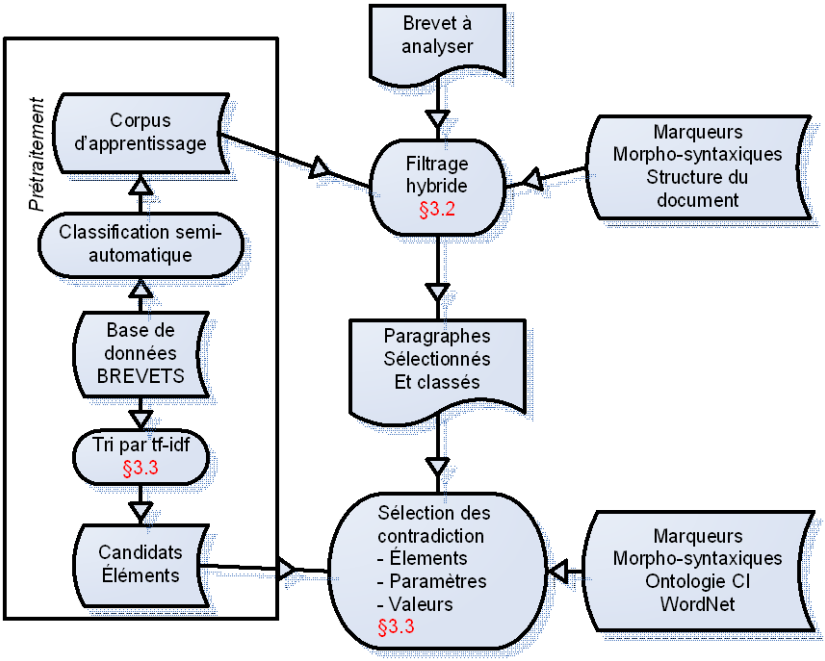


Figure 1. Vue générale de l'agencement des tâches

Module linguistique

Les éléments identifiés par le filtrage statistique doivent de toute façon être filtrés dans une phase ultérieure sur des critères linguistiques. Le module linguistique intégré au système est implémenté comme un ensemble de transducteurs, il permet de créer les liens entre éléments, paramètres et valeurs et de fournir des annotations correspondant aux rôles qu'ils remplissent selon la CI.

Nous avons utilisé plusieurs outils : l'outil LIKES (Rousselot *et al.* 2007) pour retrouver automatiquement les segments répétés, le concordancier Oxford Wordsmith Tools[7] pour sa fonction de tri des concordances, Tree Tagger (Schmid 1994) pour l'étiquetage du corpus.

Après l'analyse linguistique du corpus, nous avons collecté les marqueurs susceptibles d'être des candidats, puis nous avons sélectionné les plus efficaces. NOOJ[8] a servi à les mettre en œuvre pour l'annotation et

7 En vente à Oxford University Press
8 Téléchargeable à http://www.nooj4nlp.net/pages/nooj.html

l'extraction. Nous avons retenu 60 verbes, 137 adverbes, 473 adjectifs et 273 noms. Nous avons constaté que les verbes utilisés dans les brevets sont, dans la plupart des cas, des verbes d'action, plus précisément, des verbes de changement ou des verbes indiquant un changement d'état. Ce sont les plus productifs quant à la détection des valeurs et des paramètres. Les verbes modaux fréquents dans ces textes expriment généralement différents degrés de possibilité (nécessité ou certitude). Lorsque ces modaux sont accompagnés de l'auxiliaire "be" dans le corpus et qu'ils sont suivis de certains indices grammaticaux, ils permettent de localiser certaines informations recherchées.

Les adjectifs sont souvent porteurs de valeurs. Nous avons remarqué également l'usage fréquent des oppositions entre adjectifs dans le corpus.

Les adverbes sont une catégorie généralement difficile à étudier, car le sens de la phrase dépend de leur portée. Cependant, la sélection des adverbes qui nous intéressent a été plus facile, car, ici, seuls nous intéressent les adverbes d'évaluation.

Les rédacteurs de brevets utilisent des noms composés très complexes afin de véhiculer un maximum d'information dans une phrase et ils utilisent un grand nombre de termes pour décrire l'artefact et ses composants. De ce fait, nous ne retenons dans la liste des noms que ceux qui correspondent à des paramètres ou à des valeurs.

Nous avons implémenté les grammaires d'annotation dans NOOJ. Les grammaires correspondent à des transducteurs enrichis (usage de variables et de contraintes, consultation de dictionnaires lors de l'analyse). Nous avons constitué deux dictionnaires spécifiques et édité 46 graphes sur la base des résultats de l'analyse. Les graphes définissent les contraintes à respecter pour effectuer l'annotation. Par exemple, les oppositions doivent être au même endroit ; les verbes doivent être accompagnés de certains indices pour pouvoir être annotés ; l'annotation s'effectue seulement dans le cas où deux notions recherchées au moins existent, etc.

L'application des grammaires d'annotation fournit finalement en sortie le texte annoté en format XML exportable.

Consultation de Wordnet – recherche d'antonymes

Après avoir identifié les éléments qui participent aux changements, il reste encore à chercher les paramètres et les valeurs qui y sont attachées. Un module qui fait appel à Wordnet va essayer de repérer des valeurs opposées, sachant que les oppositions se trouvent entre les descriptions du Prior Art et de l'invention. Les paramètres qui changent de valeur peuvent

s'exprimer d'une part, au niveau lexical, soit par des adjectifs ou participes antonymiques, soit par des paires de verbes : affirmatif – négatif, d'autre part, au niveau syntaxique, par des marqueurs syntaxiques complexes, qui indiquent les rôles joués par les entités situées dans leur contexte proche.

Les oppositions lexicales sont plus faciles à identifier : il s'agit de paires d'adjectifs ou de participes antonymiques, qui sont liés (référentiellement ou syntaxiquement) aux mêmes éléments (syntagmes nominaux). Par exemple :

However, the plastic materials which can be released by resiliently deforming such an undercut area in the prior art injection blow molding process are limited to relatively ***soft*** *plastic materials.*

Another object of the present invention is to provide an injection mold which can release the core mold by resiliently deforming the undercut formed on the lip portion even if it is molded of a relatively ***hard*** *plastic material.*

À part les antonymes, présents à l'intérieur de groupes nominaux ayant une structure identique, nous avons aussi remarqué la présence fréquente de marqueurs d'opposition : par exemple '*limited to*' vs '*even if*' qui permettent d'exprimer des valeurs opposées.

Les antonymies exprimées par des adjectifs à l'intérieur des groupes nominaux ('*hard plastic materials*') ainsi qu'entre les adjectifs ou participes qui ont une fonction prédicative en tant que têtes syntaxiques sont utiles. Il existe également des cas où des substantifs réfèrent à des propriétés exprimant des valeurs de paramètres. Le module fait appel à Wordnet et sélectionne des couples d'adjectifs parmi ceux trouvés dans WordNet, en excluant les adjectifs qui réfèrent à la position, à l'ordre, etc. (p.ex. first-second/last, inner-outer) et qui précisent généralement les éléments du système technique sans exprimer de jugement de valeur.

Les oppositions syntaxiques sont, elles, plus difficiles à localiser. Elles se manifestent souvent par des répétitions lexicales dans des contextes différents, par exemple la même action exprimée une fois dans un contexte affirmatif, et plus tard reprise dans un contexte négatif :

However, such a mold structure <u>disables the release of a molding</u> from the mold. Namely, the undercut of the molded preform as well as the mold <u>will be damaged</u> when the injection core mold is drawn out from the interior of the molded preform.

It is therefore an object of the present invention to provide an injection mold which can injection mold a preform to be biaxially stretch blow molded with a lip portion having an undercut and also which <u>can release the core mold</u> <u>without damaging of the undercut</u> .

Les deux paires d'oppositions syntaxiques sont les suivantes :

disables the release of a molding	*vs*	*can release the core mold*
the undercut will be damaged	*vs*	*without damaging the undercut*

Ces structures peuvent être trouvées par des expressions régulières, comme vu plus haut. Cependant, alors que les éléments et leurs paramètres, ainsi que les paramètres et leurs valeurs sont toujours à chercher dans la même phrase, les oppositions doivent être cherchées dans des paragraphes différents. La recherche doit donc tenir compte des répétitions lexicales et examiner les contextes grammaticaux des segments répétés pour en extraire les oppositions potentiellement pertinentes.

4. Conclusion et directions futures

Nous avons présenté les différentes tâches à effectuer pour extraire des connaissances orientées changement à partir des textes de brevets. Nous avons mis en place une méthode hybride s'appuyant sur des ressources et des outils de traitement de langues pour extraire les informations pertinentes. Notre objectif final est la conception d'un prototype logiciel qui permettra aux inventeurs de connaître l'évolution d'un artefact à un instant donné et comme point de départ pour une future invention.

Le système, encore en cours de développement, comprend des modules de prétraitement linguistique (étiquetage morphologique, analyse syntaxique de surface), un module de fouille de texte statistique, une série de grammaires régulières et, finalement, un module de consultation de WordNet. L'ajout d'un deuxième module linguistique est envisagé pour améliorer les résultats sur le repérage des oppositions. Le contrôle du lancement des différents modules se fait manuellement pour l'instant.

Les études et les expérimentations effectuées ont permis de voir quels éléments incorporer dans la future chaîne de traitement et quelles tâches étaient automatisables ou non.

Elles ont montré qu'il est nécessaire de disposer d'un outil capable de prendre en compte le résultat d'une ou plusieurs expressions régulières et de raisonner sur les contextes dans lesquels on les a trouvés. Pour passer à une plus grande échelle, un module qui facilite l'accès aux bases de données de brevets accessibles sur Internet est également souhaitable.

La chaîne de traitement est en cours de développement autour de LIKES (Rousselot *et al.* 2007) déjà cité. En effet, la toute dernière version de

LIKES possède un plug-in qui permet déjà d'accéder à Google, et bientôt à GooglePatent. LIKES intègre maintenant un système expert basé sur SNARK (Laurière 1986) qui permet d'une part de de faire de la déduction sur le résultat de la recherche d'expressions régulières et d'autre part de lancer des tâches ou des programmes, le système expert servant alors de langage de script. Le système final permettra d'intégrer alors des tâches opérées par des modules Perl (tf-idf par exemple) ou C++(TreeTagger).

Les expérimentations effectuées ont également permis de créer des ressources linguistiques génériques réutilisables, qui devront encore bien sûr être complétées afin d'améliorer la qualité des résultats. Nous savons maintenant que les résultats que nous obtenons sont intéressants, même s'ils doivent parfois être vérifiés et complétés par l'homme. C'est pourquoi, nous projetons d'adjoindre à la chaîne de traitement un module destiné à traiter les liens de références entre brevets et à visualiser ces liens grâce à une interface permettant de cheminer entre eux de manière agréable.

5. Bibliographie

Agatonovic M., Aswani N., Bontcheva K., Cunningham C., Heitz T., Li Y., Roberts I., Tablan V. (2008) : *Large-scale, Parallel Automatic Patent Annotation.* Proc. of the 1st Int. CIKM Workshop on Patent Information Retrieval - PaIR'08, Napa Valley, California, USA

Altshuller Guenrich (1999) : *The Innovation Algorithm* : *TRIZ. Systematic innovation and technical creativity*, Worchester, Mass, Technical Innovation Center

Cascini G. and D. Russo D. (2007) : *Computer-aided analysis of patents and search for TRIZ contradictions.* Int. J. of Product Dev. Vol.4, no.1/2, pp.52-67

Cavallucci Denis, Rousselot François, Zanni-Merk Cecilia (2008) : *Representing and selecting problems in Contradiction Network*, in Proc of the 2nd IFIP Session on Computer-Aided Innovation

Cunningham H., Maynard D., Bontcheva K., Tablan V. (2002) : *GATE : A Framework and Graphical Development Environment for Robust NLP Tools and Applications*. Proceedings. of ACL'02. Philadelphia

Feldman R., Fresko M., Hirsh H., Aumann Y., Liphstat O., Schler Y., Rajman M. (1998) : *Knowledge Management : A Text Mining Approach*. Proc.of the 2nd Int. Conf. on Pract. Aspects of Knowledge Management, Basel

Ghoula Nizar, Khelif Khaled & Dieng-Kuntz Rose (2007) : *Supporting Patent Mining by using Ontology-based Semantic Annotations*, Proc. of IEEE/WIC/ACM International Conf. on Web Intelligence, Silicon Valley, USA

Goujon Bénédicte (1999) : *Extraction d'informations pour la veille technologique avec le système VIGITEXT*, Actes de RECITAL, Cargese, France

Guyot Brigitte, Normand Sylvie (2004) : *Le document brevet, un passage entre plusieurs mondes*, Actes du Forum pluridisciplinaire document et organisation, Semaine document numérique, La Rochelle

Laurière Jean-Louis (1986) : *Un Langage déclaratif : Snark*, Technique et science informatique, vol. 5, no 3

Mille Simon, Wanner Leo (2008) : *Making Text Resources Accesible to the Reader*, *The Case of Patent Claims*, Proceedings of LREC, Marakesh (Morocco)

Rousselot François, Montessuit Nicolas (2007) : *LIKES un environnement d'ingénierie linguistique et d'ingénierie des connaissances*", Formaliser Les Langues Avec L'ordinateur : De Intex À Nooj", Koeva Svetla, Maurel Denis, Silberztein Max Presses Université de Franche-Comté, Cahiers De La MSH Ledoux ,ISBN 2848671890

Schmid Helmut (1994) : *Probabilistic Part-of-Speech Tagging Using Decision Trees*, Proc. of the Int. Conference on New Methods in Language Processing, pp. 44-49

Sheremetyeva Svetlana (2003) : *Natural Language Analysis of Patent Claims*, Proc. of the ACL-2003 workshop on Patent corpus processing, Sapporo, Japan

Verhaegen P-A., D'hondt J., Vertommen J., Dewulf S., Duflou J.R. (2008) : *Searching for Similar Products through Patent Analysis*. Proc. of the ETRIA TRIZ Future 2008 Conf, Twente

Zanni Cecilia, Cavallucci Denis, Rousselot François (2009) : *An ontological basis for computer aided innovation*, Computers in Industry, ISSN 01663615

A propos des auteurs

Kata Gábor

LGECO – INSA de Strasbourg
24, bd. De la Victoire
67000 Strasbourg
http://lgeco.insa-strasbourg.fr
katagood@gmail.com

François Rousselot

LGECO – INSA de Strasbourg
24, bd. De la Victoire
67000 Strasbourg
http://lgeco.insa-strasbourg.fr
francois.rousselot@insa-strasbourg.fr

François de Bertrand de Beuvron

LGECO – INSA de Strasbourg
24, bd. De la Victoire
67000 Strasbourg
http://lgeco.insa-strasbourg.fr
francoisdebeuvron@insa-strasbourg.fr

Denis Cavallucci

LGECO – INSA de Strasbourg
24, bd. De la Victoire
67000 Strasbourg
http://lgeco.insa-strasbourg.fr
dénis.cavallucci@insa-strasbourg.fr

Dildar Wu

LGECO – INSA de Strasbourg
24, bd. De la Victoire
67000 Strasbourg
http://lgeco.insa-strasbourg.fr
angellawooh@gmail.com

Corpus et Web : deux alliés pour la construction et l'enrichissement automatique de classes conceptuelles

Nicolas Béchet, Mathieu Roche, Jacques Chauché

Résumé : Cet article présente des méthodes permettant de construire et d'enrichir des classes conceptuelles. Ces classes sont construites en utilisant les informations syntaxico-sémantiques issues d'un corpus. La première méthode d'enrichissement se fonde sur l'utilisation du corpus et d'objets de verbes dits complémentaires. Nous présentons alors un protocole d'évaluation automatique permettant de valider la pertinence de ces objets réduisant ainsi le travail de l'expert. La seconde méthode permet d'enrichir les concepts avec des termes plus généraux en s'appuyant sur les ressources du Web.

Mots-clés : terminologie, classes conceptuelles, enrichissement, Web

1. Introduction

La terminologie est un domaine ayant de nombreuses applications en TAL (Traitement Automatique des Langues). Elle peut-être vue comme l'étude des mots techniques propres à un domaine et de leurs significations. Nous distinguons deux types d'études terminologiques : l'approche *sémasiologique* et l'approche *onomasiologique*. La première s'intéresse à l'étude des significations partant du *mot* pour en étudier le sens. La seconde, propose de partir du *concept*.

Un concept peut être défini comme la *représentation mentale d'une chose ou d'un objet* (Desrosiers-Sabbath 1984). Nous proposons de définir un concept comme un *ensemble de connaissances partageant des caractéristiques sémantiques communes*. Nous utilisons dans nos travaux l'approche *sémasiologique* en apportant un début de réponse aux problèmes générés par ce type d'approches. La terminologie ainsi extraite est en effet très dépendante du corpus. Cela implique alors qu'une terminologie répondant à des besoins spécifiques est vouée à une faible réutilisabilité (Roche 2005).

Nous proposons dans nos travaux de construire, dans un premier temps, des classes conceptuelles spécifiques en nous appuyant sur les données issues de corpus. Pour cette tâche, nous construisons des classes conceptuelles en étudiant la dépendance syntaxique des termes d'un corpus (**section 2**). Pour cela, nous nous fondons sur les relations syntaxiques *Verbe-Objets*. De telles relations sont assez représentatives d'un domaine. Par exemple dans un texte du domaine de l'informatique, le verbe *charger* prendra comme objet un nom qui appartient à la classe conceptuelle *logiciels* (L'Homme 1998).

Une extraction terminologique ainsi effectuée s'apparente à une analyse distributionnelle "à la Harris" (Harris 1968). Il existe de nombreux travaux effectuant une telle analyse pour l'acquisition de ressources terminologiques ou ontologiques à partir de textes. Citons par exemple (Bourigault et Lame 2002) dans le domaine du droit et (Nazarenko *et al.* 2001) dans le domaine médicale.

Une fois les classes conceptuelles constituées, nous proposons de les enrichir avec deux méthodes distinctes. Une première utilise les ressources syntaxico-sémantiques du corpus afin de proposer de nouveaux termes (**section 3**). Ces termes sont alors ordonnés automatiquement avant d'être soumis à un expert. L'autre méthode d'enrichissement se fonde sur l'utilisation du Web afin de générer de nouvelles ressources terminologiques qui sont moins dépendantes du corpus (**section 4**).

2. Méthode pour la construction de classes conceptuelles

Cette section présente une méthode de construction de classes conceptuelles fondée sur l'utilisation d'informations syntaxiques d'un corpus : les relations syntaxiques de type Verbe-Objet. Ainsi un concept est formé avec les objets des verbes d'un corpus qui ont été jugés "proches". Nous détaillons ci-dessous la construction de tels concepts.

La première étape consiste à extraire les relations syntaxiques d'un corpus. Nous utilisons pour cela l'analyseur morpho-syntaxique *Sygfran* (Chauché 1984). Ainsi, avec cette analyse syntaxique, nous avons par exemple extrait de la phrase : "*Thierry Dusautoir brandissant le drapeau tricolore sur la pelouse de Cardiff après la victoire*". la relation syntaxique : "*verbe : brandir, objet : drapeau*"

La seconde étape de notre méthode consiste à rassembler les objets des verbes jugés proches. Pour cela, nous émettons l'hypothèse que des verbes peuvent être considérés comme sémantiquement proches s'ils partagent un nombre important d'objets en communs. Ainsi, nous utilisons le score d'Asium, se fondant sur cette hypothèse (Faure 2000) afin de mesurer la proximité des verbes de notre corpus. Le principe d'Asium est similaire à (Bourigault *et al.* 2002), se fondant sur une analyse distributionnelle.

Alors, nous regroupons tous les objets dont les verbes ont été jugés proches. Nous illustrons dans la figure 1 un exemple de verbes sémantiquement proches : *agiter* et *brandir*. Les **objets communs** à ces deux verbes constituent le concept : *Objets symboliques*. Notons que le concept formé ici est d'une thématique peu fréquente montrant l'intérêt d'une telle construction de classes conceptuelles spécifiques.

La troisième étape de notre méthode propose de distinguer deux types d'objets pour construire une classe conceptuelle. Les objets **communs** aux deux verbes de la figure 1 *drapeau* et *fleur* sont des instances de qualité du concept *Objets symboliques* mais qu'en est-il des deux autres objets *pancarte* et *rasoir* ? Ces objets, qui ne sont objets que d'un seul verbe sont appelés objets **complémentaires.** Ils permettent d'induire de l'information. En effet, la relations syntaxique *brandir pancarte* n'est pas originalement présente dans le corpus et est induite de la relation *agiter pancarte* et du verbe *brandir*.

La méthode définie dans cet article considère « nativement » un objet **commun** comme instance d'un concept. Néanmoins, les objets **complémentaires** nécessitent d'être validés. En effet, l'objet *pancarte* est une instance pertinente du concept *Objets symboliques*, contrairement à *rasoir*.

Nous présentons dans la section suivante des solutions de validation des objets complémentaires ainsi que des protocoles d'évaluation afin d'estimer la qualité de ces validations.

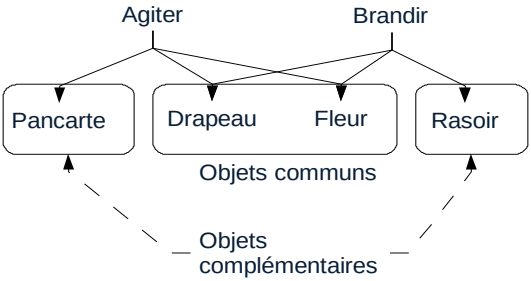


Figure 1. Objets communs et complémentaires des verbes "agiter" et "brandir"

3. Traitement des objets complémentaires

3.1. Les mesures de validation

Cette section présente brièvement les approches permettant d'ordonner en termes de qualité les relations syntaxiques induites. Elles sont présentées en détail dans (Béchet *et al.* 2009a). Nous émettons l'hypothèse qu'une relation syntaxique induite cohérente signifie que l'objet complémentaire la composant est une instance pertinente du concept formé par les objets communs des deux verbes.

Vecteurs sémantiques (VS)

Une première validation consiste à mesurer la pertinence de l'association d'un verbe avec son objet complémentaire. Ainsi, nous allons valider la proximité sémantique entre le verbe et l'objet d'une relation induite avec l'autre verbe et le même objet. Sur l'exemple de la figure 1, il s'agit de mesurer la proximité sémantique des relations *agiter pancarte* (relation originale) et *brandir pancarte* (relation induite). Les relations syntaxiques sont représentées par des **vecteurs sémantiques**.

Un tel vecteur est construit en représentant un ou plusieurs termes en le(s) projetant sur un espace de dimension finie de 873 concepts. Ces concepts sont organisés comme une ontologie de concepts définis dans le thésaurus (Larousse 1992). Chaque mot est indexé par un ou plusieurs éléments. Par exemple, "brandir" est associé à "agitation" et "drapeau" à "paix, armée, funérailles, signe, cirque". La représentation d'une relation syntaxique résulte d'une combinaison linéaire entre la représentation du verbe et de l'objet, dont les coefficients prennent en compte la structure syntaxique

(Chauché 1990). La proximité des vecteurs est finalement mesurée par un calcul de cosinus entres les deux vecteurs sémantiques (vecteurs propres aux relations originales et aux relations induites).

Validation Web (VW)

Une autre validation utilise le **Web** afin de mesurer la dépendance entre verbe et objet d'une relation syntaxique induite. Elle s'inspire des travaux de (Turney 2001). Ainsi, une requête est soumise à un moteur de recherche, sous forme de chaîne de caractère. Le nombre de pages de résultats retournés constitue la mesure de dépendance. De plus, nous appliquons diverses mesures statistiques telles que l'*Information Mutuelle* (Church et Hanks 1990) ou le *coefficient de Dice* (Smadja *et al.* 1996). Ceci permet de pondérer l'importance de la relation syntaxique en fonction du verbe et de l'objet dans les résultats obtenus. Nous utiliserons dans cet article uniquement l'Information Mutuelle suite aux expérimentations effectuées. L'Information Mutuelle adaptée à notre approche est définie comme suit :

$$IM(v,o) = \frac{nb(v,o)}{nb(v)nb(o)}$$

Avec *nb(v), nb(o) et nb(v,o)* étant respectivement le nombre de pages retourné par le moteur de recherche lors de la soumission du verbe *v,* de l'objet *o* et de la relation syntaxique *vo*. La particularité de la validation Web est qu'elle utilise des ressources extérieures afin de mesurer la cohérence d'un candidat à un concept. Ainsi, à partir d'informations spécifiques propres à un corpus, nous obtenons une évaluation plus globale de la pertinence des concepts.

Les combinaisons

Nous proposons de combiner les deux approches présentées précédemment avec deux types de combinaisons.

- La première combinaison introduit un paramètre $k \in [0,1]$ pour donner un poids supplémentaire à l'une ou l'autre des approches. Pour une relation syntaxique *r*, nous combinons les approches avec le calcul suivant :

$$combine_score_r = k \times VS + (1 - k) \times VW$$

Avec VS et VW étant respectivement les scores obtenus pour la relation syntaxique *r* avec l'approche des vecteurs sémantique et la validation Web.

- La seconde combinaison consiste à classer la totalité des relations syntaxiques par l'approche VS. Puis, les *n* premières relations syntaxiques sont de nouveau ordonnées avec l'autre approche VW.

L'ensemble des approches présentées dans cette section vont fournir une liste de candidats aux différents concepts ordonnés par valeurs décroissantes des différentes mesures (VS, VW, combinaison 1, combinaison 2). Nous ne présenterons dans cet article que les résultats de la validation Web et de la seconde combinaison, mesures obtenant les meilleures performances (Béchet *et al.* 2009b). Les autres résultats sont proposés en annexe.

La section suivante présente différents protocoles d'évaluation afin de mesurer la qualité de nos approches.

3.2. Protocoles expérimentaux

Nous avons montré dans de précédents travaux la qualité de ces approches (Béchet *et al.* 2009b) en les évaluant par un protocole automatique. Nous proposons dans cet article d'effectuer une évaluation manuelle sur un échantillon de relations syntaxiques. Ainsi, nous pourrons confirmer la qualité de nos approches et également la qualité du protocole d'évaluation automatique décrit ci-dessous.

Nos expérimentations utilisent un corpus écrit en français. Il est extrait du site Web d'informations de Yahoo (http://fr.news.yahoo.com) appartenant au domaine *actualités avec un style journalistique*. Il contient 8 948 articles. Les expérimentations ont été effectuées à partir de 60 000 relations produites dans (Béchet *et al.* 2009b). Dans nos expérimentations, nous sélectionnons *manuellement* cinq concepts, dont les instances sont les objets communs des verbes ayant généré le concept[1], présentés dans le tableau 1.

Concepts	*Organisme /Administration*	*Fonction*	*Objets symboliques*	*Sentiment*	*Manifestation de protestation*
Instances	parquet	négociateur	drapeau	mécontentement	protestation
	mairie	cinéaste	fleur	souhait	grincement
	gendarme	écrivain	spectre	déception	indignation
	préfecture	orateur		désaccord	émotion
	pompier			désir	remous
	onu				tollé
					émoi
					panique

Tableau 1. Les cinq concepts sélectionnés et leurs instances

1 Issus des concepts de verbes ayant un score de plus de 0,7 avec le score d'Asium (Faure 2000)

L'objectif de nos protocoles présentés ci-dessous est d'évaluer la qualité des objets complémentaires pouvant enrichir les cinq concepts définis. L'objectif est donc de proposer à l'expert les objets complémentaires les plus pertinents. Dans ce but, nous allons utiliser les approches présentées en section 3.1 qui classeront ces objets. Dans la section suivante, nous proposons d'évaluer la qualité du classement obtenu.

Protocole d'évaluation automatique

Le principe de l'évaluation automatique est d'utiliser un second corpus écrit également en français, de taille plus conséquente que celui d'où proviennent les relations induites. Les deux corpus sont du même domaine. Nous jugeons alors comme bien formés des relations induites qui vont être présentes *nativement* dans le second corpus. Une telle relation sera alors qualifiée de **positive**. Notons que les relations jugées négatives peuvent être de faux négatifs. En effet, une relation qui n'a pas été retrouvée dans le second corpus n'est pas pour autant non pertinente. De plus, un objet complémentaire dans une relation syntaxique jugée pertinente peut également ne pas être un « bon » candidat pour un concept. C'est pourquoi nous présentons dans la section suivante un protocole d'évaluation manuel.

Protocole d'évaluation manuel

Pour mesurer la qualité de notre protocole automatique et vérifier le bon comportement des approches de validation des relations syntaxiques induites, nous proposons d'effectuer une validation manuelle des relations. Nous disposons de huit évaluateurs. Nous leurs avons alors soumis un formulaire. Celui-ci a pour objectif de faire valider manuellement des termes pouvant appartenir à un concept. Pour chacun des cinq concepts extraits, nous soumettons aux évaluateurs les objets candidats, qui ne sont autres que les objets **complémentaires** de l'un ou l'autre des verbes, tel que défini dans la section 2.2. L'évaluateur doit alors mesurer la pertinence d'un terme pour un concept donné en respectant le barème suivant :

- 2 : Parfaitement pertinent
- 1 : Susceptible d'être pertinent
- 0 : Non pertinent
- N : Ne se prononce pas

La figure 2 présente une capture d'écran du formulaire soumis aux experts.

Nous présentons alors deux variantes permettant d'utiliser les scores attribués par les juges : une moyenne des scores obtenus et un système de votes.

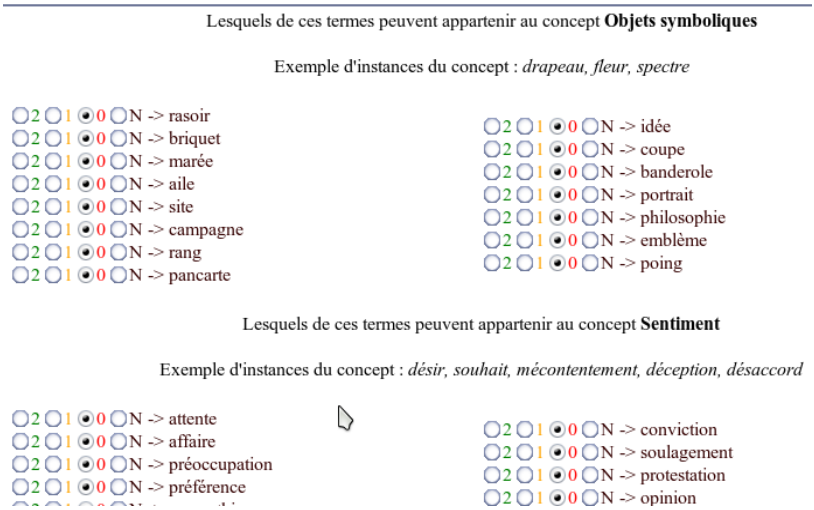


Figure 2. Capture d'écran du formulaire d'évaluation manuelle

La moyenne. Après l'évaluation des objets candidats (553 termes) par les experts, nous effectuons une moyenne des résultats obtenus en faisant varier la tolérance aux résultats obtenus. Nous distinguons alors différents intervalles afin de considérer un résultats comme positif ou non. Par exemple, un terme peut-être positif si son score est supérieur à 1.

Le vote. Une autre manière de qualifier un candidat de positif est de soumettre les scores donnés par les juges à un système de vote. Nous qualifions alors de pertinent un candidat si un pourcentage *p* de juges l'ont jugé pertinent. Un score pertinent d'un juge peut alors être *1* ou *2*.

Une fois la notion de *candidats pertinents* définie avec les protocoles présentés, nous proposons d'évaluer le classement issu de nos différentes approches en utilisant les courbes ROC. Cette méthode est décrite ci-dessous.

Les courbes ROC

Terme	***Validation Manuelle***
Conviction	+
Opinion	+
Pr□f□rence	-
Attente	-
Col□re	+

Tableau 2. Exemple de classement de termes du concept *Sentiment*

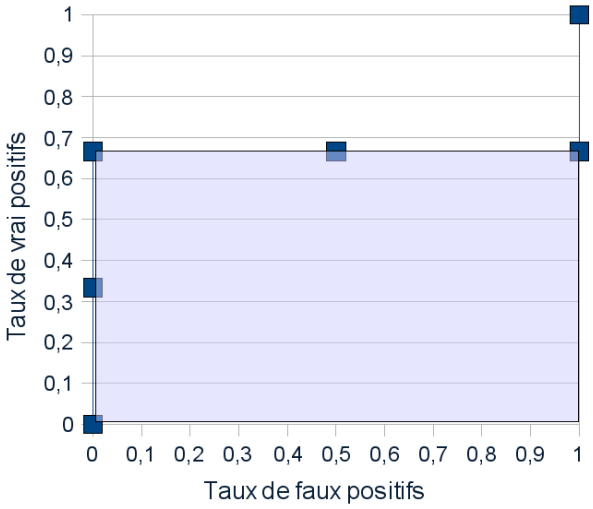


Figure 3. Courbe ROC de l'exemple du tableau 2

La méthode des courbes ROC (Receiver Operating Characteristic), détaillée par (Ferri *et al.* 2002), fut utilisée à l'origine dans le domaine du traitement du signal. Cette méthode est fréquemment employée en médecine afin d'évaluer automatiquement la validité d'un diagnostique de tests. On trouve en abscisse des axes représentant une courbe ROC le taux de faux positifs et l'on trouve en ordonnée le taux de vrais positifs. La surface sous la courbe ROC ainsi créée est appelée AUC (Area Under the Curve). Un des avantages de l'utilisation des courbes ROC réside dans leur résistance à la non parité de la répartition du nombre d'exemples positifs et négatifs.

Une courbe ROC représentée par une diagonale correspond à un système où les relations syntaxiques ont une distribution aléatoire, la progression du taux de vrais positifs est accompagnée par la dégradation du taux de faux positifs. Considérons le cas d'une validation de relations syntaxiques induites. Si toutes les relations sont positives (ou pertinentes), l'AUC vaudrait 1, ce qui signifie avoir toutes les relations pertinentes en début de liste, donc ordonnées de manière optimale.

Le tableau 2 présente un exemple de termes ordonnés avec la seconde combinaison évalués par une validation manuelle pour le concept *Sentiment*. La courbe ROC alors obtenue avec cette validation est présentée dans la figure 3. Nous obtenons finalement une AUC (aire sous la courbe ROC, bleutée sur la figure 3) de 2/3 avec cet exemple.

3.3. Résultats expérimentaux

L'objectif que nous nous fixons dans ces travaux est de réduire la tâche de l'expert en filtrant le nombre de relations syntaxiques induites candidates à un concept. Les expérimentations ci-dessous ont pour but de montrer dans quelle mesure nos approches de validations sont intéressantes. Ainsi nous introduisons un seuil qui n'est autre que le nombre de relations syntaxiques considérées.

Un seuil fixé à 100 indique que l'on ne mesure l'AUC que pour les 100 premières relations syntaxiques. Le tableau 3 présente les AUC obtenues avec les approches validation Web et combinaison 2. La combinaison 2 obtient les meilleures performances. Les résultats des autres approches sont présentés en annexe. Ce tableau compare le protocole d'évaluation manuel avec l'automatique. Pour le protocole manuel, nous ne présentons que les résultats obtenus avec le système de *vote*. Une relation positive est validée si 75% des experts ont attribué la note de 2. Les résultats utilisant la *moyenne*, présentés en annexe, sont assez similaires, et donc non reportés ici.

Avec la seconde combinaison, l'évaluation manuelle donne d'excellents résultats pour les premières relations (AUC jusqu'à 0,83). Les résultats sont de bonne qualité (AUC de 0,70) jusqu'au seuil de 350, pour se dégrader avec la totalité des candidats (AUC proche de l'aléatoire 0,5). Nous ne pouvons ainsi pas fournir à l'expert une liste triée de l'ensemble des candidats mais une liste contenant un sous ensemble. Ainsi, nous privilégions la précision et la qualité de la liste fournie à l'expert en réduisant en contre partie le nombre de candidats disponibles initialement (plus faible rappel).

	Validation Web		***Combinaison 2***	
	Vote	**Auto**	**Vote**	**Auto**
nb relations	***AUC***		***AUC***	
50	0,64	0,59	**0,81**	**0,90**
100	0,50	0,60	**0,83**	**0,87**
150	0,62	0,66	**0,80**	**0,84**
200	0,61	0,65	**0,76**	**0,79**
250	0,56	0,66	**0,71**	**0,75**
300	0,51	0,65	**0,70**	**0,74**
350	0,57	0,67	**0,69**	**0,75**
400	0,59	0,67	**0,67**	**0,74**
450	0,61	0,67	**0,65**	**0,71**
500	0,56	0,68	**0,57**	**0,70**
550	**0,52**	**0,69**	**0,52**	**0,69**

Tableau 3. AUC obtenues avec la seconde combinaison pour le protocole manuel et automatique

Nous proposons maintenant de comparer les scores de l'évaluation manuelle avec l'automatique. Nous constatons que les résultats sont du même ordre pour les deux approches. En effet, les résultats de la combinaison 2 sont de très bonne qualité pour les faibles seuils et se dégradent avec la totalité des résultats. Pour la validation Web, les résultats sont assez réguliers de l'ordre de 0,60 pour l'évaluation manuelle et 0,65 pour l'automatique. Nous montrons alors, sur cet échantillon de candidats que notre protocole de validation automatique est de bonne qualité. Il permet en effet de montrer que les premières relations sont les mieux classées avec la seconde combinaison. Il montre également que cette approche fournit les meilleurs classements (Cf. tableaux en annexe).

Néanmoins, les scores obtenus ont tendance à être surévalués avec le protocole automatique. Ces scores s'expliquent notamment par la diversité des tâches effectuées par les deux protocoles. Le protocole manuel cherche à connaître la pertinence d'un terme dans un concept. Le protocole automatique propose de mesurer la cohérence d'une relation syntaxique formée d'un verbe et d'un objet complémentaire. Ces tâches, bien qu'assez proches, ne visent pas les mêmes objectifs. Il est en effet plus difficile de mesurer de manière automatique la qualité d'un candidat potentiel à un concept que la qualité d'une relation syntaxique.

Nous avons présenté précédemment une méthode afin de construire et d'enrichir des classes conceptuelles en utilisant les objets de verbes jugés proches. Nous présentons dans la section suivante une autre méthode d'enrichissement de ces classes utilisant le Web.

4. Enrichissement via le Web

Avec notre précédente méthode d'enrichissement, nous utilisons les informations d'un corpus afin de proposer de nouveaux termes pour enrichir des concepts. Une telle approche utilise des connaissances spécifiques pour enrichir les concepts. Elle est donc propre au corpus. Nos approches fondées sur le Web permettent une validation utilisant des ressources plus générales. Cependant, les termes proposés suite à ces validations sont limités à la thématique du corpus d'où ils sont extraits. Nous présentons dans cette section une autre approche d'enrichissement fondée sur le Web utilisant ainsi des ressources de domaines plus généraux que celles d'un corpus.

4.1. Méthode d'enrichissement

L'objectif de cette méthode est de fournir de nouveaux candidats aux concepts formés tel que décrit section 2. Elle se fonde sur l'énumération de termes sémantiquement proches présents sur le Web. Par exemple, en saisissant dans un moteur de recherche la requête (chaîne de caractères) "lundi, mardi et", nous obtenons d'autres jours de la semaine en résultats.

Afin d'appliquer cette méthode, nous considérons dans un premier temps les objets communs des verbes jugés sémantiquement proches. Ils constituent les *instances de références* des classes ainsi formées. Nous proposons alors d'utiliser le Web afin d'acquérir de nouveaux candidats. Cette méthode présente l'avantage de ne plus se limiter aux termes du corpus dont les classes conceptuelles sont issues.

Considérons alors les N concepts $C_{i \in \{1,N\}}$ et leurs instances respectives $I_j(C_i)$. Pour chaque concept C_i nous soumettons alors à un moteur de recherche les requêtes suivantes :

"$I_{jA}(C_i)$, $I_{jB}(C_i)$ **et"** *et* "$I_{jA}(C_i)$, $I_{jB}(C_i)$ **ou"**

avec jA et jB$\in \{1, NbInstanceC_i\}$ et jA $\neq$ jB. Plus concrètement avec l'exemple de la figure 1, nous fournissons au moteur de recherche les requêtes : "*drapeau, fleur et*", "*drapeau, fleur ou*", "*fleur, drapeau et*", "*fleur, drapeau ou*".

Le moteur de recherche nous retourne alors un ensemble de résultats desquels nous extrayons de nouveaux candidats à un concept. Après avoir identifié la requête dans nos résultats, le terme qui suit notre requête constitue une nouvelle instance du concept, tel qu'illustré dans l'exemple suivant.

Considérons la requête : "*drapeau, fleur et*", le moteur nous retourne alors :

"Tu joues version normale (Carreau, pique, coeur et trèfle) ou version bourbi... heu... suisse-allemande (Gland, **Drapeau, Fleur et** Grelot)".

Après avoir identifié notre requête dans le résultat retourné (en gras sur notre exemple), nous ajoutons au concept le terme suivant directement la requête (ici, le terme *Grelot*).

4.2. Protocole et résultats expérimentaux

Nous avons expérimenté cette seconde méthode d'acquisition de termes afin d'enrichir nos cinq concepts déjà expérimentés dans la section

3.3. Nous utilisons pour nos expérimentations l'API du moteur de recherche *Yahoo!* afin d'obtenir nos nouveaux termes.

Trente cinq nouveaux termes ont été obtenus, sur lesquelles nous avons appliqué différents traitements. Tout d'abord, nous appliquons un filtrage grammatical afin de ne conserver que des noms, puis nous appliquons un élagage en supprimant les termes génériques, tels que *même, chose, avoir, etc.* Il nous reste alors trente termes.

Nous faisons alors évaluer à 3 experts les différents termes en leurs demandant si un terme peut être considéré comme pertinent (respectivement, non pertinent) dans un concept. Nous obtenons alors trois évaluations. Nous calculons pour chacune le taux de positifs défini par le nombre de termes pertinents sur le nombre total de termes. Notons que dans notre cas, le taux de positifs n'est autre que la précision. Finalement, nous effectuons une moyenne de la précision qui atteint **0,70** dans nos expérimentations.

Ce résultat est assez encourageant bien que pouvant être amélioré. En effet, la thématique même du concept présenté aux experts est discutable car elle est établie manuellement. Alors la subjectivité de l'évaluation humaine joue un rôle non négligeable dans cette évaluation.

Citons par exemple le concept « *manifestation de protestation* ». La question posée est alors : le terme "*adhésion*" est-il une instance correcte de ce concept ? Une définition du terme adhésion (provenant du TLFI) est : "Reconnaissance implicite ou explicite de l'autorité d'une loi, d'un gouvernement, etc.". D'une manière triviale en se fondant sur cette définition, on aurait plutôt tendance à dire que ce terme appartiendrait à un concept opposé à celui-ci. Mais si l'on considère une adhésion comme un engagement politique ou associatif s'opposant aux règles ou aux lois établies, il peut être perçu comme un moyen de protestation. Cette subjectivité humaine pose là une question importante au niveau de la qualité de l'évaluation humaine pour des systèmes de fouilles de textes.

5. Synthèse

Nous proposons dans cette section de présenter une synthèse des deux approches d'enrichissement des concepts.

Nous allons nous appuyer sur le concept *Sentiment*. Ce concept a été formé par les objets communs de deux verbes jugés proches : *exprimer* et *manifester*. Les instances de ce concept sont : *désir, souhait, mécontentement,*

déception, désaccord. Nous proposons alors d'utiliser les deux méthodes présentées dans cet article afin d'enrichir ce concept.

Première méthode : Induction d'informations provenant du corpus.

Nous *fabriquons* une liste d'objet dits *complémentaires* à partir du corpus. Ces objets vont alors être ordonnés avec la seconde combinaison des approches validation Web et vecteurs sémantiques. Nous les soumettons alors à un expert qui va sélectionner les plus pertinentes.

Les objets retenus par l'expert sont : sympathie, regrets, doute, crainte, exaspération, satisfaction, sensibilité, espoir, indignation, dédain, joie, amertume, désarroi, solidarité, confiance, colère.

Seconde méthode : Enrichissement via le Web.

La seconde méthode propose d'utiliser des ressources extérieures *généralisant* ainsi notre concept original. Elle se fonde sur l'envoi de requête à un moteur de recherche. Les candidats obtenus par le Web après validation de l'expert sont : *fatalisme, stabilité, angoisse, inconscience*.

6. Conclusion

Cet article a proposé de joindre les ressources d'un corpus à celles du Web afin d'enrichir des classes conceptuelles. De telles classes ont été formées par des objets communs de verbes jugés sémantiquement proches. Nous avons alors proposé deux approches d'enrichissement afin d'extraire de nouveaux termes. Une première utilisant les objets des verbes dits complémentaires. Nous avons alors validé ces approches avec différents protocoles, dont un totalement automatique, se révélant assez efficace. Cependant, un tel enrichissement limite la construction des classes à un domaine spécialisé, à l'image du corpus dont les termes sont extraits. Nous avons alors présenté une autre méthode, se fondant non plus sur le corpus mais sur le Web. Cette méthode s'est avérée pertinente. Mais une évaluation plus poussée doit être menée afin de confirmer sa qualité.

Nous envisageons comme futurs travaux pour la première méthode, d'introduire le contexte dans nos différentes approches de validation. Pour la validation Web, il s'agit introduire le contexte dans la requête fournie au moteur de recherche. Pour les vecteurs sémantiques, il s'agit d'utiliser des vecteurs dit *contextualisés* prenant en compte la structure morphosyntaxique de la phrase dont le terme à valider est issu. De plus, nous pensons intégrer la notion de *nominalisation* afin d'acquérir une plus grande quantité

de relations syntaxiques. Citons par exemple "consommer un fruit", plus fréquemment rencontré comme suit "consommation de fruits".

La seconde approche mérite quant à elle d'être approfondie, en effectuant notamment des expérimentations sur des corpus d'autres thématiques. Nous souhaitons également pouvoir exécuter à plusieurs reprises l'acquisition de termes par le Web. Nous devrons alors faire sélectionner les termes par un expert ou bien avec une méthode automatique qui devra être proposée. Ces termes vont effet être utilisés afin de former la requête Web d'où la nécessité de les sélectionner rigoureusement.

Les travaux présentés dans cet article posent la question ouverte de la qualité de l'évaluation des ressources terminologiques, sémantiques ou autres. Bien que la plupart de ces ressources furent conçues avec la validation d'experts des domaines traités, la subjectivité humaine peut mettre en doute la qualité de certaines de ces ressources ainsi que les évaluations telles que celles présentées dans ce papier. Par exemple, un domaine est-il défini par toutes personnes de manière analogue ? Ou encore un concept a-t-il toujours la même caractérisation ?

Bibliographie

Béchet, N. *Comment valider automatiquement des relations syntaxiques induites*. In EvalECD'09, acte des ateliers de EGC'09, p. A5-25 - A5-33, 2009a.

Béchet, N., Roche M., Chauché J. Towards the Selection of Induced Syntactic Relations. In ECIR'09, poster proceedings, to appear, 2009b.

Bourigault D. et Fabre C. *Approche linguistique pour l'analyse syntaxique de corpus*. Cahiers de Grammaires, 25, 131–151., 2000.

Bourigault D., Lame G. *Analyse distributionnelle et structuration de terminologie. Application à la construction d'une ontologie documentaire du Droit*, in TAL, 43-51, 2002.

Chauché, J. *Un outil multidimensionnel de l'analyse du discours*. In Proceedings of Coling, Standford University, California, p. 11–15, 1984.

Chauché, J. *Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance*. In TA Information, pp. 17–24, 1990.

Church, K. W. et Hanks P. *Word association norms, mutual information, and lexicography*. In Computational Linguistics, Volume 16, pp. 22–29, 1990.

Desrosiers-Sabbath *Comment enseigner les concepts* - Sillery: Presses de l'Université du Québec, 1984.

Faure, D. *Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM*. Ph. D. thesis, Université Paris-Sud, 2000.

Ferri, C., Flach P., et Hernandez-Orallo J.. *Learning decision trees using the area under the ROC curve*. In Proceedings of ICML'02, pp. 139–146., 2002.

Harris Z. *Mathematical Structures of Language*, New-York, John Wiley & Sons, 1968.

Larousse, T. *Thésaurus Larousse - des idées aux mots, des mots aux idées*. Ed.Larousse, Paris, 1992.

L'Homme M. -C. *Le statut du verbe en langue de spécialité et sa description lexicographique*. Cahiers de Lexicologie. 73. 61-84, 1998.

Nazarenko A., Zweigenbaum P., Habert B, Bouaud J. *Corpus-based Extension of a Terminological Semantic Lexicon*. In Recent Advances in Computational Terminology, 327-351, 2001.

Smadja, F., McKeown K. R., et Hatzivassiloglou V. *Translating collocations for bilingual lexicons : A statistical approach*. Computational Linguistics 22(1), 1–38, 1996.

Roche C. *Terminologie et ontologie*. Revue Langages n°157, pp. 48-62, Éditions Larousse, mars 2005.

Turney, P. Mining the Web for synonyms : PMI– R versus LSA on TOEFL. Proc of ECML, LNCS, 2167, 491–502, 2001.

A propos des auteurs

Équipe TAL - LIRMM
UMR 5506, CNRS, Univ. Montpellier 2
34392 Montpellier Cedex 5 - France
{bechet,mroche,*chauche*}@*lirmm.fr*
http://www.lirmm.fr/~{bechet,mroche,chauche}

Annexes

Automatique	**VS**	**VW**	**C. 1**	**C. 2**
nb relations	***AUC***			
50	0,54	0,59	0,65	**0,90**
100	0,54	0,60	0,65	**0,87**
150	0,55	0,66	0,73	**0,84**
200	0,48	0,65	**0,80**	0,79
250	0,52	0,66	0,66	**0,75**
300	0,47	0,65	0,64	**0,74**
350	0,50	0,67	0,62	**0,75**
400	0,51	0,67	0,64	**0,74**
450	0,50	0,67	0,62	**0,71**
500	0,53	0,68	0,62	**0,70**
550	0,55	**0,69**	0,62	**0,69**

Tab. 4: AUC obtenues avec le protocole automatique.

Vote	**VS**	**VW**	**C. 1**	**C. 2**
nb relations	***AUC***			
50	0,50	0,64	0,78	**0,81**
100	0,66	0,50	0,54	**0,83**
150	0,55	0,62	0,69	**0,80**
200	0,57	0,61	0,74	**0,76**
250	0,53	0,56	0,58	**0,71**
300	0,35	0,51	0,55	**0,70**
350	0,42	0,57	0,53	**0,69**
400	0,46	0,59	0,53	**0,67**
450	0,46	0,61	0,53	**0,65**
500	0,41	0,56	0,46	**0,57**
550	0,39	**0,52**	0,43	**0,52**

Tab. 5: AUC obtenues avec le protocole manuel, en utilisant le *vote*.

Moyenne	**VS**	**VW**	**C. 1**	**C. 2**
nb relations	***AUC***			
50	0,54	0,62	**0,79**	0,74
100	0,57	0,53	0,54	**0,82**
150	0,52	0,58	0,69	**0,79**
200	0,50	0,61	**0,75**	0,75
250	0,42	0,57	0,56	**0,65**
300	0,34	0,52	0,49	**0,68**
350	0,42	0,56	0,51	**0,67**
400	0,47	0,58	0,53	**0,66**
450	0,46	0,60	0,52	**0,63**
500	0,41	0,57	0,47	**0,57**
550	0,39	**0,53**	0,43	**0,53**

Tab. 6: *AUC obtenues avec le protocole manuel, en utilisant la moyenne (score positif pour une moyenne supérieur ou égale à 1,75).*

SESSION 2

Following the path between conceptual maps and visual thesauri

Olga Bessa Mendes

Abstract : One of the challenges in digital libraries is to give access to more efficient ways of information retrieval. In the environment of special information in digital libraries there is a demanding for a new approach of information management because terminology is a *unit of information* and also the *access point to information*. Facing the context of special libraries in digital environment and focusing on the confront of schemes for information classification, we propose using a *visual thesaurus* as guide at a search engine as means for comprehension of information. The relevant issue is about the classification structure needed to offer a frame of reference of information which can be built in a conceptual map. The goals of the visual thesaurus within a special library are to offer a more dynamic search of information to users, as well as to guide them on searching and also to contribute to special information literacy.

Keywords : Information Science, digital special library, information literacy, visual thesaurus, Terminology

1. Introduction

The library social and pedagogical mission that combines information preservation and dissemination turned out to be more visible with current technologies. Considering digital library as a place of access to knowledge for users with different profiles, our concern is to know *how to improve access to technical and scientific information* to specialists and non-specialists users in digital special library.

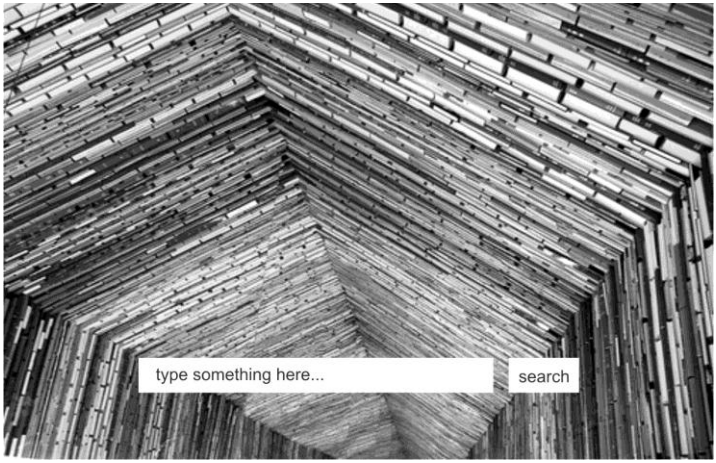
In a recent investigation work (Mendes 2008) we have analysed the procedures of information management in Information Science and methodologies for terminology organization, in Terminology.

In Information Science, librarian plays a role of mediation between information and users. In this context, of special information, we have realized that there is a complementary bond between terminology organisation, in Terminology, and information management, in Information Science, for *terminology* is the means of communication connecting user and information, or librarian and user or librarian and information.

In this article, we focus on some conclusions of the referred investigation work, underlying the *classification scheme* as the main issue to achieve information visualization and knowledge representation in digital library. Beginning with a synthesis of context and problematic on information retrieval in digital special library, we emphasize the need for improving criteria in thesaurus buillding; then, we analyse the common structure of some reference thesauri and the problematic of search and information visualization. The conceptual map scheme arises as contribute for the visual thesaurus architecture.

2. Digital special library

A digital library is, first, a repository of digital documents and/or digitalized. Contents and terminology organization for information access is an obstacle either to user or to information manager, because of information amount and diversity. The digital special library accentuates the need for specialized terminology use and classification instruments to establish a terminology connection for specialists and non-specialists users. In a technical and scientific domain, a vocabulary guide is considered an essential tool for search orientation.



Matej Krén - *Book cell*

Figure 1. A *magic* search box in a digital library

2.1. Context & problem

Libraries, in particular of public access, undertake an ethical code of coherent and harmonized criteria use shared between affiliated institutions. In this context, we use normalization tools for information cataloguing and indexing. The librarian has to establish criteria for content analysis in a clear and consensual way between the existing collection in library and the community of practice.

Indexing is the action that consists in describing a document in relation to its content, representing it in a formal language, or documental language. The concern for a coherent practice of indexing based on the presentation and organization of preferred terms is the purpose to achieve effectiveness in information retrieval. The need for criteria normalization in thesaurus elaboration arises under the concern for a communication and share of controlled languages between libraries.

In order to use these principles which assist a thesaurus methodology, we have selected the standards currently used in libraries, namely - *ISO2788 1986-Documentation. Guidelines for the establishment and development of monolingual thesauri*; and *ISO5963 1985-Documentation. Methods for examining documents, determining their subjects, and selecting indexing terms*. However, we have identified some gaps on the methodology guidelines to elaborate controlled vocabularies and to organise information in the point of view of a special library. From this analysis, we focus the following aspects :

difficult reading of the graphical display of relationships between concepts ;

associative relationship lacks identification of the relationship typology ;

preparation and use phases of the thesaurus coexist and bring out ambiguity in content analysis ;

the *scope note* is not adequate for the technical and scientific information arrangement because it is not mandatory and also requires elaborating criteria.

2.2. Improving methodology

In the environment of technical and scientific information, we may conclude that criteria for content analysis and identification of terminology are the core work needing development on the standards referred above [1].

Being a thesaurus a vocabulary of a controlled indexing language, formally organized so that the a priori relationships between concepts are made explicit (ISO2788 1986 : 5) and, considering that a vocabulary like a dictionary, is a product of Terminology work (ISO1087-1 2000 : 12), we have analysed the standards that present the course of actions for terminology organization. The ISO704 (2000) standard presents a description of the relationships between objects, concepts and their representation; furthermore it settles that a conceptual system performs for model concept structures based on specialized knowledge of a field (ISO704 2000 : 12). We find these orientations suitable to the work developed in Information Science on thesaurus elaboration to special information.

On one side, for there is the need to define all terms before the organisation of vocabulary for indexing language (choosing the *descriptor* and the *non-descriptor*). And on the other side, considering the progression of technical and scientific information a constant, the thesaurus of a specific domain may well be a tool of systematization of terminology evolution and of knowledge representation.

The presentation structure elected for a thesaurus provides a uniform use as information indexing tool. In general, thesauri are structured for indexer understanding and use with a special purpose. Nevertheless, each thesaurus has to be adapted to subject organization and dissemination as

1 Under development is the ISO/CD 25964-1 - Information and documentation: Thesauri and interoperability with other vocabularies - Part 1: Thesauri for information retrieval which revises the standards ISO2788 1986 and ISO5964 1985. Available from www : <URL:http://www .iso.org>.

required by information service. In our opinion, to a special library is recommended an information classification scheme focused on domain and built in user's perspective.

3. Visual thesaurus architecture

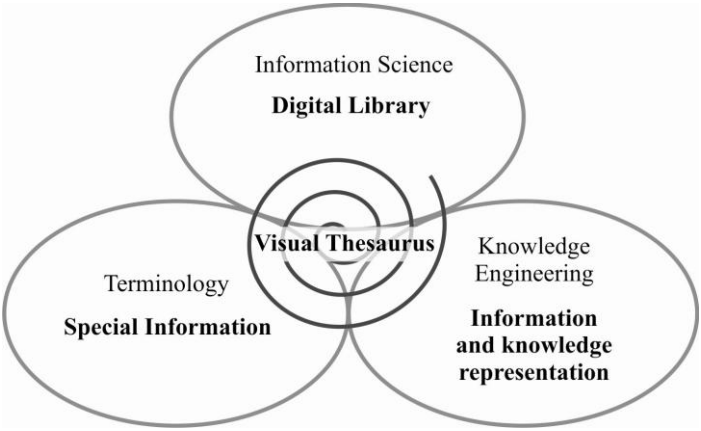


Figure 2. Visual thesaurus outline

Classifications are interrelated with the evolution of scientific knowledge and scientific concepts change as theories evolve. Therefore, special information management will bring in more coherence if terminology that evolves through time is incorporated in a classification system, according to demand.

To identify the relationships between concepts is a form of classification, since classifications are geographic elements of mind and *only them allow us orientation in the world around us, to establish habits, similarities and differences, recognize places, beings, events; to arrange them, group them, to draw near from each other, to keep them together or keep them away hopelessly* (Pombo 1998 : 1). To organize knowledge is, afterwards, a need for comprehension of a context which is accomplished through the association and distinction of concepts.

3.1. Thesaurus structure

Nowadays, some thesauri consent the navigation and visualization of terms and its relationships, however do not enable to identify the delimitation of concepts. The alphabetic presentation and identification of relationships between terms are confined to descriptor in hierarchical and

associative relationships. In the case of *NASA thesaurus*[2], the alphabetic and hierarchical presentation reflects the concern to present definitions (Figure 3).

absorptance
DEF The ratio of the radiant flux absorbed by a body to that incident upon it.
GS electromagnetic properties
. optical properties
. . **absorptance**
RT albedo
capture effect
cosmic ray albedo
density (mass/volume)
Earth albedo
electromagnetic absorption
light transmission
lunar albedo
microwave absorption
opacity
reflectance
surface properties

ammonia
magnetic cooling
refrigerants

absorption cross sections
DEF In radar, cross sections characterized by the amount of power removed from a beam by absorption of radio energy by a target to the power in the beam incident upon the target. Used for capture cross sections.
UF *capture cross sections*
RT ∞ absorption
∞ cross sections
ionization cross sections
neutron cross sections
radiation absorption
scattering cross sections
stopping power

Figure 3. ***NASA Thesaurus***

We choose *Thesaurus of Nations*[3] as an example that combines image, equivalence and partitive relationships that allow us to *see the world* with other details. Its structure is based on ontological relationships: *is a, a division of, included in, country of, neighbour*; and includes also an alphabetic presentation of the equivalence and hierarchical relationships (Figure 4).

Romania
Serbia
Turkey
Prs بلغارستان

Burkina Faso
Country of Africa
Is a Geographical division
A division of Nations
Capital Ouagadougou
Neighbor Benin
Cote d'Ivoire
Ghana
Mali
Niger
Togo
Prs بورکینافاسو

Canada
Country of North America
Included in America
Is a Geographical division
A division of Nations
Capital Ottawa
Neighbor Greenland
United States
Prs کانادا

Figure 4. ***Thesaurus of Nations***

These instruments constitute relevant sources of information, but in our opinion they still do not establish a good connection with user. These organization formats are not appealing and friendly in the perspective of an interactive access and elucidative about information/knowledge for

2 Available from www: <URL:http://www.sti.nasa.gov /98Thesaurus/ vol1.pdf>
3 Available from www: <URL http://www.thesaurusbuilder.com>

user. Even if combining various relationships, they do not offer yet a whole perspective of the universe, that is, of the domain under analysis.

3.2. Information search and information visualization

Nowadays there are several technological tools that allow constructing interactivity on digital library. Some reference thesauri[4], from FAO, UNESCO, NASA, among others, which deal with specific domains but combining several sub-domains, already present changes on terms of organization and search. For example, the structure of UNESCO thesaurus[5] permits *online* access for the theme *Performing arts* as micro-thesaurus (MT) and terms list in hierarchy (Figure 5).

Term: **Performing arts** [289]
Terme français: Arts du spectacle
Término español: Artes escénicas
Русский термин : Исполнительские виды искусства

MT 3.55 Performing arts
UF Acrobatics
UF Circuses
RT Ballet [24]
RT Cinema [433]
RT Cultural activities [478]
RT Cultural industry [296]
RT Entertainment [72]
RT Music [414]
RT Performers [130]
RT Theatre [283]

Figure 5. ***UNESCO thesaurus***

However, we have realized that the proposed presentation scheme for the elaboration of thesaurus, in ISO 2788 : 1986 standard, that these thesauri use, is insufficient to knowledge representation in special library perspective.

We consider that using the ontological relationships structure and combining orientations and also the relationships present on the referred standards, both for Information Science and Terminology, we can build a management instrument for information and knowledge representation to assist the information indexing and classification as well as the search guide.

Founding on available technologies, and also depending on purposes, we can mention either information visualization or knowledge visualization.

4 FAO-Food Agriculture organization; UNESCO-United Nations Educational, Scientific and Cultural Organization; NASA-National Aeronautics and Space Administration.
5 Available from www: <URL : http://databases.unesco.org/thesaurus/>

As suggests Card *et al.*[6], information visualization *is the use of computer-supported, interactive, visual representations of abstract data to amplify cognition*. The employment of images to follow and support thought (*visual thinking*) is a method used at a pedagogical level, which communication in scientific context, as in vulgarization speech context, recuperates to create better interaction between sender and receiver.

Visual perception is an issue discussed in cognitive sciences and that involves also the knowledge apprehension. In Terminology, the terms arrangement in hierarchy and association also offers us a graphical image. Although, we question whether from this information representation we may obtain a knowledge representation. Burkhard (2006 : 2) suggests that *knowledge visualization examines the use of visual representations to improve the transfer of knowledge between at least two persons or group of persons*.

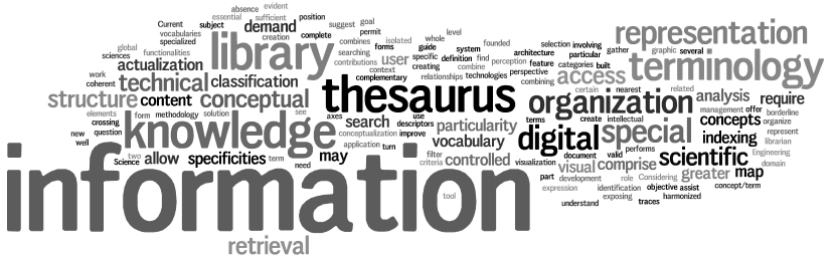


Figure 6. Frequency representation [www.wordle.net]

In fact, on the various specialities and sciences, the information representation tools are distinct and adapted to communication needs. Where in Architecture we have the sketch or in Statistics a graphic bar and a thematic map, in Terminology we can have a conceptual map.

The common information representation of language studies is given in statistical graphics for frequency, occurrence and co-occurrence of terms. On the contrary, more than to know how many records there are for each *subject* (descriptors) in the digital library, our goal is to identify a theoretical framework to create the infography of thesaurus' content in order to present a formal representation of concepts/terms related to a specific subject.

6 Card et al. — Readings in information visualization : using vision to think, 1999 apud Burkhard, (2006 : 2)

3.3. Conceptual map

The infrastructure of conceptual map and its graphic display allow having the interactive visualization of conceptual relationships that provides the understanding of concepts/terms.

We can use a conceptual map in different stages of information arrangement (Novak 2008 : 2)[7].

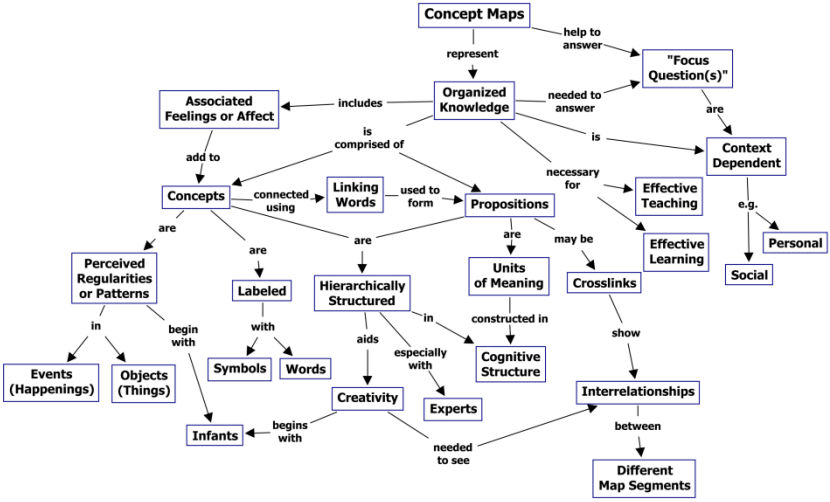


Figure 7. Conceptual map of conceptual map

First, when we build a conceptual system to systematize and communicate the relationships between concepts, and later too we use it to navigate among concepts of a certain domain, in order to search and apprehend information. As Tricot (2006 : 50) suggests *les cartes conceptuelles sont généralement utilisées pour organiser des idées, concevoir et communiquer une structure complexe et résoudre des problèmes.*

The conceptual map may constitute a visual guide for searching information joined to communication and knowledge discover for user. The intended hypertext organization of terminology management, which offers search dynamism, is close to the organization of our own conceptual model of knowledge representation. The advantage for representing ontological relationships stands in the guidance offer to user and to indexer of the connecting lines between concepts.

7 Available from www: <URL:http://cmap.ihmc.us/Publications/ResearchPapers/TheoryCmaps/TheoryUnderlyingConceptMaps.htm>.

In this context, the architecture structure of knowledge organization is built with complementary traces from Terminology, Information Science and Knowledge Engineering. Taking into account the contributions of these sciences we gather some elements for a work methodology on controlled vocabulary organization which combines the information organization and the knowledge representation.

4. Special classification schemes

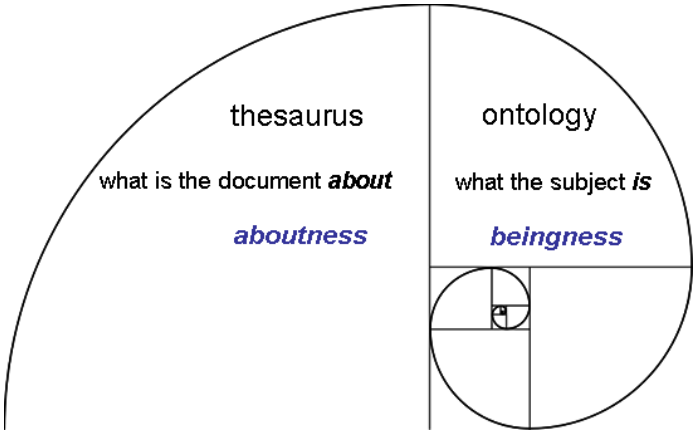


Figure 8. Classification framework of the visual thesaurus

Considering that ontology is a structure applied to information management providing information representation through the established relationships, we might have this same structure tailored to special information management needs in digital library. We are facing a thesaurus evolution form and we consider that an ontology synthesis might be a *special classification* of a domain, which can be presented in a dynamic form and online access.

Beyond hierarchical and associative relationships, different researchers refer the relationships that provide *faceted classification*, or by classes, or by categories, that allow a specific domain organization. In fact, we recall that Otlet (1934) and Briet (1951)[8] made reference to information classification into *information nodes*. Might we be under the same question about the knowledge organization structure? Can it be *only* the use of distinct terms for a similar need of information arrangement?

8 Otlet, P. Traité de documentation: le livre sur le livre, théorie et pratique. Bruxelles : Editions Mundaneum, 1934. Briet, S. What is documentation? [online]. 1951 [Cited 20-05-2008]. Available from www <URL:http://ella.slis.indiana.edu/~ roday/ what% 20is% 20 documentation. pdf>

Thesaurus and ontology are structures to describe formally information and to represent knowledge, although thesaurus' formal language is not sufficiently formal to computer system.

In building an ontology for thesaurus we may achieve a balance between a conceptual scheme and a presentation scheme of information (Smith, 2001:63), where descriptors stand for the identification of *document's subject* which are complemented by the concept map display as access to *what the subject is* (Figure 8).

In this approach the *visual thesaurus* is a controlled vocabulary of an *indexing language* [i. e. thesaurus] that establishes formal links to *language in speciality* displayed in a concept map.

The need for special classification structures was referred a long time ago for various researchers like Vickery (1960) who gives accent to faceted classification preparation, and recently a group of researchers gave a renewed attention to this debate (Gnoli 2008) underlying the need for a broader understanding information universe.

However, their relevant reflections about classifications created by Ranganathan, Bliss and Dahlberg are concerned with generalist libraries arrangement. Nevertheless, these methodology discussions are an incentive to compare efforts and create synergies between disciplines that deal with classifications and knowledge representation.

Future steps on this path between conceptual maps and visual thesauri, will take place on analysis of faceted classification perspective in Information Science and its contribute to visual thesaurus structure.

5. Conclusion

All concepts have at least one documental sense (civilization) and one expressive dimension (cultural)[9]

In digital special library, the technical and scientific terminology performs a borderline and development role for information access.

Before the question of *how to improve access to technical and scientific information in the digital library*, we suggest the conceptual map as new feature on thesaurus structure for information and knowledge representation.

Thesaurus is a tool that can combine two functionalities: as controlled vocabulary, to indexing, and as guide to assist on information searching.

9 Prado Coelho (2003 : 4). Available from www: <URL : www.ciberscopio .net>.

The classification particularity and demand that we find in the specialized library require specificities in content analysis and greater terminology actualization. To access the term as an isolated form, in a thesaurus, is not sufficient. To understand the use and application of technical and scientific information we need to have a whole perspective of the nearest related axes of concept/term to see its position in conceptual system.

Classification particularity and demand founded in the special library require specificities in content analysis and greater actualization of terminology. The conceptualization of thesaurus organization, while intellectual creation valid to indexing and to information retrieval, does not comprise with the absence of harmonized criteria for terminology management.

Current technologies allow creating information visualization but comprise only a part of solution. To digital special library, the information retrieval ought to be more complete on identification of relationships between concepts for which definition is essential. The information representation and knowledge representation are graphic or visual forms, with the objective of exposing information and turn it evident.

In this context, the architecture structure of knowledge organization is built with complementary traces from Terminology, Information Science and Knowledge Engineering. Considering the contributions of these sciences we gather some elements for a work methodology on controlled vocabulary organization which combines the information organization and the knowledge representation.

The visual thesaurus may have the conceptual map structure combining descriptors, as categories of concepts, with specific terms of a certain subject and it will permit to organize and to represent knowledge of a domain. At the information retrieval level, the thesaurus selection as search filter may allow crossing several vocabularies (involving document, librarian and user) and to create a more coherent search expression.

The purpose is to offer a global as well as particular perception of information to user in digital special library.

Bibliography

Burkhard Remo (2006) : *Learning from architects: the difference between knowledge visualization and information visualization*, Information Visualization Journal [online], Vol. 5, nº 3, [Cited 08-05-2008], Available from www : URL : http:// www.palgrave-journals.com/ivs/index.html, ISSN 1473-8716

Costa Rute (2006) : *Plurality of theoretical approaches to terminology*, In PICHT, Herbert, ed. -Modern approaches to terminological theories and applications, Bern : Peter Lang, ISBN 3-03911-156-6, pp. 77-89

Gnoli Claudio (2008) : ed. *Axiomathes*, vol. 18, nº 2, [online], Springer Netherlands, ISSN 1572-8390

ISO5963 (1985) : *Documentation. Methods for examining documents, determining their subjects, and selecting indexing terms*, Genève : ISO

ISO 2788 (1986) : *Documentation. Guidelines for the establishment and development of monolingual thesauri*, Genève : ISO

ISO 1087-1 (2000) : *Terminology work : vocabulary. Part 1: Theory and application*. Genève : ISO

ISO 704 (2000) : *Terminology work. Principles and methods*, Genève : ISO

Lerat Pierre (2005) : *Terme et microcontexte : les prédications spécialisées*, In 7es Journées scientifiques AUF-LTT : Mots, termes et contextes, [online], [Cited 03-11-2007], Available from www : <URL:http://perso.univ-lyon2.fr/~thoiron /JS%20 LTT%202005/pdf/Lerat.pdf>

Mendes Olga Bessa (2008) : *Information management in special library – the borderlines on searching information*, [master thesis], Lisboa, Universidade Nova de Lisboa

Mendes Olga Bessa (2009) : *Case study. From survey to indicators : the role of terminology*, [working paper] Lisboa : INE

Pombo Olga (1998) : *Da classificação dos seres à classificação dos saberes*, Leituras, Revista da Biblioteca Nacional de Lisboa, ISSN 0873-7045, Nº 2, Primavera, pp. 19-33

Roche Christophe (2007) : *Le terme et le concept: fondements d'une ontoterminologie*, In Conférence TOTh 2007 - *Terminologie & Ontologies : Théories et Applications*. Annecy, Institut Porphyre

Roche Christophe (2008) : *Quelle terminologie pour les sociétés de l'information ?* Lexique, dictionnaire et connaissance dans une société multilingue, Cahiers de linguistique, ISBN 978-2-930481-52-4. Vol. 43/1, pp. 139-160

Smith Barry (2001) : *Ontology for information systems* [online], [Cited 10-05-2009], Available from www : <RL:http://www.ecor.uni-saarland.de/otherpublicati ons.html>

Tricot Christophe (2006) : *Cartographie sémantique : des connaissances à la carte* [online], Annecy, Université de Savoie, [PhD thesis]. [Cited 03-05-2008], Available from www : < http://www.knowledge-mapping.net/>

Vickery Brian Campbell (1960) : *Faceted classification. A guide to the construction of special schemes*. London, Aslib

A propos des auteurs

Olga Bessa Mendès

INE – Instituto Nacioanl de Estatística
CLUNL – Centro de Linguística da Universidade Nova de Lisboa
Lisboa, Portugal
olgabessa@gmail.com

Dynamic Concept Relations : A Definition and Representation Proposal

Chiara Messina

Résumé : Terminology standards and reference books mainly deal with static concept relations, such as generic (or hierarchical) and partitive relations. Ontological (or associative) concept relations have been seldom dealt with. Still, they play a fundamental role in the concept systems of many domains, thus requiring a deeper analysis. In particular, standards for categorization and representation of ontological relations are needed, in order to allow both terminologists and knowledge engineers to manage these relations in cooperative work.

In this paper we take up the distinction between static and dynamic situation proposed by (Jouis 2007) and further analyze dynamic concept relations and their representation. Our aim is to contribute to this research topic by suggesting a starting point for further analysis and by highlighting the issues that have to be tackled. After a brief overview of the previous work, we will outline a definition proposal for dynamic concept relations and propose a way of representing them using conceptual graphs. We shall then present a case study where conceptual graphs are used to represent dynamic relations between terms belonging to law field. In this case study, we shall use some of the concepts collected at the University of Genoa within the last edition of our Specialization degree course and University Master degree course in Legal Translation and their Italian and English designations.

Mots-clés : terminology, dynamic concept relations, conceptual graphs, ontology

1. Introduction

Most terminology standards and reference books mainly deal with static concept relations and leave little room to non-hierarchical or associative concept relations, which are usually dealt with in a concise and less detailed way.

Hierarchical relations offer the fundamental structure to design a concept system, as they help classify super-ordinate and subordinate concepts, but they still lack the necessary expressive power to describe complex concept systems or dynamic subject fields. For example, they are suitable for describing domains featuring clear structures, such as highly technical product information, catalogues, etc. With regard to this, we fully agree with (Arntz *et al.* 2004) : "Da die Terminologielehre in ihrer Entstehungsphase stark durch die technische Normung beeinflußt wurde, standen die hierarchischen Begriffsbeziehungen, die eine eher statische Sprachbetrachtung widerspiegeln, zunächst im Mittelpunkt". Still, the increasing complexity of the subject fields that terminology has to tackle nowadays shows the need for further research on non-hierarchical relations. This is especially true if we consider the growing importance that terminology is gaining in the field of knowledge management and in knowledge representation systems, where the needs of translation management have led not only to the use of specific terminology management tools, but also to an increasing integration between terminology and complex knowledge representation systems. Indeed, a domain terminology, considered as an organized concept system including concept designations and relations, is a good starting point for the development of a KRS. We are especially referring to formal ontologies, as they share a number of features with terminology. Terminology provides the conceptual structure of a domain, a formal ontology, defined as "an explicit specification of a conceptualization" (Gruber 1993), represents this knowledge formally, thus providing an operationalisation of the terminology. The main components of formal ontologies are: classes, relations, functions, formal axioms and instances. Terminology shares the first two components with ontology, classes (which correspond to concepts) and relations, but it still lacks the power to represent all kind of relations. Indeed, hierarchical relations are clearly represented in both terminology and ontology ; on the the other hand, a standard for non-hierarchical and thematic relations has not yet been defined in terminology. From this point of view, formal ontologies seem to be the most natural output for terminologies; this idea arises not only from the integration between conceptualization and specification, but also from the

common background of terminology and ontology. Indeed, both terminology and ontology are based on an epistemological theory, i.e., they represent a world model – or, in other words, they both have an ontological commitment.

This paper is not meant to propose any definitive solution, but it is aimed at contributing to the research in field of dynamic concept relations by highlighting some issues that are worth being dealt with in future work. Starting from the distinction between static and dynamic situation proposed by (Jouis 2007), we shall focus on dynamic relations and outline a definition proposal. This paper introduces then a representation of dynamic concept relations by means of conceptual graphs (CGs). A similar approach was first introduced by (Gerzymisch-Arbogast 1996) ; the notation adopted in this paper, however, is based mainly on a purely logical approach (Sowa 1976, 1999) and features a higher degree of formality. We indeed consider CGs as a "formal notation that serves as an intermediary between the human and the computer" (Sowa 1976). Conceptual graphs can be expressed in a quite clear and intuitive display form that can be easily used in every-day terminology practice or translated to other logical notations, thus offering the possibility to operationalise terminology while preserving user-friendliness for terminologists and formalization possibilities for engineers.

A case study shows how conceptual graphs are used to represent dynamic relations between terms belonging to law field. In this case study, we shall use some of the concepts collected at the University of Genoa within the last edition of our Specialization degree course and University Master degree course in Legal Translation and their Italian and English designations.

2. Dynamic Concept Relations

As a beginning of this section, we shall introduce a short overview of the standards and some previous work about non-hierarchical concept relations, in order to give an insight into the current situation and into the issues that have to be clarified.

According to the standard ISO 704, "an associative relation exists when a thematic connection can be established between concepts by virtue of experience. […] Some associative relations exist when dependence is established between concepts with respect to their proximity in space or time. […] Some relations involve events in time such as a process dependent on time or sequence, others relate cause and effect" (ISO 704

2000). Space and time proximity are also considered in the German standards (DIN 2330 1993) and (DIN 2342-1 1992), which define non-hierarchical concept relations as the relations existing when spatial, temporal or causal connections are considered. The DIN standards divide non-hierarchical relations into sequential and pragmatic relations, providing a short classification: causal relations (cause-effect), genetic relations (producer-product), stages of a process, and transmission relations between sender and receiver belong to sequential relations, while thematic relations, which are nor hierarchic nor sequential, belong to pragmatic relations.

The classification proposed by (Felber 1984) is derived from Wüster's ideas. In Felber's view, ontological relations are essentially relations "characterized by contiguity (juxtaposition) in space or time or by the connection cause-effect". Felber divides concept relations into logical relationships, ontological relationships and relationships of effect. He ranks partitive relations, relations of succession and relations of material – product under ontological relations ; while causality, tooling and descent (with the subtypes genealogic descent, ontogenetic descent and descent between stages of substances) are ranked under the relations of effect. In this classification, thematic relations are not considered. This schema is quite different from the one proposed by ISO, which ranks partitive relations among hierarchical relations. Causal relations have been deeper analysed by Nuopponen (Nuopponen 1994a, 1994b). Nuopponen's categories are based essentially on Wüster's previous work (Wüster 1974), which she extends with a more detailed classification of causal relations and of the concepts involved.

A different meta-terminology is used in (Cabré 1999). Cabré distinguishes between logical and ontological concept relations. The latter are divided into coordination relationships ("(part-whole relationships), which describe two types of relationships : those established between a whole and its parts […], and those among the various parts of a single whole […]. These relationships are based on the contiguity of objects in space, and are thus produced simultaneously") and "chain relationships, which are based on the succession of objects in time (cause-effect relationships), which are thus sequential". Thus, this classification is based mainly on time variation issues, as it distinguishes between relations that happen at the same time and relations produced over a certain lenght of time. (Cabré 1999) further specifies that "ontological relationships can also exist between the formatives or words that make up complex terms". The dynamic of term formation is beyond the scope of this paper ; for insights in this topic (Kageura, 1997, 2002) can be referred to.

(Arntz *et al.* 2004) essentially follow the division of the standard DIN 2330 between sequential and pragmatic relations. To the sequential relations proposed by the DIN standard, Arntz et al. add "chronologische Beziehung", "Herstellungsbeziehung", "instrumentelle Beziehung" and "funktionelle Beziehung". Furthermore, they rank the relation concerning the stages of a process as a subordinate relation of the chronological relation.

A totally different account of dynamic concept relations is given in (Kageura 1997). The subject is dealt from an operational point of view and the categories proposed for concept system classifications are closer to the components of a formal ontology than to the traditional terminological classifications. Indeed, the categories introduced are entities, activities, and quality and relations; a further division mirrors an ontological commitment. This approach integrates terminology issues with ontological categories.

Finally, with regard to dynamic concepts we would like to mention (Pilke 2000).

If we try to sum up what we know about non-hierarchical concept relations from the previous work and the standards mentioned, we are first of all faced with a terminological question: how are we to name relations which are other than hierarchical relations ? As candidate terms we have "associative relations" (ISO 704 2000), "nichthierarchische Begriffsbeziehungen" (non-hierarchical relations) (DIN 2330 1993 ; DIN 2342-1 1992), "ontological relationships" (Cabré 1999) and "relationships of effect" (Felber 1984). Furthermore, previous literature does not clarify thematic relations in details. According to the standard ISO 704, they correspond to associative relations, therefore as a candidate term for naming this kind of relations we also have "thematic relations".

The brief overview outlined above shows clearly that there is absolutely no full consensus about the classification of non-hierarchical relations, nor on their definition. In some cases, the same relations are mentioned, but they are ranked in a different way by the various authors. Thematic relations mentioned in the DIN standards are neither further analyzed nor mentioned by other authors. Indeed, they are highly domain-dependent, which prevents general classification. Still, as they are at the core of most domain terminologies, thematic relations need to be dealt with, especially as far as their representation is concerned. In addition to the disagreement on theoretical issues concerning dynamic relations, it is not always easy to decide at which point of the classification a relation should be located exactly. For example, some relations are sequential both in space and time

and, at the same time, they may produce an effect. This aspect is described in (Kageura 1997) as follows : "Within the classificatory organization of concept systems, we have to distinguish two different kinds of multidimensionality : multidimensionality introduced by the co-existence of different means of concept classification, i.e., generic/specific, part/whole, and type/value ; multidimensionality introduced by the application of different types of characteristics or facets at the same level in the generic/specific structure. Generic/specific and part/whole classifications have been widely accepted in terminological studies, as these relations are clearly recognizable among entity concepts, while type/value relations have not been given much attention as a means for classificatory organization of concepts". If we were to include all relation categories resulting from such combinations in a taxonomy, the classification of non-hierarchical relations would grow exponentially, assuring the granularity of the classification but preventing flexibility and clarity. Thus, an approach is needed which on one side relies on a solid theoretical foundation, and, on the other side, assures a certain degree of flexibility in the representation of specific domains.

In this paper we take up the classification proposed by (Jouis 2007) into "situation statique ("état de choses") et situation dynamique (modification et changement dans le domaine)" (Jouis 2007). If we look carefully at the classifications above, we may notice that, even if under different designations, they all describe relations that modify the subject field in some way and to same extent. According to such classifications, concepts belonging to the same subject field may have a temporal or spatial succession or they can affect each other in a causal, functional or more generic way. The division between static and dynamic relations allows us to comprise thematic relations in the classification, since the definition of dynamic situations of (Jouis 2007) is broad enough to include domain-dependent relations. From a terminological point of view, the definition of dynamic relations as relations modifying the subject field faces us with two important issues. First of all, concepts succeeding and affecting each other determine a multidimensional concept system, which has to be represented accordingly : "Since the characteristics of a concept are frequently specified from different points of view or facets (function, material, shape, weight, etc.) a set of characteristics that constitutes a concept is normally multidimensional. From this point alone, we can expect a concept system to be multidimensional" (Kageura 1997). Secondly, the question arises whether and how the characteristics of concept influencing each other are affected. Indeed, concept relations determine the intension of the concepts taking part in the relation and therefore their definition : "The intensional

definition should be based on the concept relations determined during analysis. A definition based on a generic relation shall state the generic concept sharing the same dimension, either immediately above or at some higher level, followed by the essential characteristics that differentiate the given concept from coordinate concepts in a generic concept system" (ISO 704, 2000). According to the standard (DIN 2330), "Kann eine Definition nicht über hierarchische Beziehungen erstellt werden, ist es ausnahmensweise zulässig, nichthierarchische Begriffsbeziehungen zur Formulierung heranzuziehen. So könnte z.B. ein Produkt über seine genetische Beziehung zum Produzenten definiert werden. BEISPIEL : Gammastrahlung entsteht durch eine Isotopenkollision" (DIN 2330). This standard states the possibility that non-hierarchical relations are used in definitions, but still it does not state precisely how. According to (Arntz et al., 2004), "dynamische Begriffe (Pilke 2000 : 179ff) sind jedoch vielfach Teil eines Beziehungsgeflechts, das sich nur in der flexibleren Form eines nichthierarchischen Begriffssystems darstellen läßt. Bei der Strukturierung und Darstellung solcher Systeme wird grundsätzlich die gleiche Methodik angewandt wie bei den hierarchischen Systemen, diese wird jedoch von Fall zu Fall an die speziellen Gegebenheiten des konkreten Falles angepasst". If we are to modify the method used for hierarchical relations so as to apply it to non-hierarchical relations and to the concept involved, we have first of all to determine how non-hierarchical relations can affect the intension of concepts and how multidimensionality can be addressed in terminological definitions. Still in question is whether definitions deriving from dynamic relations can be written according to general patterns or to domain-dependent patterns.

To sum up what is outlined above, we can state that dynamic relations describe changes in the subject field and mostly imply multidimensionality ; that they cannot be represented with a hierarchical concept system, and that their description requires the standard method to be adjusted according to the specific case. Thus, we can maintain that dynamic relations can be both domain-independent (such as, for example, causal relations) and highly domain-dependent, given that they also include many domain-specific thematic relations. Therefore, we may summarise the above stating that dynamic relations are relations describing the way concepts belonging to the same subject field affect each other at any level of specification.

If we consider such definition a starting point to work with dynamic relations, we may use the ISO recommendation for intensional definitions as a basis for writing definitions according to the requirements of dynamic relations: a definition based on a dynamic relation shall state the concept

taking part in the relation, followed by the way it is affected by the related concepts. In defining the concepts affected by a dynamic relation this way, two elements are to be highlighted : the cardinality of the relation, i.e., the number of concepts involved (to say it formally, the relation arguments), and the context of the concept, i.e., the situation to which the relation belongs to. Both elements can be represented formally by conceptual graphs and translated respectively into predicates and contexts in other forms of logics. Such a definition could be a trade-off between the traditional intensional definition and the recent operationalisation needs.

3. Representation of dynamic concept relations

Using conceptual graphs as a representation tool for dynamic concept relations has several advantages. As outlined above, they are an intermediary notation between highly informal and highly formal representation systems; they can be expressed in a quite intuitive display form or translated to predicate calculus and other forms of logic. However, it is important that conceptual graphs are well built ; otherwise they can cause serious problems if translated or implemented into a knowledge representation system. For this reason, a short tutorial may be necessary for terminologists who would like to work with conceptual graphs. Their difficulty degree is still much lower than that of other knowledge representation systems.

Conceptual graphs (CGs) are a "network of concepts and conceptual relations that describe the domain roles" (Sowa 1976). They are an intensional formalism as the meaning of the relation they express is called the intension, while the surrogates corresponding to the concepts involved are called the extension. We may notice a similarity with intension and class in terminology, being intension the set of characteristics of a concepts and the class the set of objects the concept refers to. Terminologies deal with the concepts of a subject field and their designations, defining the concept with their relation to the other concept belonging to the subject field ; CGs offer a representation for the concept network based on *roles*. CGs derive from Peirce's existential graphs, but, "beside Peirce's primitives, conceptual graphs provide means of representing *case relations*, *generalized quantifiers*, *indexicals* and other aspects of natural language" (Sowa 1999).

3.1. Brief overview on CGs

A more formal definition of CGs is supplied in (Sowa 1999) : "A *conceptual graph* g is a bipartite graph that has two kinds of nodes called, *concepts* and *conceptual relations*". Concepts are represented by boxes, while relations are represented by circles, for example :

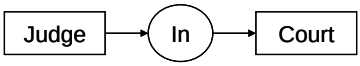


Figure 1. Conceptual graph for the sentence "A judge is in a court"

The default quantifier used in CGs is the existential quantifier ∃. The arrows between the boxes and the circle are called *arcs* and they are said to belong to the relation, not to the concept ; every arc links the concepts of the graph to the relation. The number of arcs determines the *valence* of the relation, i.e., a nonnegative integer number representing the number of the relation arguments. A relation with only one arc is said monadic, a relation with two arcs dyadic, a relation with three arcs triadic, and so on. CGs having concepts and relations can be linked together, if they feature one identical concept, by overlaying the identical concepts :

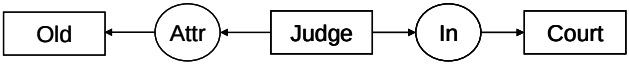


Figure 2. Conceptual graph for the sentence "An old judge is in a court"

Concepts feature a type and a referent. Types are surrogates for classes, while referents are surrogates for individuals (instances). With a "terminological terminology", we could say that types are the proper concepts, while referents are the objects referred to by the concept. The designation coincides in this case with the concept graphically, as it is written in the box. Concept types may be designated either by a type label or by a lambda expression. Usually, types precede referents and are followed by a colon. If no referent is indicated, the existential quantifier ∃ is supposed to rule. CGs thus allow to represent concepts (or classes) and referents (or instances). Types can be defined for relations, too ; furthermore, relations may have a *signature*, i.e., a constraint on the concepts that can take part in the relation.

By means of concept and relation types it is also possible to express hierarchies. To this aim it is necessary to introduce the lambda expression *e*, which may be defined as "a conceptual graph, called the body of *e*, in which *n* concepts have been designated as formal parameters of *e*" (Sowa 1999). The lambda expression is usually indicated with the Greek letter λ

and, if more formal parameters have been established, a number. Hierarchies are created with subtype relations, where the symbol ≤ indicates subtype, the symbol < indicated proper subtype, the symbol ≥ indicates supertype and the symbol > indicates proper supertype. The same rules and notations are applied to hierarchies of relation types. "The definitional mechanism introduces new type labels, whose place in the hierarchy is determined by their definitions […]. The formal parameter is always a supertype of the newly defined type : Farmer ≥ MaineFarmer. As an alternate notation, type labels can be defined with the keyword type and a variable" (Sowa 1999). Finally, CGs can also represent contexts : "A context $\mathscr{C}$ is a concept whose designator is a nonblank conceptual graph g" (Sowa 1999). In the context $\mathscr{C}$, it is possible to nest other conceptual graphs, as in the following example :

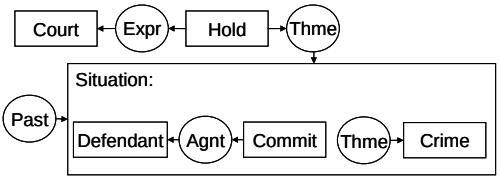


Figure 3. CG for "A court holds that a defendant committed a crime"

Identity between concepts outside the context and concepts nested in the context can be established with the so-called co-reference link, which is represented by a dotted line in the display form and by a x* in the linear form :

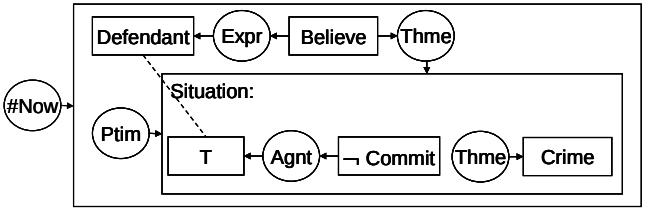


Figure 4. CG for "A defendant believes that he did not commit a crime"

In this example, the symbol ¬ indicates negation. The dotted line indicates the linked concepts are the same. This example illustrates in addiction that time can be represented in CGs. Many different notations have been proposed for time representation; an overview is given in (Schärfe 2003).

In conclusion, we can state that CGs can represent concepts, relations, and referents; concepts and relations can be arranged in hierarchies by means of their type labels ; last but not least, CGs can also represent contexts and time dimension.

3.2. Case study

In the following section, conceptual graphs are used to represent dynamic concept relations in the law field. For this case study, we chose three particularly interesting terms from the terminology collected during the last edition of the Specialization degree course and University Master degree course in Legal Translation of the University of Genoa. The course was divided into thematic units corresponding to the different branches of the law ; for every thematic unit (e.g. "criminal law") students were taught lessons of translation, law and terminology. At the end of every thematic unit, students had to produce a glossary including a number of multilingual terminology records. For this case study, we are analyzing some Italian terms featuring subtle differences from each other, which cannot be expressed in terms of characteristics deriving from hierarchic relations. The subtlety of the term differences caused the students a lot of troubles in identifying the right English equivalents. Beside proposing a way of representing dynamic concept relations, this case study is aimed to illustrate how CGs can make the terminological analysis easier.

3.3. Diffamazione

The first term analyzed is *diffamazione*. According to the Italian Penal Code, *diffamazione* is a crime and is defined as follows: "Chiunque, fuori dei casi indicati nell'articolo precedente, comunicando con piu' persone, offende l'altrui reputazione, e' punito con la reclusione fino a un anno o con la multa fino a lire due milioni […]". It is clear that this article cannot be considered a terminological definition, as it conveys a lot of information which is not essential and, at the same time, does not explicitly states other essential information. For example, the noun phrase *l'altrui reputazione* has the implicit meaning that the person offended is not present at the time of the offence. This is an essential characteristic of the concept underlying the term *diffamazione*, as it turns out to be one of the characteristics differentiating it from other terms of the same domain. Such information has to be stated in an explicit way in any knowledge representation system. At this point, we would like to recall the above mentioned definition of ontology as "an explicit specification of a conceptualization" (Gruber 1993). Thus, it is important that CGs, as an intermediate notation between plain text and formal knowledge representation systems, represent such information in an adequate way. The extraction of the relevant concepts from the above shows how implicit is the information we have. The only concepts we may obtain from text mining are [Offendere] and [Reputazione]. This means that a CG should make explicit the underlying information, as shown in the following graph :

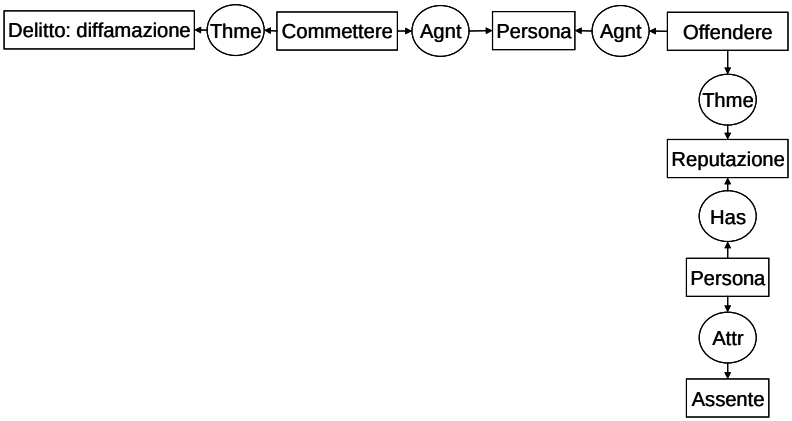


Figure 5. CG for "Diffamazione è il delitto che commette chi offende la reputazione di una persona assente"

This graph conveys the essential information for the terminological definition in an explicit way, representing both static and dynamic relations: the indication of a hierarchy at the beginning of the graph belongs to the first kind of relation; the colon allows indeed to state that *diffamazione* is a subtype of *delitto*. Thus, this CG represents both the static and the dynamic dimension of the concept intension, thus adding information to the classical intensional definition as "generic concept sharing the same dimension, either immediately above or at some higher level, followed by the essential characteristics that differentiate the given concept from the coordinate concept in a generic concept system" (ISO 704 2000). The essential characteristics differentiating the concept behind the term *diffamazione* derive from dynamic relations, which are represented in the CG by means of relation nodes. Relation nodes state the concepts of the intension and the way they interact in an explicit way. The presence of an apparently identical concept [Persona] does not represent a problem in CGs, as the identity of two concepts carrying the same name must be expressed with the coreference link. Thus, CGs allow managing concepts independently from their names, a feature that, however, should be taken into consideration when translating CGs into systems featuring unique name convention.

As an equivalent for *diffamazione*, some students proposed the English term *libel*, defining it as follows: "a written or oral defamatory statement that conveys an unjustly unfavourable impression". If we design a CG according to this definition, we may notice some differences in the intension:

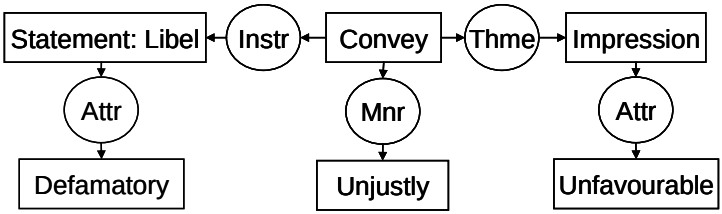


Figure 6. CG for "Libel is a defamatory statement that conveys an unjustly unfavourable impression"

First of all, it should be highlighted that this CG represents *libel* as an instrument to convey an unfavourable impression and not as an agent as *libel* is not an animate being, thus expliciting the relations in the intension and avoiding attributing intentionality to *libel* itself, which could cause serious problems in the translation to a KR system.

If we consider the content itself, we may notice that the concepts of offence, reputation and absence of the person being offended do not belong to the intension. Furthermore, a libel is defined as a *statement*, not even as a *crime*. This should led us to a reflection about the adequacy of the equivalent proposed, of the definition outlined, or, at least, about the differences of the two law systems involved. Thus, CGs provide a good feedback to check whether the conceptual analysis and the resulting definition are adequate, complete and efficient.

3.4. Ingiuria

The second term analyzed is *ingiuria*. According to the Italian Penal Code, *ingiuria* is a crime described as follows : "Chiunque offende l'onore o il decoro di una persona presente è punito con la reclusione fino a sei mesi o con la multa fino a lire un milione. […]". If we proceed as in the case of *diffamazione*, we obtain the following graph :

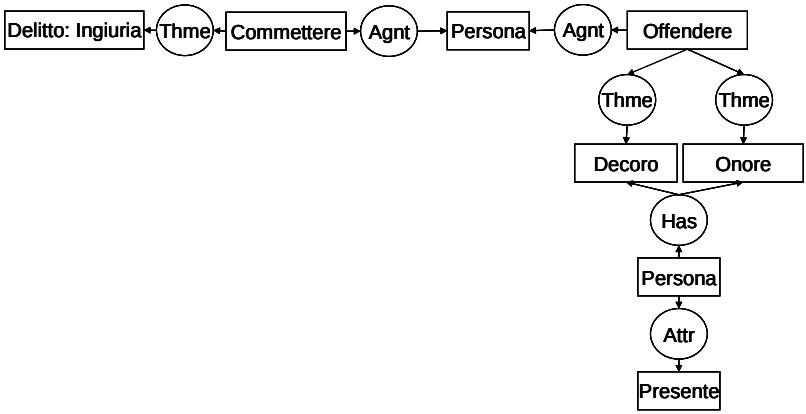


Figure 7. CG for "Ingiuria è il delitto che commette chi offende l'onore e il decoro di una persona presente"

This CG illustrates clearly the differences between the intensions of *diffamazione* and *ingiuria*, i.e., theme and attribute. Both graphs feature the same relations, but the relation values are different. Thus far the graph allows us to analyze not only the kind of relation underlying the essential characteristics of a concept, but also which values they assume. In this case, for example, the dynamic relation between [Offendere] and what is being offended is of the same type agent-theme, but the values are different: in the case of *diffamazione*, *reputazione* is theme of the action; in the case of *ingiuria*, *onore* and *decoro*. This is a crucial point for a further analysis of the specification level of thematic relations. We may notice that CGs frame dynamic relations within general relation types, to which more specific information is added in the boxes. On one side, such a representation enables to manage different specification levels at the same time depending on the focus of the analysis. On the other side, this division allows to represent the relation types in a language independent way (granted that English has been used as a basis for representing the relation nodes), while the boxes contain language-specific information, thus allowing interesting comparative analysis. From this point of view, a conceptual structure expressed in CGs is a clear output where relations, concepts and roles are stated in an explicit way, thus being an ideal basis for terminological analysis and comparison between possible equivalents. Such a bipartite structure gives terminologists an insight into the conceptual super-structure into which the domain-specific information can be mapped, thus allowing contrastive analysis focusing either on language or on the structure itself, even if we grant for the sake of argument that the CGs are designed in a consequent way with regard to both notation and *ratio*.

The equivalent for *ingiuria* proposed by the students is *insult*, which was defined as "expression, statement or behavior that is considered offensive, rude or degrading. An insult is an act that offends a person's sense of pride or dignity. Insults may be intentional or accidental".

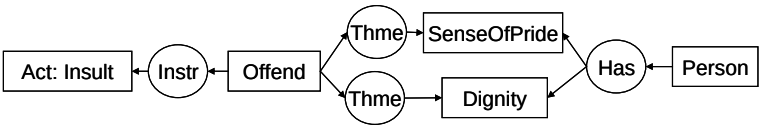


Figure 8. CG for "An insult is an act that offends a person's sense of pride or dignity"

The CG only represents the terminologically relevant information. The relations are quite close to the relations defined in the CG for *ingiuria* ; still, some concepts are missing or simply not defined (e.g. we have no information about the presence of the person being offended).

3.5. Calunnia

The last term of this case study is *calunnia*, which can be defined as follows on the basis of the Italian Penal Code : "Delitto commesso da chi, con i mezzi previsti dalla legge, accusi falsamente qualcuno di un reato, ovvero ne simuli a suo carico le tracce".

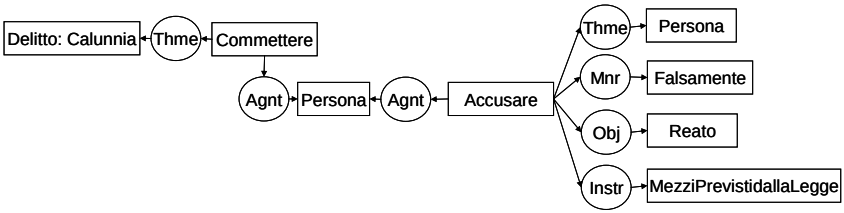


Figure 9. CG for "Calunnia è il delitto commesso da chi, con i mezzi previsti dalla legge, accusi falsamente qualcuno di un reato"

So far, we notice that the intension of *calunnia* is quite complex, as it features a 5-adic relation with an agent, a theme, a manner, an instrument and an objective (*reato*). The latter is the most controversial argument of the relation, as it describes the object of the charge, not the patient or the theme. Actually, this objective entails a whole situation, and could therefore be represented as a context.

The English equivalent proposed during the course is calumny, which was defined as follows : "A false charge or imputation".

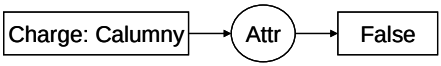


Figure 10. CG for "A calumny is a false charge"

This graph seems to represent a simpler intension if compared with *calunnia*; however, this is due to the fact that great part of the intension is entailed in the concept expressed by the term charge, which has to be made explicit accordingly :

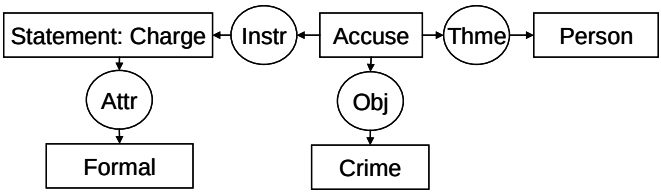


Figure 11. CG for "A charge is a formal statement to accuse a person of a crime"

At this stage we can restrict and join the graphs as follows :

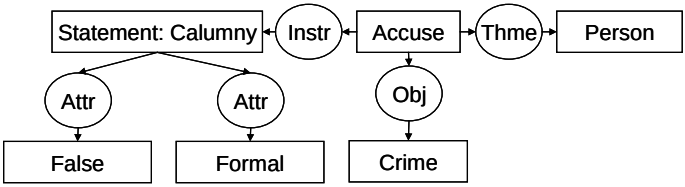


Figure 12. CG for "A calumny is a false, formal statement to accuse a person of a crime"

This graph allows a comparison in order to evaluate the equivalent proposed.

So far we have used CGs to represent the conceptual structure underlying terms which were already defined; still, from the observations we have made, it is clear that CGs are an extremely useful instrument to represent conceptual structures during the analysis of conceptual systems, with regard to both monolingual and multilingual terminology. Thus far, two methods of using CGs can be outlined. In the first place, CGs are used as a means for representing the conceptual structure, and in particular the dynamic concept relations, during the concept system analysis. This is the proper use of CGs as a representation means. This method allows terminologists to work with dynamic concept relations and complex subject fields to manage both hierarchies and thematic relations, thus encoding both static and dynamic aspects of the terminology they are working on. During terminological concept analysis, the rules outlined in (Sowa 1976) for designing and merging CGs may also apply. Firstly, the relevant concepts should be chosen and represented with the boxes (the criteria for the choice of the concepts are beyond the scope of this article ; nevertheless, we would like to mention the role that an ontological approach to terminology can play in this phase); secondly, the relations among them are made clear and represented with the corresponding

nodes. If any restriction has to be imposed, signatures and lambda expressions can be added. Finally, the conceptual structure obtained this way can be modified by joining, copying, etc. Upon comparison between the graphs representing different concepts, definitions can be written according to the intension outlined in the graphs – i.e., definitions should describe clearly the relations involved and highlight the relations differentiating concepts from each other according to the predicate representation sketched by means of CGs. In this way, during the conceptual analysis itself it is possible to develop a quite formal and semantic consistent representation of the conceptual structure underlying the analyzed subject field; dynamic concept relations are represented in a bipartite way, by means of a generic relation node and of a concept box which specifies the node. It may be noticed that the relation nodes, e.g. "Agnt", "Ptnt", etc. correspond to what is called thematic roles in semantics: "Linguists often talk of thematic roles or roles […], which have to do with the participants in the events or states the sentence describe" (Chierchia *et al.* 2000). If applied to CGs, θ-roles, as they are often called, help mapping thematic information to a more general structure having a quite solid theoretical basis. Indeed, "thematic roles can be thought of as kinds of specific roles, so we could define them formally as properties common to sets of specific roles" (Chierchia *et al.* 2000). θ-roles convey both syntactic and semantic information, as they illustrate syntactic positions and, at the same time, semantic argument structure, thus acting as an interface between syntax and semantics. We are not dealing with the interactions between syntax and semantics, which is far beyond the scope of this paper; nevertheless, we would like to highlight how using thematic roles for CGs relation nodes allows different levels of analysis. Indeed, CGs represent at the same time the conceptual structure of the subject field, the relations defining a concept, and some syntactic information about the terms chosen to express the concept.

In second place, CGs can be used as a check means, too, as showed in the examples above; indeed, if designed in a consequent way, they allow to compare intensions both in monolingual and in multilingual terminology. Still, in our opinion the first approach, i.e. using CGs during the conceptual analysis, permits saving a considerable amount of time, which is a quite important factor if we consider that terminology work is an extremely time-intensive task.

4. Conclusions and future work

In the introduction we outlined a brief overview of the previous work, which highlighted the lack of a systematic definition and classification of non-hierarchical concept relations, in particular thematic relations.

Starting from the division between static and dynamic relations proposed by (Jouis 2007), we defined dynamic relations as relations describing the way concepts belonging to the same subject field affect each other at any level of specification.

In the third section, we proposed a way of representing dynamic concept relations by means of conceptual graphs, being CGs an intermediate notation between plain text terminologies and formal knowledge representation systems. We outlined a brief introduction to CGs and highlighted how relation nodes can be expressed by thematic roles, which allows a mapping of domain-dependent relations into more general relations. Thus far, we underlined the possibilities offered by thematic roles in terms of syntactic and semantic analysis.

Finally, we presented a short case study that illustrates the representation of some dynamic concept relations in law field; in the reflections on the case study we outlined two methods of using CGs in terminology work, i.e. as a proper representation means during conceptual analysis and as a check tool of definitions and concept structure.

Thus far, some issues have been tackled which need further and deeper analysis.

In the first place, it must be investigated how to represent entailments by means of CGs in a way that can be also managed in terminology work. The procedure we adopted in the case of *calunnia* may not always apply ; as it is not possible (and not even reasonable) to repeat the characteristics of every super-ordinate concept. A way of representing inheritance and inference rules should be outlined, which also takes the needs of terminologists in consideration.

In the second place, future work may also deal with the problem of distinguishing concepts and their designations, as they overlap in the graphic representation proposed so far. Still, in our opinion this is a fundamental issue, as it is connected with the more general problem of language dependency of instruments which are supposed to be language-independent.

Another interesting research topic connected with this paper is the investigation of how terminological representations by means of CGs could be extended with Petri Nets to represent processes. Indeed, tackling

processes in terminology is a quite difficult task, as concepts are defined through different transitions and stages.

Finally, we would like to remark the necessity of continuing the theoretical research on dynamic concepts and non-hierarchical relations, in order to pin down what has been stated so far – and which steps still have to be taken to reach, if possible, a standard, shared approach to the topic.

5. Bibliographie

Arntz R., Picht H., Mayer F. (2004) : *Einführung in die Terminologiearbeit*, Hildesheim et al., Georg Olms

Cabré M.T. (1999) : *Terminology. Theory, Methods and Applications*, Amsterdam, John Benjamins

Chierchia G., McConnelly-Ginet S. (2000) : *Meaning and Grammar : an introduction to semantics*, Cambridge, Mass., MIT Press

De Nova G. (2004) : *Codice Civile e leggi collegate – con CD-Rom 4 codici e leggi complementari*, Bologna, Zanichelli

DIN 2330 (1993) : *Begriffe und Benennungen. Allgemeine Grundsätze. Deutsche Normen*. Berlin, Köln, Beuth

DIN 2342 (1992) : Teil 1, *Begriffe der Terminologielehre: Grundbegriffe*, Berlin, Köln, Beuth

Felber H. (1984) : *Terminology Manual*, Paris, Unesco

Gerzymisch-Arbogast H. (1996) : *Termini im Kontext .Verfahren zur Erschließung und Übersetzung der textspezifischen Bedeutung von fachlichen Ausdrücken*, Tübingen, Narr

Gruber T. R. (1993) : "A translation approach to portable ontology specifications", *Knowledge Acquisition*, 5, 199-220

ISO 704 (2000) : *Terminology work – Principles and methods*, Geneva, ISO

Jouis C. (2007) : "Un système logique pour les relations sémantiques entre concepts", in *Terminologie & Ontologie : Théories et Applications, Actes de la conférence TOTh 2007*, Annecy, Porphyre

Kageura K. (1997) : "Multifaceted/Multidimensional Concept Systems", in Wright S. E., Budin G. *Handbook of Terminology Management*, vol. I, Amsterdam, John Benjamins

Kageura K. (2002) : *The Dynamics of Terminology. A Descriptive theory of Term Formation and Terminological Growth*, Amsterdam, John Benjamins

Nuopponen A. (1994) : "Causal relations in terminological knowledge representation", *Terminology Science and Research*, vol. 5, n. 1, 36-44

Nuopponen, A. (1994) : "Wüster revisited: On Causal Concept Relationships and Causal Concept Systems", in: Brekke A. *et al.* (eds.), *Applications and Implications of Current LSP Research, Proceedings of the 9th European Symposium on LSP*, vol. II, 532-539, Bergen, Fagbokforlaget

Pilke N. (2000) : *Dynamiska fackbegrepp. Att strukturera vetande om handlingar och händelser inom teknik, medicin och juridik*. Vaasa, Vaasan yliopisto

Schärfe H., Øhrstrøm P. (2003) : "Representing Time and Modality in Narratives with Conceptual Graphs", De Moor, A. et al. (eds.) *Conceptual Structures for Knowledge Creation and Communication*, 201-214, Springer Verlag

Sowa J. F. (2000) : *Knowledge Representation. Logical, Philosophical, and Computational Foundations,* Pacific Groove, CA, Brooks/Cole

Sowa J. F. (1976) : "Conceptual graphs for a database interface", *IBM Journal of research and development*, vol. 20, n. 4, 336-357

Wüster E. (1974) : "Die Umkehrung einer Begriffbeziehung und ihre Kennzeichnung in Wörterbüchern", in *Nachrichten für Dokumentation*, vol. 25, n. 6, 256-263

A propos des auteurs

Chiara Messina

Università degli Studi di Genova / Universität Wien
Piazza S. Sabina, 2 / Gymnasiumstraße 50
IT-16124 Genova / A-1190 Wien
messina.chiara@gmail.com

Construction et alignement d'ontologies pour évaluer le risque alimentaire

Liliana Ibanescu, Patrice Buche, Juliette Dibie-Barthélemy

Résumé : Nous présentons dans cet article un retour sur notre expérience de construction et d'alignement d'ontologies appliqué au domaine du risque alimentaire. Plus précisément, l'application concerne l'évaluation de l'exposition d'une population cible à un risque alimentaire réalisée avec le logiciel CARAT. Cette évaluation repose sur deux sources de données : une source de données de contamination chimique des aliments, appelée CONTA, et une source de données de consommation, appelée CONSO. Pour pouvoir calculer l'exposition d'une population cible à un risque alimentaire, les produits alimentaires de la source CONTA doivent être mis en correspondance avec les produits alimentaires de la source CONSO. Nous présentons dans ce papier le logiciel CARAT, ses sources de données, leurs référentiels, appelés ontologies, et le problème d'alignement de ces ontologies. Nous rappelons la méthode que nous avons proposée pour aligner ces ontologies, qui repose sur le modèle des graphes conceptuels. Nous présentons enfin les limites de cette méthode et proposons une nouvelle formalisation des ontologies pour les contourner.

Mots-clés : Représentation des connaissances, intégration de données, construction d'ontologie, alignement d'ontologies, risque alimentaire

1. Introduction

L'intégration des données permet d'accéder de manière unifiée à des sources multiples, hétérogènes en syntaxe, schéma ou sémantique. Le but de l'intégration de données est de faciliter l'accès et la réutilisation d'un ensemble de sources. La notion centrale sur laquelle reposent les recherches actuelles en intégration *sémantique* des données est la notion d'ontologie (Wache *et al.* 2001), (Ziegler *et al.* 2004), une ontologie étant un vocabulaire qui décrit un domaine d'intérêt et attribue un sens à ses termes (Gruber 1993). L'alignement des ontologies (Euzenat *et al.* 2007) est une solution pour résoudre l'hétérogénéité sémantique des données, en proposant des correspondances entre des entités sémantiques similaires de différentes ontologies.

Nous présentons dans cet article un retour sur notre expérience de construction et d'alignement d'ontologies dans le domaine du risque alimentaire. Plus précisément, nous présentons le logiciel CARAT (Buche *et al.* 2006) qui permet d'évaluer l'exposition d'une population cible à un risque alimentaire. Ce logiciel repose sur deux sources volumineuses de données : une source de données de contamination chimique des aliments, appelée CONTA, qui contient environ 2600 produits alimentaires différents, et, une source de données de consommations, appelée CONSO, qui contient environ 500 produits alimentaires. Dans CARAT, l'utilisateur est confronté au problème de l'intégration des deux sources de données qui reposent sur des référentiels distincts.

Nous présentons dans cet article l'architecture du logiciel CARAT, ses sources de données et leurs référentiels associés, appelés dans la suite ontologies, ainsi que le problème d'alignement de ses ontologies. Nous présentons ensuite notre méthode d'alignement des ontologies (Buche *et al.* 2008), qui repose sur le modèle des graphes conceptuels, et discutons des deux problèmes majeurs soulevés par cette méthode : i) un problème de modélisation : la description des produits alimentaires est au niveau instance (existentiel) ; ii) un problème concernant les résultats : les résultats ne sont pas satisfaisants et ils ne peuvent pas être comparés avec l'état de l'art. Enfin, nous proposons une nouvelle formalisation, en OWL (Dean *et al.* 2004), (Smith *et al.* 2004), des ontologies, et nous présentons la méthodologie utilisée pour construire ces ontologies à partir des sources de données existantes. Nous concluons et nous présentons les perspectives.

2. Le logiciel CARAT

Le logiciel CARAT (Buche *et al.* 2006) développé dans notre unité de recherche permet d'évaluer l'exposition des consommateurs à un risque alimentaire à partir de deux sources de données : une source de données de contamination chimique des aliments, appelée CONTA, et une source de données de consommation, appelée CONSO. Pour pouvoir calculer un niveau d'exposition, il faut définir trois éléments i) le contaminant (par exemple le méthyle-mercure); ii) le groupe d'individus étudié, appelé population cible (par exemple les enfants de moins de 15 ans); iii) le groupe de produits alimentaires (par exemple le groupe des poissons). Ensuite, il faut choisir la méthode de calcul statistique à utiliser pour évaluer le niveau de l'exposition de la population cible au contaminant étudié pour le groupe de produits alimentaires donné.

La Figure 7 présente l'architecture du système d'intégration de données du logiciel CARAT. Dans les deux sources de données, CONSO et CONTA, les noms des produits alimentaires ne sont pas identiques. Pour que CARAT puisse croiser les données de consommation avec les données de contamination, il est nécessaire de mettre en correspondance les noms de produits des référentiels de ces deux sources. Cette mise en correspondance est appelée alignement des ontologies CONSO et CONTA dans la Figure 7. Cet alignement permet ensuite d'interroger les deux sources de données en utilisant un seul vocabulaire. L'opération d'alignement repose sur la création de groupes de produits.

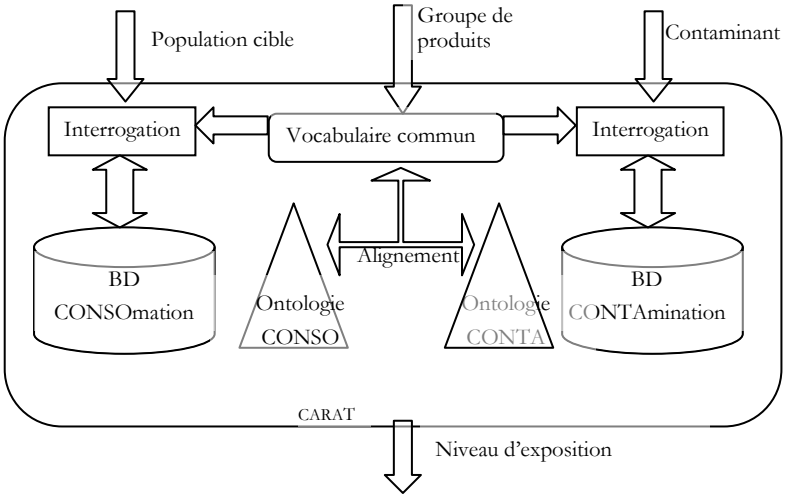


Figure 7. Architecture du système d'intégration de données du logiciel CARAT

La Figure 8 montre une capture d'écran du processus de construction d'un groupe d'aliments étudiés. Dans cette opération, un groupe de produits est défini comme deux ensembles de produits alimentaires : le premier ensemble contient les noms de produits du référentiel de la source CONSO, le second ceux du référentiel de la source CONTA. Cette opération permet donc d'aligner les référentiels de noms de produits des deux sources au niveau de granularité 'Groupe de produits'. Ces correspondances entre les noms de produits de la source CONSO et les noms de produits de la source CONTA sont actuellement établies à la main, par l'utilisateur du logiciel CARAT, ce qui représente un travail fastidieux. En effet, pour réaliser cette opération, l'utilisateur doit naviguer dans des référentiels volumineux : environ 500 termes pour la source CONSO et environ 2600 termes pour la source CONTA. Il est donc nécessaire de mettre à disposition de l'utilisateur un outil d'alignement semi-automatique des noms de produits alimentaires des deux sources afin de le soulager dans la tâche de création des groupes de produits.

Figure 8. Alignement entre les produits consommés et les produits contaminés pour définir le groupe de produits *Fruits et légumes secs*

Dans un premier temps, nous avons regardé s'il était possible d'effectuer une comparaison lexicale des noms de produits des deux ontologies CONSO et CONTA pour réaliser l'alignement. Le résultat de cette étude était loin d'être satisfaisant : l'alignement lexical des noms de produits a donné 13 correspondances exactes (égalité des chaînes de caractères) et 50

correspondances en comparant des sacs de mots (ensemble de mots lemmatisés contenus dans la chaîne de caractères correspondant au terme) sur un total de 3248 correspondances à trouver. Ce résultat médiocre s'explique principalement par la différence dans le niveau de granularité de la description des produits alimentaires dans les deux ontologies. L'ontologie CONSO (environ 500 termes désignant des noms de produits) est en effet beaucoup moins détaillée que l'ontologie CONTA (environ 2600 termes). Par exemple, *Poisson frais* dans l'ontologie CONSO doit être mis en correspondance avec *Cabillaud cru* dans l'ontologie CONTA. Nous avons alors décidé d'exploiter une description complémentaire des produits alimentaires disponible dans les deux sources de données pour effectuer l'alignement.

En effet, dans les deux sources de données, CONSO et CONTA, un produit alimentaire est décrit par une liste de triplets (produit, caractéristique, valeur). La Figure 9 présente des exemples de description. L'ensemble des triplets d'une source constitue l'ontologie des produits alimentaires associée à cette source.

Description d'un produit consommé **Poisson frais**	Description d'un produit contaminé **Cabillaud cru**
(Poisson frais, Présentation, Entier) (Poisson frais, Quel poisson ?, Cabillaud) (Poisson frais, Quel poisson ?, Saumon)	(Cabillaud cru, Origine de l'ingrédient principal, Morue ou cabillaud) (Cabillaud cru, Etat physique ou forme, Entier de forme naturelle) (Cabillaud cru, Méthode de conservation, Conservé par stockage en réfrigérant)

Figure 9. Exemples de triplets de description (produit, caractéristique, valeur) pour un produit consommé et respectivement un produit contaminé

Plus précisément, la source de données de consommation CONSO est produite par TNS Worldpanel France qui fournit un référentiel de 472 noms de produits alimentaires décrits avec 30 505 triplets (produit, caractéristique, valeur). Le nombre de caractéristiques distinctes est de 141 et celui de valeurs distinctes est de 15 116. Dans la source de données de contamination, CONTA, les produits alimentaires sont indexés selon le référentiel REGAL de l'AFSSA (Agence Française de Sécurité Sanitaire des Aliments) et décrits en utilisant Langual (Ireland *et al.* 2000), un thesaurus multilingue de description des aliments. Les 2595 noms de produits alimentaires sont décrits avec 25 069 triplets (produit, caractéristique, valeur). Le nombre des caractéristiques distinctes est de 14 et celui des valeurs distinctes est de 939. Toutes les valeurs d'une

caractéristique sont organisées dans une hiérarchie par la relation "sorte-de" (voir la Figure 10).

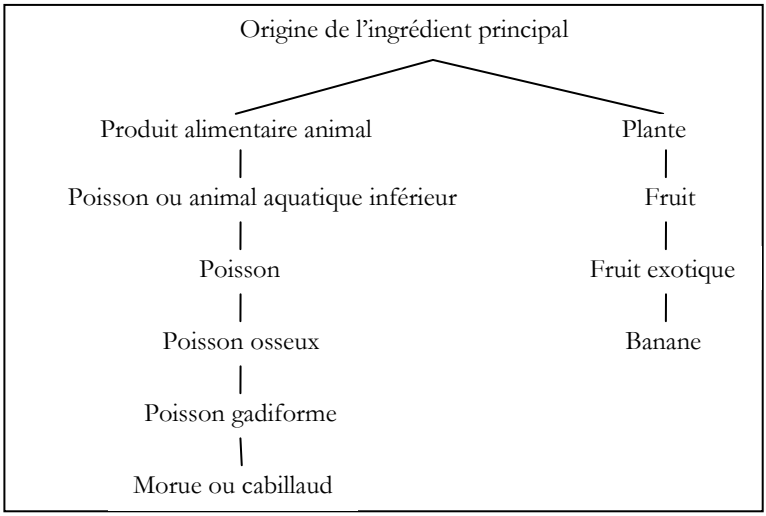


Figure 10. Un fragment de la hiérarchie des valeurs pour la caractéristique *Origine de l'ingrédient principal*

3. Alignement d'ontologies en utilisant les graphes conceptuels

Dans (Buche *et al.* 2008) nous proposons une méthode d'alignement d'ontologies, reposant sur le modèle des graphes conceptuels (Mugnier *et al.* 1996) dont nous rappelons dans cette section les grands principes.

Nous avons choisi le modèle des graphes conceptuels pour les raisons suivantes : i) les triplets (produit, caractéristique, valeur) de description d'un produit peuvent être aisément représentés sous la forme de graphes conceptuels; ii) la taxonomie des valeurs peut être directement représentée dans le support terminologique du modèle ; iii) la relation de projection du modèle peut être utilisée pour retrouver les alignements en exploitant les taxonomies de valeurs.

Les deux ontologies étudiées dans la méthode d'alignement ne sont pas symétriques. En effet, l'ontologie CONTA est stable dans le temps, elle est alors considérée comme l'ontologie cible, alors que l'ontologie CONSO évolue fréquemment et elle est considérée comme l'ontologie source. Le principe de la méthode proposée consiste à comparer la description d'un produit de l'ontologie CONSO à celle d'un produit de l'ontologie

CONTA en utilisant l'opération de projection du modèle des graphes conceptuels. Pour ce faire, les deux produits doivent être décrits dans le même vocabulaire. L'ontologie CONTA étant l'ontologie cible, nous avons choisi de traduire de manière semi-automatique chaque produit de l'ontologie CONSO dans le vocabulaire de l'ontologie CONTA. Plus précisément, la "traduction" d'un produit de l'ontologie CONSO est obtenue en remplaçant sa description en termes de caractéristiques et valeurs de CONSO par une description en termes de caractéristiques et valeurs similaires de l'ontologie CONTA. Les 4 grandes étapes de l'algorithme sont décrites ci-dessous.

La première étape consiste à trouver un alignement entre les caractéristiques des deux ontologies, en prenant en compte leurs valeurs. Par exemple : la caractéristique *Présentation* de l'ontologie CONSO est mise en correspondance avec la caractéristique *Etat physique ou forme* de l'ontologie CONTA, et la caractéristique *Quel poisson ?* est mise en correspondance avec la caractéristique *Origine de l'ingrédient principal*. L'établissement des correspondances entre caractéristiques est fait de manière semi-automatique à partir de similarités lexicales trouvées entre les valeurs qui leur sont associées. La validation manuelle de cette mise en correspondance par l'utilisateur est possible compte tenu du nombre restreint de caractéristiques. A la fin de cette étape, chaque caractéristique de l'ontologie CONSO est associée à une caractéristique de l'ontologie CONTA et chaque valeur de l'ontologie CONSO est associée à une liste de valeurs de l'ontologie CONTA pondérées par leur similarité lexicale (avec la valeur de l'ontologie CONSO).

Dans la deuxième étape, chaque description de produit de l'ontologie CONSO est "traduite" et représentée par un graphe conceptuel flou (Thomopoulos *et al.* 2003). Le support terminologique utilisé est défini de la manière suivante : i) l'ensemble des types de concepts contient les noms des produits des deux ontologies, l'ensemble des caractéristiques de l'ontologie CONTA, la hiérarchie des valeurs de l'ontologie CONTA et le type de concept *ValNum* (qui permet de représenter les valeurs numériques) ; ii) l'ensemble des types de relations contient les quatre types de relation *APourCarac*, *APourValeur*, *EstAnnoté* et *APourScore* ; iii) l'ensemble des marqueurs individuels contient les valeurs réelles. Le produit *Poisson frais* de l'ontologie CONSO (voir la Figure 9) est " traduit " et représenté par le graphe conceptuel de la Figure 11, avec les caractéristiques et les valeurs de l'ontologie CONTA. Dans cette traduction, chaque caractéristique de l'ontologie CONSO est remplacée par celle qui lui est associée dans l'ontologie CONTA et chaque valeur de

l'ontologie CONSO est remplacée par la liste pondérée de valeurs de CONTA associée.

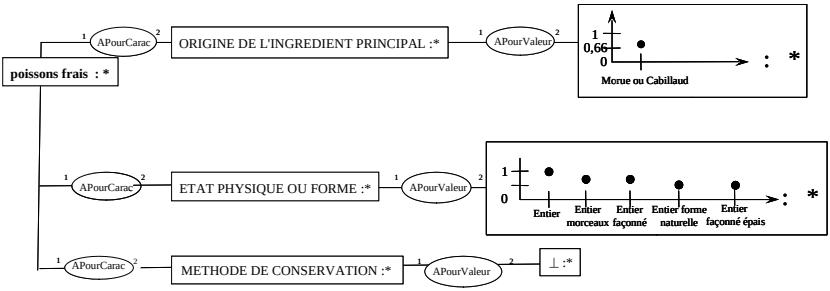


Figure 11. Représentation du produit *Poisson frais* de l'ontologie CONSO sous forme de graphe conceptuel

La troisième étape consiste à créer des règles pour représenter les produits de l'ontologie CONTA. A chaque produit de l'ontologie CONTA est associée une règle dont la partie condition contient la description du produit et dont la partie conclusion contient l'annotation qui est ajoutée à un graphe représentant un produit de l'ontologie CONSO si la règle est déclenchée. La règle de la Figure 12 est associée au produit *Cabillaud cru* de l'ontologie CONTA.

La quatrième étape consiste à appliquer toutes les règles créées dans l'étape précédente à l'ensemble des graphes associés aux produits de l'ontologie CONSO. L'application d'une règle a pour résultat d'ajouter au graphe associé à un produit de l'ontologie CONSO une annotation indiquant le produit de l'ontologie CONTA avec lequel il est aligné ainsi qu'un score d'alignement.. Par exemple, suite à l'application de la règle de la Figure 12, le produit *Poisson frais* de l'ontologie CONSO est annoté avec le produit *Cabillaud cru* de l'ontologie CONTA avec un degré de 0.5.

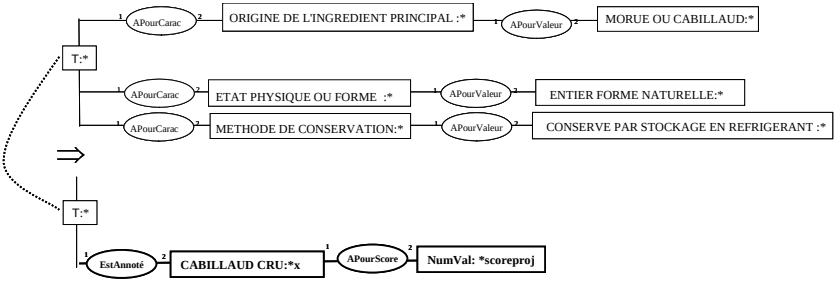


Figure 12. Règle associée au produit *Cabillaud cru* de l'ontologie CONTA

4. Limites de notre méthode d'alignement d'ontologies

Nous avons identifié deux problèmes majeurs soulevés par cette approche. Le premier problème concerne la modélisation. Si l'on met de côté l'extension floue proposée, la description sous forme de graphe conceptuel (voir la Figure 11) de la traduction d'un produit de l'ontologie CONSO, dans la formalisation proposée par (Mugnier *et al.* 1996), a une interprétation de fermeture existentielle conjonctive en logique des prédicats (il existe un poisson frais, …). Cela ne correspond pas à l'intuition que l'on peut avoir de la description d'un produit auquel on voudrait pouvoir associer une interprétation universelle (quel que soit le poisson frais, il a pour description…). Le deuxième problème concerne les résultats expérimentaux obtenus. Dans (Buche *et al.* 2008) nous avons calculé la précision[1] (2,08%) et la couverture[2] (77,04%) en comparant les résultats obtenus par notre méthode d'alignement à un alignement de référence que nous avons construit (manuellement) entre l'ontologie CONSO et l'ontologie CONTA. Des travaux ultérieurs, non encore publiés, nous ont permis d'améliorer la précision à 29,93%, avec une couverture de 59,85%, en ajoutant une étape supplémentaire de vérification de contraintes, pour éliminer des alignements incorrects. Ces résultats ne sont pas satisfaisants, dans la mesure où le travail de validation des résultats intermédiaires de l'alignement demandé à l'utilisateur est trop important. La raison essentielle que nous avons identifiée et qui pourrait expliquer ces mauvais résultats est l'existence de caractéristiques et de valeurs peu discriminantes mais génératrices de beaucoup de bruit. Par exemple, la caractéristique *Présentation* de l'ontologie CONSO qui, à la suite de la première étape de l'algorithme d'alignement, est mise en correspondance avec la caractéristique *Etat physique ou forme* de l'ontologie CONTA, permet par la suite de retrouver des alignements entre le produit *Poisson frais* de CONSO avec tous les produits de CONTA qui sont décrits avec le mot *Entier* parmi lesquels figurent tous les fromages. Cependant, pour pouvoir évaluer nos résultats expérimentaux, il faut pouvoir les comparer avec les résultats d'autres méthodes d'alignement. Or nos ontologies ne sont pas représentées dans un format adapté (RDF(S) ou OWL) pour pouvoir les aligner avec des outils existants d'alignement d'ontologies, comme ceux recensés dans (Euzenat *et al.* 2007), (Kalfoglou *et al.* 2005) et (Noy 2004).

1 La précision est le pourcentage d'alignements corrects générés par rapport à tous les alignements générés.

2 La couverture est le pourcentage d'alignements corrects générés par rapport à tous les alignements corrects.

Dans les deux sections suivantes nous présentons une nouvelle approche pour modéliser les ontologies CONSO et CONTA en OWL à partir des sources de données.

5. Construction de l'ontologie CONTA en OWL

Dans la source de données CONTA, qui est stockée sous la forme d'une base de données relationnelle, l'information concernant les produits alimentaires contaminés se trouve dans les trois tables suivantes (les clés primaires sont soulignées) :

produit_REGAL(nomProduit, nomFamille)

description(nomProduit, nomCaracteristique, valeurCaracteristique)

taxonomie_valeurs(valeur, valeurPere)

Les classes, les propriétés et les restrictions OWL associées à ces tables sont définies selon les étapes suivantes :

- Trois classes génériques, sous-classe de la classe Object, sont définies en OWL : *Famille*, *Produit*, *Valeur*.
- Pour chaque nom de famille 'nomFamille' distinct de la table produit_REGAL une nouvelle classe appelée 'nomFamille'[3] est créée. Cette classe est une sous-classe de la classe générique *Famille* et son étiquette est 'nomFamille'. Par exemple, pour la famille *Poissons et batraciens* le code généré en OWL est le suivant :

```
<owl:Class rdf:about= "#Poissons_et_batraciens">
   <rdfs:label xml:lang="fr">  Poissons et batraciens </rdfs:label>
   <rdfs:subClassOf rdf:resource="#Famille"/>
</owl:Class>
```

- Pour chaque nom de produit 'nomProduit' de la table produit_REGAL une nouvelle classe appelée 'nomProduit' est créée. Cette classe est une sous-classe de la classe générique *Produit* et son étiquette est 'nomProduit'. Par exemple, pour le produit *Cabillaud cru* le code généré en OWL est le suivant :

```
<owl:Class rdf:ID="Cabillaud_cru">
        <rdfs:label xml:lang="fr">Cabillaud, cru</rdfs:label>
        <rdfs:subClassOf rdf:resource="#Produit"/>
</owl:Class>
```

3 Aucun espace n'étant autorisé dans le nom des classes en OWL, on considère, dans la suite, que chaque espace dans le nom d'une classe est remplacé par le caractère spécial '_'.

- Pour chaque valeur 'valeur' d'une caractéristique décrite dans la table taxonomie_valeurs, une nouvelle classe de nom 'valeur' est créée. La représentation de la relation "sorte-de", codée par les tuples (valeur, valeurPere) de la table taxonomie_valeurs, est traduite en OWL par le fait que la classe 'valeur' est une sous-classe de la classe 'valeurPere'. Par exemple, le tuple (*Morue ou cabillaud*, *Poisson gadiforme*) de la table taxonomie_valeurs (qui code la hiérarchie de la Figure 4) et représenté en OWL par le code suivant :

```
<owl:Class rdf:about="#morue_ou_cabillaud">
   <rdfs:label xml:lang="fr">MORUE OU CABILLAUD</rdfs:label>
   <rdfs:subClassOf rdf:resource="# poisson_gadiforme"/>
</owl:Class>
```

- Les valeurs de chaque caractéristique sont organisées dans une hiérarchie dont la racine est une sous-classe de la classe générique *Valeur*. Par exemple, pour la hiérarchie de la caractéristique *Origine de l'ingrédient principal* (cf. Figure 4) le code généré en OWL est le suivant :

```
<owl:Class rdf:about="#Val_origine_ingredient_principal">
  <rdfs:label xml:lang="fr">Origine de l'ingrédient
principal</rdfs:label>
   <rdfs:subClassOf rdf:resource="#Valeur"/>
</owl:Class>
```

- La relation exprimée par la table produit_REGAL entre un nom de produit et un nom de famille est représentée en OWL par une propriété 'Appartient_a_famille' dont le domaine est la classe générique *Produit* et le co-domaine est la classe générique *Famille*.

```
<owl:ObjectProperty rdf:about="#APPARTIENT_A_FAMILLE">
<rdfs:domain rdf:resource="#Produit"/>
<rdfs:range rdf:resource="#Famille"/>
</owl:ObjectProperty>
```

- Chaque tuple (nomProduit, nomFamille) de la table produit_REGAL est représentée en OWL par une restriction sur la propriété APPARTIENT_A_FAMILLE, cette restriction étant ajoutée à la définition de la classe 'nomProduit'. Par exemple, le tuple (*Cabillaud cru*, *Poissons et batraciens*) est décrit dans la définition de la classe associée au produit *Cabillaud cru* par le code suivant en OWL :

```
<owl:Class rdf:ID="Cabillaud_cru">
… (cf. code de l'étape 3)
```

```
<rdfs:subClassOf>  <owl:Restriction>
 <owl:allValuesFrom rdf:resource="#Poissons_et_batraciens"/>
   <owl:onProperty>
     <owl:ObjectProperty rdf:about="#APPARTIENT_A_FAMILLE"/>
   </owl:onProperty>
</owl:Restriction> </rdfs:subClassOf>
</owl:Class>
```

- Pour chaque relation distincte de la table description entre un nom de produit 'nomProduit' et une caractéristique 'nomCaracteristique', une propriété appelée 'nomCaracteristique' est créée en OWL. Cette propriété a pour domaine la classe générique *Produit* et pour co-domaine la classe associée à la racine de la hiérarchie de valeurs de la caractéristique 'nomCaracteristique' (définie dans l'étape 5).

```
<owl:ObjectProperty
rdf:about="#ORIGINE_INGREDIENT_PRINCIPAL">
   <rdfs:domain rdf:resource="#Produit"/>
   <rdfs:range rdf:resource="# Val_origine_ingredient_principal "/>
</owl:ObjectProperty>
```

- Chaque tuple (nomProduit, nomCaracteristique, valeurCaracteristique) de la table description est représenté en OWL par une restriction sur la propriété 'nomCaracteristique', cette restriction étant ajoutée à la définition de la classe 'nomProduit'. Par exemple, le tuple (*Cabillaud cru*, *Origine de l'ingrédient principal*, *Morue ou cabillaud*) de la Figure 3 est décrit dans la définition de la classe associée au produit *Cabillaud cru* par le code suivant en OWL :

```
<owl:Class rdf:ID="Cabillaud_cru">
… (cf. code des étapes 3 et 7)
<rdfs:subClassOf>  <owl:Restriction>
   <owl:onProperty>
      <owl:ObjectProperty
rdf:about="#ORIGINE_INGREDIENT_PRINCIPAL "/>
   </owl:onProperty>
   <owl:allValuesFrom>
     <owl:Class rdf:ID="morue_ou_cabillaud">
   </owl:allValuesFrom>
</owl:Restriction> </rdfs:subClassOf>
</owl:Class>
```

L'ontologie CONTA générée en OWL compte 7 495 classes (dont 43 pour les familles, 2 595 pour les produits et 4 854 pour les valeurs) et 25 112 propriétés.

6. Construction de l'ontologie CONSO en OWL

Dans la source de données CONSO, qui est stockée sous la forme d'une base de données relationnelle, l'information concernant les produits alimentaires consommés se trouve dans les deux tables suivantes :

produit_SECODIP(nomProduit, nomFamille)

description(nomProduit, nomCaracteristique, valeurCaracteristique)

Les étapes de construction de l'ontologie CONSO sont à peu près les mêmes que pour l'ontologie CONTA. On peut identifier deux grandes différences : i) dans l'étape 4, la génération de la taxonomie de valeurs ne sera pas possible car les valeurs de l'ontologie CONSO ne sont pas organisées en hiérarchies ; ii) dans l'étape 9, la restriction des propriétés ne peut pas se faire avec l'opérateur **allValuesFrom,** mais avec l'opérateur **someValueFrom**, car, pour un produit donné, l'ensemble des tuples (nomProduit, nomCaracteristique, valeurCaracteristique) de la table description de l'ontologie CONSO ne donne pas une description d'un produit alimentaire, mais toutes les possibilités pour le décrire. Par exemple, pour le produit *Poisson frais* de la Figure 3 le code OWL généré est le suivant :

```
<owl:Class rdf:ID="POISSON_FRAIS">
<rdfs:label xml:lang="fr">POISSON FRAIS</rdfs:label>
<rdfs:subClassOf> <owl:Restriction>
        <owl:onProperty>
  <owl:ObjectProperty rdf:about="#PRESENTATION"/>
        </owl:onProperty>
        <owl:someValuesFrom rdf:resource="#entier"/>
</owl:Restriction> </rdfs:subClassOf>
<rdfs:subClassOf> <owl:Restriction>
        <owl:onProperty>
  <owl:ObjectProperty rdf:about="#QUEL_POISSON"/>
        </owl:onProperty>
        <owl:someValuesFrom rdf:resource="#cabillaud"/>
</owl:Restriction> </rdfs:subClassOf>
<rdfs:subClassOf> <owl:Restriction>
        <owl:onProperty>
          <owl:ObjectProperty rdf:about="#QUEL_POISSON"/>
        </owl:onProperty>
        <owl:someValuesFrom rdf:resource="#saumon"/>
</owl:Restriction> </rdfs:subClassOf>
</owl:Class>
```

L'ontologie CONSO générée en OWL compte 15 655 classes (dont 64 pour les familles, 472 pour les produits et 15 116 pour les valeurs) et 30 569 propriétés.

Cette représentation en OWL résout le problème de modélisation rencontré avec le modèle des graphes conceptuels dans la section précédente. L'interprétation logique de la description d'un produit dans la modélisation OWL proposée est bien de type universelle (quel que soit le cabillaud cru, il a pour description…).

7. Conclusion et perspectives

Dans le domaine du risque alimentaire, l'évaluation de l'exposition d'une population à un risque chimique requiert le croisement de sources de données de consommation et de contamination des aliments. Afin d'effectuer ce croisement, il est préalablement nécessaire d'aligner les noms de produits alimentaire utilisés dans les deux sources pour indexer les données. Nous avons présenté dans cet article un retour sur notre expérience de construction et d'alignement d'ontologies de produits alimentaires que nous aimerions utiliser pour étendre le système d'intégration de données du logiciel CARAT (Buche *et al.* 2006).

Pour trouver de manière semi-automatique les correspondances entre les noms des produits des deux sources nous avons proposé dans (Buche *et al.* 2008) une méthode d'alignement d'ontologies basée sur le modèle des graphes conceptuels (Mugnier *et al.* 1996). Cette méthode combine des techniques syntaxiques (comme la lemmatisation des noms de produits) avec des techniques structurelles utilisant la taxonomie des valeurs. Dans cet article nous avons rappelé les étapes de l'algorithme et nous avons mis en évidence et discuté des problèmes rencontrés : i) un problème de modélisation dû au choix initial du modèle des graphes conceptuels ; ii) un problème concernant les résultats expérimentaux qui ne sont pas satisfaisants.

Nous avons ensuite proposé dans cet article une nouvelle formalisation, en OWL (Dean *et al.* 2004) (Smith *et al.* 2004), des ontologies associées aux sources de données, en précisant les règles de transformation des métadonnées et données extraits de chaque base de données relationnelle en classes et propriétés. Notre méthode est automatique. Elle se différencie des propositions récentes dans le domaine de la génération automatique d'ontologies à partir d'une base relationnelle (Astrova 2007), (Lubyte *et al.* 2007). En effet, nous voulons conserver dans l'ontologie la représentation des taxonomies stockées dans les tables de la base en

utilisant la relation de spécialisation entre classes. Or ces méthodes représentent sous forme d'instances les tuples de données stockées dans les tables.

Les perspectives à court terme consistent à comparer les performances de notre méthode d'alignement avec celles de l'état de l'art grâce à la modélisation OWL de nos ontologies proposée dans cet article. Les perspectives à plus long terme sont d'étudier comment il est possible d'intégrer dans les méthodes d'alignement d'ontologies les évolutions périodiques de l'ontologie CONSO.

Bibliographie

Astrova I. (2007) : *Rules for Mapping SQL Relational Databases to OWL Ontologies. MTSR 2007*: 415-424

Buche P., Soler L., Tressou J. (2006) : *Le logiciel CARAT*. Dans : Bertail P., Feinberg M., Tressou J., Verger P., Analyse des Risques alimentaires, Lavoisier Tech&Doc, pp. 305-333

Buche P., Dibie-Barthélemy J., Ibanescu L. (2008) : *Ontology Mapping Using Fuzzy Conceptual Graphs and Rules*. ICCS'08 Supplement, pp. 17-24

Dean M., Schreiber G. (Editors) (2004) : *OWL Web Ontology Language Reference*, W3C Recommendation, 10 February 2004

Euzenat, J., Shvaiko, P. (2007) : *Ontology Matching*. Berlin: Springer

Gruber T.R. (1993) : *A Translation Approach to Portable Ontology Specifications*. Knowledge Acquisition, 5(2):199-220

Ireland, J. D., Moller, A. (2000) : *Review of International Food Classification and Description. Journal of Food Composition and Analysis*, 33, pp. 529-538

Kalfoglou, Y., Schorlemmer M. (2005) : *Ontology Mapping: The State of the Art*. Semantic Interoperability and Integration

Lubyte L., Tessaris S. (2007) : *Extracting Ontologies from Relational Databases*. Description Logics 2007

Mugnier, M., Chein, M. (1996) : *Représenter des connaissances et raisonner avec des graphes*. Revue d'Intelligence Artificielle 10 (1), 7-56

Noy, N. F. (2004) : *Semantic Integration: A Survey of Ontology-Based Approaches*. ACM SIGMOD Record, 33(4), pp. 65-70

Smith M. K., Welty C., McGuinness D.L . (Editors) (2004) : *OWL Web Ontology Language Guide*, W3C Recommendation, 10 February 2004

Thomopoulos T., Buche P., Haemmerlé O. (2003) : *Different Kinds of Comparisons between Fuzzy Conceptual Graphs*. ICCS 2003: 54-68

Wache H., Vögele T., Visser U., Stuckenschmidt H., Schuster G., Neumann H., Hübner S. (2001) : *Ontology-Based Integration of Information - A Survey of Existing Approaches*, pp. 108-117

Ziegler P., Dittrich K. R. (2004) : *Three Decades of Data Integration - All Problems Solved*, WCC 2004, 12, pp. 3-12

A propos des auteurs

Liliana Ibanescu
Mét@risk – INRA, UFR Informatique, AgroParisTech
16 rue Claude Bernard, F-75231 Paris Cedex 05
Liliana.Ibanescu@agroparistech.fr

Patrice Buche
Mét@risk – INRA
16 rue Claude Bernard, F-75231 Paris Cedex 05
Patrice.Buche@paris.inra.fr

Juliette Dibie-Barthélemy
Mét@risk – INRA, UFR Informatique, AgroParisTech
16 rue Claude Bernard, F-75231 Paris Cedex 05
Juliette.Dibie@agroparistech.fr

Multilinguïsation d'ontologies par des correspondances avec un lexique pivot

David Rouquet - Hong-Thai Nguyen

Résumé : Les ontologies sont parmi les représentations formelles de connaissance les plus utilisées en informatique. Le besoin d'un accès multilingue à ces connaissances apparait tant au niveau des applications (Web Sémantique, RI, TA, etc.) que de la création et de l'enrichissement des ontologies. Après avoir précisément délimité le problème de la multilinguïsation d'ontologies, nous définissons formellement les objets de notre étude. Une rapide revue de l'état de l'art montre que les approches existantes pour répondre au problème posé ne sont pas satisfaisantes. Nous proposons une méthode basée sur une correspondance entre l'ontologie et un langage pivot, disposant d'un espace lexical autonome, et pouvant représenter les énoncés de la langue de façon formelle. Une application de cette méthode dans le projet OMNIA pour l'extraction d'information de textes multilingues, en vue d'une indexation et d'une recherche d'images est présentée. Nous exposons enfin la mise en œuvre de notre méthode, supportée par la plate-forme de gestion de ressources lexicales PIVAX et utilisant le langage pivot UNL.

Mots-clés : ontologie, multilinguisme, langage pivot

1. Introduction

Les ontologies formelles font partie des représentations informatiques de connaissances les plus utilisées. Les domaines d'application sont de plus en plus nombreux et mettent en avant des difficultés bien spécifiques. On peut se demander, avec l'avènement du Web Sémantique, comment les utilisateurs pourront contribuer à la construction d'ontologies et les utiliser sans pré-requis en logique formelle ? Comment utiliser automatiquement les ontologies dans des applications (RI, TALN, etc.) qui traitent des textes "tout venant" ? Comment les conceptualisations décrites avec des terminologies monolingues peuvent-elles être accessibles dans d'autres langues (CLIR, TA, etc.) ?

On voit que l'accès (contribution et utilisation) aux ontologies par le biais de la langue naturelle est un problème clef, tant pour les humains que pour les logiciels traitant des textes. Le bénéfice d'un accès multilingue est alors évident dans le contexte d'utilisation distribuée des ontologies. Les méthodes visant à la "multiliguïsation d'ontologies" doivent selon nous répondre à certains critères que nous précisons dans cet article. Elles doivent être *modulaires* et *dynamiques,* sans contraindre l'ontologie ni interférer avec la conceptualisation.

On commencera dans cet article par définir précisément le problème de la multilinguïsation d'ontologies. Nous proposerons ensuite une définition formelle des ontologies informatiques adaptée au problème traité et suffisamment générale pour englober toutes les ontologies rencontrées. Une brève revue de l'état de l'art montrera qu'aucune des approches que nous avons trouvé ne répond au problème dans son intégralité. Nous proposerons donc une méthode de multilinguïsation d'ontologies par des correspondances avec un langage pivot. Cette méthode est illustrée par son utilisation dans le projet ANR OMNIA (recherche et indexation d'images accompagnées de textes), pour une extraction d'information dans des textes multilingues. Nous décrirons enfin la mise en œuvre de notre méthode, avec le langage pivot UNL et des ressources complémentaires comme WordNet, au sein de la plate-forme de gestion de ressources lexicales PIVAX.

2. Définition du problème

Selon (Gruber 1993), les ontologies "informatiques" sont des spécifications explicites et formelles d'une conceptualisation d'un domaine de connaissances.

Une ontologie comporte :

1. un treillis conceptuel, enrichi par des relations et des propriétés obéissant a des axiomes logiques (*T-box*) ;

2. un ensemble d'objets peuplant les concepts et décrits par les relations et les propriétés (*A-box*).

Chaque concept, objet et relation de l'ontologie est désigné par une *étiquette* constituée à partir de lexèmes ou locutions une langue naturelle. Les étiquettes ne sont pas nécessairement bien formées dans une langue (abréviations, mots "agglutinés", etc.) mais sont suffisamment explicites pour permettre une interprétation en langue naturelle. Idéalement, les ambiguïtés dans l'interprétation de ces étiquettes sont levées par le contexte au sein de l'ontologie. La connaissance contenue dans une ontologie est accessible tant par des agents logiciels (grâce à une sémantique formelle) que par des humains (grâce aux étiquettes et à leur interprétation pragmatique). Par *accès* à une ontologie, nous entendons : contribuer à la connaissance ontologique et utiliser cette connaissance.

L'ajout de connaissances dans une ontologie peut, en outre, se faire manuellement via un éditeur spécialisé (par exemple Protégé[1]) ou de façon automatisée grâce à des processus d'extraction d'information. La contribution manuelle est effectuée par un humain qui retranscrit naturellement une conceptualisation via sa langue maternelle ; l'extraction d'information est souvent faite à partir de textes. Aussi, dans ces deux cas, le traitement d'une langue naturelle est nécessaire. D'autre part, la majorité des utilisateurs humains ne sont pas en mesure de formuler des requêtes formelles SPARQL[2] pour consulter une ontologie et certaines applications (RI, TA, etc.) doivent utiliser automatiquement une ontologie pour traiter des textes. Ici encore, l'ontologie doit être accessible dans la langue de l'utilisateur ou du document.

Il est donc intéressant de proposer des méthodes permettant l'ajout de données multilingues à une ontologie donnée, pour en favoriser l'accès (par des agents humains ou logiciels). Selon nous, ces méthodes doivent répondre à certains critères. Premièrement, elles ne doivent pas contraindre la création de l'ontologie pour ne pas décourager les contributions humaines ou complexifier les contributions automatiques. Ensuite, la conceptualisation spécifiée dans l'ontologie sert un but bien précis que nous ne connaissons pas nécessairement, les méthodes de

1 http://protege.stanford.edu/
2 www.w3.org/TR/rdf-sparql-query/

multilinguïsation ne doivent donc pas interférer avec cette conceptualisation. Enfin, les méthodes doivent idéalement être *modulaires* et *dynamiques*. C'est à dire qu'elles doivent respectivement : permettre l'ajout de nouvelles langues sans recours aux spécialistes qui ont fourni la conceptualisation et s'adapter automatiquement à des modifications incrémentales de l'ontologie sans repartir de zéro.

Nous cherchons donc à résoudre le problème suivant :

Étant donnée une ontologie informatique quelconque, permettre l'ajout de données multilingues à cette ontologie, sans interférer avec la conceptualisation, de façon modulaire et dynamique.

3. Définition formelle des ontologies

3.1. Objectifs

Bien que certains aspects, comme la nécessité d'un caractère consensuel, soient encore sujets à débat, la définition intuitive de (Grubber 1993) est aujourd'hui communément admise pour les ontologies informatiques. Cependant, aucune définition formelle des ontologies ne s'est imposée et l'on en trouve de multiples concurrentes (par exemple (Maedche *et al.* 2003) et (Euzenat *et al.* 2007)). Il parait en fait raisonnable d'en choisir une selon ses besoins spécifiques.

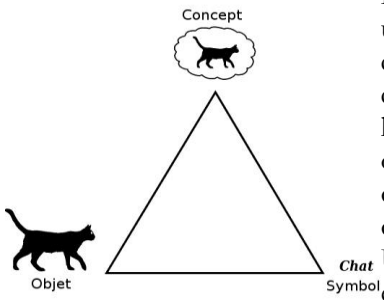


Figure 1. Le triangle sémiotique

Dans notre cas, il s'agit de développer une méthode pour permettre l'ajout de données multilingues. La définition doit donc identifier clairement les éléments linguistiquement pertinents d'une ontologie. D'autre part, une ontologie est constituée de trois éléments principaux : concept, objet, symbole (ou étiquette). Une bonne définition formelle doit donc également faire une distinction claire entre ces éléments, comme présenté dans le triangle sémiotique de la figure 1 (Ogden *et al.* 1923).

La définition doit enfin être suffisamment générale pour englober toutes les ontologies informatiques, indépendamment du point de vue adopté pour leur utilisation ou leur développement.

3.2. Définitions formelles

Nous adoptons comme définition formelle pour les ontologies informatiques celle de (Maedche *et al.* 2003). Nous l'avons complétée pour répondre au mieux aux critères présentés précédemment. Elle est composée des notions d'*ontologie abstraite*, d'i*nstanciation* et de *lexique*.

Définition 1 : Une *ontologie abstraite* est une structure $\mathcal{O} = (C, \top, R, \sigma, \leq_C, \leq_R, \mathcal{L}, \mathcal{T})$ avec :

- C et R des ensembles finis disjoints d'étiquettes de concepts et d'étiquettes de relations ;
- $\sigma : R \rightarrow C \times C$ une fonction signature, retournant le domaine d'une relation ;
- $\leq_C$ et $\leq_R$ des ordres partiels sur C et R ;
- $\top \in C$ une borne supérieure pour $\leq_C$ de sorte que cet ordre forme un semi-treillis sur C, nommé treillis conceptuel ;
- $\mathcal{L}$ une théorie logique, dotée d'une sémantique formelle, dont la signature contient les constantes de C et R, les ordres $\leq_C$ et $\leq_R$ ainsi que σ ;
- $\mathcal{T}$ un ensemble d'axiomes exprimés dans la logique $\mathcal{L}$. On l'appelle aussi la T-box (terminological box) ;

Exemple : voici une ontologie très simple :

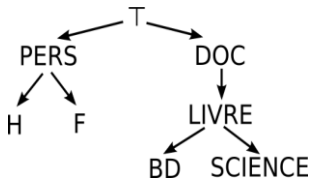


Figure 2. Treillis des concepts

Les étiquettes des concepts s'interprètent comme : personne, homme, femme, document, livre, bande dessinée, ouvrage scientifique. La relation AUT exprime qu'une personne est l'auteur d'un document.

$R = \{AUT\}$

$C = \{PERS, H, F, DOC, LIVRE, BD, SCIENCE\}$

$\sigma(AUT) = PERS \times DOC$

La T-box, exprimée avec la théorie des ensembles, est :

1. $H \cap F = \emptyset$
2. $PERS = H \cup F$
3. $\forall d \in DOC\{a \in PERS, AUT(a, d)\} \neq \emptyset$

Les deux premiers axiomes expriment qu'une personne est soit un homme soit une femme. Le troisième exprime qu'un document a au moins un auteur.

Définition 2 : une *instanciation* pour une ontologie abstraite $\mathcal{O} = (C, \top, R, \sigma, \leq_C, \leq_R, \mathcal{L}, \mathcal{T})$ est une structure $Inst = (E, \mathcal{A})$ avec:

- E un ensemble fini d'individus ;
- $\mathcal{A}$ un ensemble d'axiomes exprimés dans la logique $\mathcal{L}$. On l'appelle aussi la A-box (assertional box) ;

L'ontologie *instanciée* pourra être notée :

$$\mathcal{O} = (C, \top, R, E, \sigma, \leq_C, \leq_R, \mathcal{L}, \mathcal{T}, \mathcal{A})$$

Exemple : une instanciation de l'ontologie prise en exemple peut être donnée par la A-box suivante :

- $NicolasBourbaki \in PERS$
- $FrankMiller \in PERS$
- $ElementsdeMathematiques \in SCIENCE$
- $SinCity \in BD$
- $AUT\,(ElementsdeMathematiques, NicolasBourbaki)$
- $AUT\,(SinCity, FranckMiller)$

Définition 3 : un *lexique* pour une ontologie $\mathcal{O} = (C, \top, R, E, \sigma, \leq_C, \leq_R, \mathcal{L}, \mathcal{T}, \mathcal{A})$ est une structure $Lex = (D, Ref)$ avec :

- D un ensemble de lexies (mots ou locutions) ;
- $Ref \subseteq ((C \cup R \cup E) \times D)$ une relation appelée affectation lexicale.

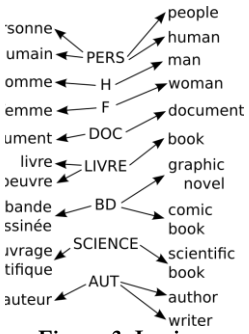


Figure 3. Lexiques

Exemple : la figure 3 présente deux lexiques (français et anglais) pour l'ontologie prise en exemple.

Maintenant que nous avons délimité le problème et défini formellement les objets de notre étude, nous présentons les approches existantes pour l'accès multilingue à une ontologie et montrons en quoi elles ne semblent pas satisfaisantes.

4. Approches existantes du problème

4.1. Traduction vers des langues cibles

(Espinoza *et al.* 2008) propose une méthode pour traduire les étiquettes de l'ontologie, directement vers d'autres langues. Un service adapté propose d'abord des traductions possibles en consultant des ressources linguistiques (dictionnaires bilingues, bases lexicales, etc.) ; la liste des traductions est ensuite classée selon leur qualité probable en utilisant les voisinages dans le treillis conceptuel.

Les méthodes de désambiguïsation utilisant le treillis conceptuel sont intéressantes mais doivent être appliquées de nouveau pour chaque langue cible. Le travail de désambiguïsation n'est pas factorisé entre les différentes langues et le simple apport de ressources lexicales ne suffit pas à augmenter le nombre de langues couvertes.

4.2. Greffe de "sous-ontologies" linguistiques

(Buitelaar *et al.* 2006) propose une trame de "sous-ontologie" linguistique permettant de stocker la traduction d'un concept accompagnée de données morpho-syntactiques. Il faut, pour chaque concept de l'ontologie source, instancier la trame dans la langue cible et la greffer au concept comme illustré dans les figures 4 et 5.

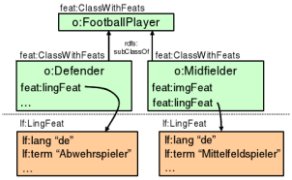


Figure 4. Greffe de la trame linguistique

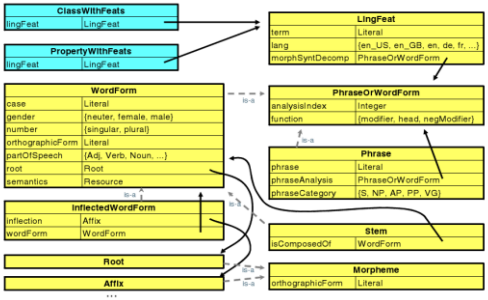


Figure 5. Détail de la trame linguistique

Cette approche rend cependant l'ontologie source bien plus complexe. De plus, l'instanciation et la greffe des sous-ontologies linguistiques doivent être faites pour chaque langue cible sans qu'aucune méthode automatisée n'ait été proposée.

4.3. Relier l'ontologie à des WordNets[3]

(Niels *et al.* 2003) décrit la correspondance entre SUMO (Suggested Upper Merged Ontology) et WordNet, calculée pour rendre l'ontologie accessible à des humains et utilisable automatiquement par des applications traitant des textes. Cette correspondance comprend des relations de synonymie, d'hyperonymie et d'instanciation entre les concepts de l'ontologie et les *synsets* de WordNet.

Cette méthode ne permet pas la multilinguïsation . Elle est cependant intéressante puisqu'elle permet l'ajout de données monolingues à une ontologie, ce qui est un sous-problème de celui qui nous concerne. Dans cette méthode, les ressources ontologiques et linguistiques ne sont pas clairement séparées (des relations d'hyperonymie et d'instanciation sont déjà présentes dans l'ontologie), ce qui risque d'interférer avec la conceptualisation proposée dans l'ontologie. Il se trouve d'ailleurs qu'un des objectifs de cette correspondance avec WordNet est de réviser et de compléter l'ontologie, ce qui n'est pas souhaitable pour nous. D'autre part, en vue d'une multilinguïsation, le travail décrit est à effectuer pour chaque nouvelle langue, ce qui est loin de se résumer à l'ajout de ressources linguistiques.

Le système KYOTO (Vossen *et al.* 2008) propose un environnement Wiki pour le développement collaboratif d'une ontologie interlingue et de sa correspondance avec des WordNets (actuellement sept langues

3 http://wordnet.princeton.edu/

supportées : basque, chinois, allemand, anglais, italien, japonais et espagnol).

Ici encore, l'approche ne permet pas de traiter séparément les aspects de conceptualisation et de multilinguïsation (et c'est même un de ses buts) puisque des experts de chaque langue doivent proposer, à partir des entrées des WordNets, des liens vers les concepts existants ou vers de nouveaux concepts consensuels à insérer dans l'ontologie.

5. Approche avec un langage pivot

5.1. Principe

En vue de permettre l'accès multilingue à une ontologie, développée *a priori*, nous proposons de passer par une représentation pivot de la langue. Notre objectif est de construire un lexique non ambigu pour l'ontologie, dans un langage pivot approprié, et d'utiliser ce lexique interlingue comme portail vers les langues naturelles. Pour permettre cela, le langage pivot doit disposer d'un espace lexical autonome et non ambigu qui est mis en correspondance avec les étiquettes de l'ontologie par des affectations lexicales. Il doit également permettre la construction de syntagmes pour traiter les concepts portant des étiquettes "composées".

Cette méthode présente plusieurs avantages. Premièrement, l'inévitable travail de désambiguïsation pour relier les concepts à un lexique est "factorisé". Il est nécessaire pour le calcul du lexique pivot (et ses mises à jour en cas de modification de l'ontologie), mais pas pour l'ajout de nouvelles langues. Une fois le lexique pivot calculé, l'ajout de nouvelles langues peut se faire par la simple acquisition de dictionnaires reliant la langue cible au langage pivot. En outre, la construction de ces ressources ne requiert pas un expert du domaine de l'ontologie compétent dans la langue cible, c'est une tâche linguistique (la méthode est bien modulaire). D'autre part, la méthode proposée ne contraint en aucune manière les processus de création ou de contribution pour l'ontologie puisque les correspondances sont calculées ou mises à jour *a posteriori*. Enfin, la méthode est respectueuse de la conceptualisation proposée dans l'ontologie car les affectations lexicales sont clairement distinguées des relations initialement présentes dans l'ontologie.

La mise en œuvre concrète du stockage et de la gestion des correspondances est décrite dans la partie 6. Le paragraphe suivant

explique la méthode adoptée dans le projet ANR OMNIA[4] pour permettre l'accès multilingue à une ontologie en utilisant cette approche par correspondance avec un langage pivot.

5.2. Exemple d'accès à une ontologie

Un des buts du projet OMNIA est de développer un outil de recherche pour des entrepôts d'images en ligne. A cet effet, une ontologie de catégorisation des images est construite. Les images seront décrites dans la A-box grâce à la fusion des données issues de l'analyse visuelle et de l'analyse des légendes et des textes compagnons écrits en langue naturelle "tout venant". Plusieurs langues seront par ailleurs proposées à l'utilisateur pour formuler ses requêtes librement (mots clefs, phrases, etc.).

L'accès multilingue à l'ontologie dans le projet concerne donc l'indexation des images dans la A-box à l'aide des légendes textuelles, et le traitement des requêtes de l'utilisateur. La même méthode est employée pour traiter ces deux aspects, comme illustré dans la figure 6.

Il s'agit d'annoter les textes avec les lexies interlingues non ambigües (lexèmes ou locutions) du langage pivot sans procéder à une analyse syntaxique poussée. On sait avec (Daoud 2006) que c'est une approche viable pour initier une extraction de contenu. L'annotation est réalisée de façon automatique grâce à un dictionnaire "langue naturelle" – "langage pivot" et à des procédés de désambiguïsation. On réalise ensuite l'extraction de contenu. Ce processus prend en entrée les annotations interlingues et la correspondance "langage pivot" – "ontologie". Il retourne les informations pertinentes (i.e. qui peuvent être représentées dans l'ontologie) formatées dans le langage de description ou de requête de l'ontologie. Les informations peuvent alors, selon le cas, être stockées dans la A-box ou soumises à un raisonneur pour résoudre la requête.

4 www.omnia-project.org

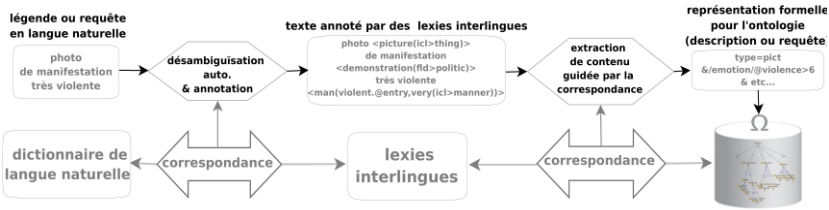


Figure 6. Accès à l'ontologie à partir de textes

6. Mise en œuvre

6.1. La plate-forme PIVAX

PIVAX (Nguyen *et al.* 2007) est une plateforme en ligne pour la gestion des ressources lexicales de systèmes de TA utilisant un pivot lexical. Elle a été développée à partir de la plateforme générique Jibiki (Sérasset 2005) qui différencie l'organisation des volumes de la base lexicale (*macrostucture*) et l'organisation des éléments de chaque volume (*microstucture*). Les ressources qui respectent une syntaxe XML peuvent être importées directement par la simple adjonction d'un ficher de métadonnées « Xpath » décrivant leur microstructure. Un exemple de fichier de métadonnées est présenté dans la figure 8 au paragraphe 6.2.

La macrostucture de PIVAX est composée de trois couches. Pour chaque langue supportée, on trouve :

- un ou plusieurs volumes de *lexies*. Les *lexies* correspondent à des sens de mots dans un dictionnaire.

- un unique volume d'*axèmes* (acceptions monolingues). Un *axème* relie des lexies synonymes dans un même langage.

- Un volume partagé d'*axies* (acceptions interlingues). Une *axie* relie des axèmes synonymes.

Au niveau de la microstructure, les lexies contiennent un lemme et des informations complémentaires (classe, définition, statut, etc.). Les axèmes et les axies sont des liens, représentés simplement comme des ensembles de lexies et d'axèmes respectivement. Nous exploiterons également la possibilité de représenter des relations entre axèmes ou entre axies.

L'organisation des notions précédentes au sein de PIVAX est illustrée dans la figure 2. Le "langage pivot" n'occupe pas une place centrale mais est représenté comme les autres langues. Les paragraphes suivant décrivent les ressources proposées pour la mise en œuvre de l'accès multilingue à une ontologie ainsi que leur intégration et leur utilisation dans la plate-forme PIVAX .

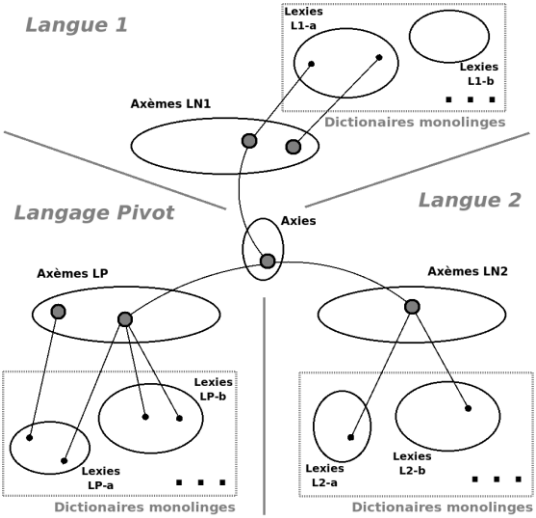


Figure 7. La plate-forme PIVAX

6.2. L'ontologie

L'ontologie que l'on cherche à multilingualiser est représentée dans PIVAX comme un dictionnaire. Les concepts et les instances forment un volume d'axèmes dans lequel les relations de l'ontologie sont également représentées (spécialisation, instanciation, etc.). Les étiquettes des concepts, instances et relations sont stockées dans un volume de lexies.

Les langages de description d'ontologies basés sur XML permettent une importation immédiate dans PIVAX. La figure 8 présente le fichier de métadonnées pour importer dans un volume de lexies les (étiquettes des) concepts d'une ontologie décrite en OWL. D'autres métadonnées doivent être ajoutées pour importer les instances et les relations dans les volumes de lexies et d'axèmes.

```
<volume-metadata>
<comments>PIVAX OWL - OMNIA Ontology for OMNIA project</comments>
  <cdm-elements>
    <cdm-volume    xpath="/rdf_RDF"/>
    <cdm-entry    xpath="/rdf_RDF/owl_Class"/>
    <cdm-entry-id    xpath="/rdf_RDF/owl_Class/@rdf_ID"/>
    <cdm-headword    d:lang="owl"
xpath="/rdf_RDF/owl_Class/text()"/>
  </cdm-elements>
  <administrators>
    <user-ref    name="nguyenht"/>
  </administrators>
  <system    name="omnia">
    <copyright>Copyright by OMNIA project</copyright>
    <description>OWL Ontology dictionary</description>
    <organization>GETALP-LIG</organization>
    <adress/>
    <responsability>Christian.Boitet@imag.fr</responsability>
    <url_link>http://www.ellemme.org/</url_link>
  </system>
  <xmlschema-ref    xlink:href="pivax_local.xsd"/>
  <volume-ref    xlink:href="Pivax_owl-omnia.xml"/>
  <template-entry-ref    xlink:href="Pivax_owl-omnia-
template.xml"/>
</volume-metadata>
```

Figure 8. Métadonnées pour PIVAX

6.3. Le langage pivot UNL[5]

Le terme UNL (Universal Networking Language) recouvre en particulier deux choses différentes :

- le projet international UNL, lancé en novembre 1996 par l'UNU (Université des Nations Unies) à Tokyo ;
- le langage UNL, qui est un langage pivot "anglo-sémantique", et pas une langue humaine naturelle ou construite (comme l'espéranto) ;

Les expressions du langage UNL représentent le sens d'un énoncé par une structure sémantique abstraite (un graphe) d'un énoncé anglais équivalent.

Le vocabulaire de UNL est constitué de lexies interlingues appelées UW (Universal Words, en français, Unités de Vocabulaire Virtuel). Idéalement, elles déterminent de manière non ambiguë un concept existant dans l'ensemble des langues considérées (certains concepts, comme atterrir ou amerrir n'existent que dans certaines langues, les autres ne référant qu'a des concepts moins fins. Une UW est composée de :

1. un *mot-vedette,* si possible dérivé de l'anglais, qui peut être un mot, une expression ou encore une phrase entière.

5 http://www.undl.org/

2. une *liste de restrictions* servant à délimiter le concept précis porté par l'UW.

Exemples :

- book(icl>thing) et book(icl>do, agt>human, obj>thing) pour lesquels le sens de l'UW est précisé par des restrictions ;
- ikebana(icl>flower_arrangement) dont le mot-vedette a été importé du japonais ;
- go_down dont le mot-vedette est une expression et dont le sens n'a pas besoin d'être précisé par des restrictions.

Les UW sont organisées dans un réseau sémantique nommé UNLKB6 (UNL Knowledge Base). Ce réseau contient des relations sémantiques et syntaxiques pondérées décrivant le comportement des UW les unes par rapport aux autres. Il facilite l'interprétation des expressions UNL.

Plusieurs dictionnaires d'UW sont disponibles et/ou en cours de développement (projet initial de l'UNU, consortium U++[7], etc.). Chacun de ces dictionnaires constitue un volume de lexies dans PIVAX. Ils devront être reliés entre eux par des axèmes. Certaines UW sont en effet équivalentes mais n'ont par exemple pas le même mot-vedette d'un volume à l'autre, comme book(agt>human,obj>thing) et reserve(agt>human,obj>thing). Le volume d'axèmes contient également les relations de l'UNLKB. Les dictionnaires "UNL" – "langue naturelle" disponibles sont représentés par des axies entre les axèmes "UNL" et les axèmes "langue naturelle".

6.4. WordNet

WordNet est une base lexicale de l'anglais développée à l'Université de Princeton (Fellbaum 1998), toujours en développement et portée dans d'autres langues (voir par exemple le projet EuroWordNet[8]). Bien que cette base lexicale ne soit pas fondamentale pour notre méthode de multilinguïsation, il est intéressant d'en disposer parmi nos ressources au sein de PIVAX. WordNet est en effet largement utilisé pour de nombreuses applications de traitement de la langue (désambiguïsation, recherche d'information, etc.) que l'on pourra alors exploiter.

Les éléments de WordNet sont des *synsets*. Ce sont des ensembles de termes qui présentent une interprétation commune dans au moins un

6 http://www.undl.org/unlsys/unl/unl2005/UNLKB.htm
7 http://www.unl.fi.upm.es/consorcio/
8 http://www.illc.uva.nl/EuroWordNet/

contexte d'utilisation. Idéalement, ces synsets représentent des concepts. Ils sont par ailleurs reliés par des relations sémantiques et lexicales.

Sous PIVAX, les lexies des volumes "WordNet" sont constituées d'un couple (mot, synset) comparable à une entrée dans un dictionnaire (mot, sens de mot). Le volume d'axèmes est constitué de l'ensemble des synsets et des relations entre ces synsets. Pour faciliter l'importation dans PIVAX, nous avons utilisé la conversion de WordNet sous forme RDFS/OWL[9] dont la syntaxe est une spécification XML. D'autre part, les UW de certains volumes "UNL" sont construites à partir de synsets de WordNet (par exemple les "UW++" du consortium U++). Ces volumes "UNL" et "WordNet" sont alors mis en correspondance par des axies.

6.5. Organisation générale

La mise en œuvre de notre méthode dans PIVAX revient donc à calculer des axies entre les volumes "ontologie" et "UNL". Pour une ontologie dont les étiquettes sont formées à partir d'une langue naturelle donnée "L1", cela est fait grâce à la combinaison des axies "ontologie" – "L1" et des axies "L1" – "UNL", ainsi qu'à des procédés de désambiguïsation. L'accès à l'ontologie par une autre langue est alors réalisé grâce aux axies entre UNL et cette autre langue.

La figure 8 décrit l'instance spécifique de la plateforme PIVAX qui supporte les correspondances entre une ontologie donnée, UNL, les langues naturelles et des ressources complémentaires (pour l'instant WordNet uniquement).

9 http://www.w3.org/TR/wordnet-rdf/

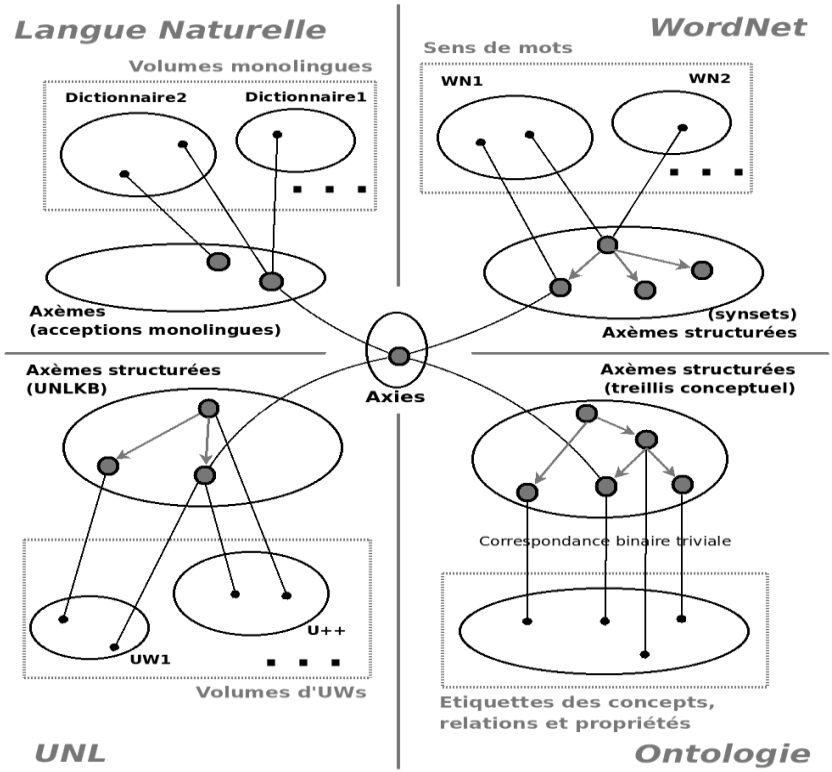


Figure 9. PIVAX pour la multilinguïsation d'ontologies

7. Conclusion et perspectives

Nous avons montré dans cet article que l'ajout de ressources multilingues aux ontologies informatiques constitue un problème en soi. Disposer de telles ressources est pertinent pour de multiples applications utilisant les ontologies (recherche d'information, traduction automatique, etc.), ainsi que pour les processus de création et d'enrichissement des ontologies. Des caractéristiques requises par les méthodes visant à résoudre ce problème ont été mises en avant (caractères modulaire et dynamique, respect de la conceptualisation formalisée, liberté dans la création et l'enrichissement des ontologies). Une rapide revue de l'état de l'art a montré en quoi les méthodes existantes ne répondaient que partiellement au problème posé, et nous avons proposé une approche plus satisfaisante par correspondance entre l'ontologie et un langage pivot dont le vocabulaire est un ensemble d'acceptions interlingues (UW++). Cette

méthode est appliquée dans le projet OMNIA pour l'extraction d'information à partir de textes multilingues "compagnons" d'images, en vue de l'indexation et de la recherche d'images. La mise en œuvre de notre méthode est supportée par la plate-forme de gestion de ressources lexicales PIVAX.

Si le stockage et la gestion des correspondances sont concrètement résolus par l'utilisation de PIVAX, le calcul et la mise à jour de ces correspondances reste une tâche à explorer en détail. En particulier, l'étude de la nature des correspondances et de leurs propriétés pour assurer un caractère dynamique à la méthode fait partie de l'objet d'une thèse en cours. Des procédés de désambiguïsation automatique sont également étudiés en utilisant des vecteurs conceptuels (Schwab 2005). Cette dernière ressource pourrait également être exploitée avec PIVAX.

Remerciements

Les auteurs tiennent à remercier leurs directeurs et/ou collègues Christian Boitet, Valérie Bellynck et Didier Schwab pour les relectures et les précieux conseils.

Notre gratitude va également aux partenaires du projet ANR OMNIA (ANR-07-MDCO-009-02) qui nous offrent le cadre applicatif et les ressources financières pour mener ces travaux.

Bibliographie

Buitelaar P., Sintek M. et Kiesel M. (2006) : A Multilingual/Multimedia Lexicon Model for Ontologies, *The Semantic Web: Research and Applications, Lecture Notes in Computer Science, vol. 4011, pp. 502-513, ISBN: 978-3540-34544-2, Springer*

Daoud D. (2006) : Il faut et on peut construire des systèmes de commerce électronique à interface en langue naturelle restreints (et multilingues) en utilisant des méthodes orientées vers les sous-langages et le contenu, *Thèse de Doctorat, 290 p., Université Joseph Fourier, Grenoble*

Espinoza M., Gómez-Pérez A. et Mena E. (2008) : Enriching an Ontology with Multilingual Information, *The Semantic Web: Research and Applications, Lecture Notes in Computer Science, vol. 4011, pp. 333-347, ISBN: 978-3-540-34544-2, Springer*

Euzenat J. and Shvaiko P. (1998) : Ontology Matching, *ISBN: 3540496114 Springer, 2007.*

Fellbaum C. WordNet : An Electronic Lexical Database *(Language, Speech, and Communication), ISBN : 026206197X, The MIT Press*

Gruber T.R. (1993) : A Translation Approach to Portable Ontology Specifications, *Knowledge Acquisition, vol. 5-2, pp. 199-220, ISSN: 1042-8143, Academic Press Ltd.*

Nguyen H.T., Boitet C., Sérasset G. (2007) : PIVAX, an online contributive lexical data base for heterogeneous MT systems using a lexical pivot, *The Seventh Symposium on Natural Language Processing (SNLP-2007), Bangkok, Thailand*

Maedche A., Neumann G. et Staab S. (2003) : Bootstrapping an ontology-based information extraction system, *Intelligent exploration of the web, pp. 345-359, ISBN: 3-7908-1529-2, Physica-Verlag GmbH*

Niles I. et Pease A. (2003) : Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology, *Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 03), pp. 23--26, Las Vegas*

Ogden C. et Richards C. (1923) : The Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism, *Harcourt*

Schwab D. (2005) : Approche hybride –lexicale et thématique– pour la modélisation, la détection et l'exploitation des fonctions lexicales en vue de l'analyse sémantique de textes, *Thèse de Doctorat, 363 p., Université Montpellier 2*

Vossen P. et al. (2008) : KYOTO : A system for Mining, Structuring and Distributing Knowledge Across Languages and Cultures, *Conference on Language Resources and Evaluation (LREC08), Marrakech, Morocco*

A propos des auteurs

David Rouquet
GETALP – LIG
385 av. de la Bibliothèque
Domaine Universitaire, BP 53
38041 Grenoble Cedex 9
David.Rouquet@imag.fr
http://www.liglab.fr

Hong-Thai Nguyen
GETALP – LIG
385 av. de la Bibliothèque
Domaine Universitaire, BP 53
38041 Grenoble Cedex 9
Hong-Thai.Nguyen@imag.fr
http://www.liglab.fr

La reformulation : processus dynamique d'acquisition des connaissances.- Le cas du discours technique arabe d'Internet

Andrée Affeich

Résumé : En arabe, les études dédiées à la reformulation sont très rares. Nous en avons trouvé une seule. Il s'agit d'une étude faite par Fayza Elqasem sur le rôle que joue la reformulation dans l'activité traduisante. C'est ainsi que nous avons voulu défricher encore plus ce terrain en examinant de notre côté des textes authentiques, non traduits, relevant du domaine d'Internet et rédigés par des spécialistes afin de voir comment ceux qui possèdent parfaitement le savoir dans ce domaine agissent en aval pour présenter l'information à un public d'apprentis ou de novices.

Mots-clés : Reformulation, reformulé, relateur, acquisition, connaissances, discours, arabe, Internet

1. Introduction

Nous partons d'une petite histoire que nous avons reprise à Daniel Jacobi dans son article "Du discours scientifique, de sa reformulation et de quelques usages sociaux de la science" (1984 : 38). Il s'agit de deux héros de Gustave Flaubert du nom de Bouvard et Pécuchet qui ont décidé d'aller à la découverte de l'immense savoir scientifique. Un soir d'été, alors qu'ils avaient bien mangé, ils se mirent à contempler le ciel étoilé et à se poser des questions. Et "*ils étaient fiers de réfléchir sur de si grands objets*[1]", comme le dit Flaubert. Leur curiosité les pousse à chercher les réponses dans les livres. Pour connaître par exemple la chimie, "ils se procurèrent le cours de Régnault et apprirent d'abord que 'des corps simples sont peut-être composés…' " ; mais ne comprenant rien ni l'un ni l'autre, "ils recoururent à un ouvrage moins difficile, celui de Girardin (…)"[2], et ainsi de suite pour les autres disciplines.

Ces deux héros de Flaubert sont, par conséquent, le prototype de ce qu'on appelle le public de la vulgarisation ou le grand public, c'est-à-dire tout simplement tout un chacun qui n'est pas spécialisé dans un domaine quelconque et qui cherche à le comprendre et à l'appréhender.

Conscients depuis longtemps de l'importance et de la nécessité du partage du savoir, des théories et des concepts techno-scientifiques au sein d'une communauté de non-spécialistes, les spécialistes ont lancé le chantier de la vulgarisation en opérant des séries d'opérations dans le discours afin d'aider les lecteurs à accéder au sens. En simplifiant les choses, ont-ils trahi et dénaturé le message scientifique ou "la vérité scientifique", comme le prétendent certains ?[3] Certainement pas. S'ils sont appelés à vulgariser, c'est avant tout dans une optique de socio-diffusion du message scientifique, pour que les non-spécialistes cessent de voir dans la science un obstacle et cessent d'accuser les terminologies d'être des sortes de jargons peu compréhensibles.

André Martinet (1967 : 9) dit : "la fonction essentielle de cet **instrument** qu'est une langue est celle de **communication** […] si toute langue se modifie au cours du temps, c'est essentiellement pour s'adapter de la façon la plus économique à la satisfaction des besoins de communication de la communauté qui la parle". Suite à André Martinet, nous pouvons affirmer

1 Flaubert G. G. Bouvard et Pécuchet, Garnier/ Flammarion, 1966, p. 104, cité par Jacobi D., 1984, p. 38.
2 Id., ibid., p. 85, 1984, p. 39.
3 Cf. Mortureux M.-F. « Paraphrase et métalangage dans le dialogue de vulgarisation », Langue Française, 1982, no 53, p. 61, qui cite un article paru dans Télérama (no 1613, 13-19/12/80) dans lequel l'auteur S. Grégeois exprime ses réserves à l'égard de la vulgarisation.

que si tout discours technique et scientifique (DTS) est amené à se modifier au cours du temps, c'est essentiellement pour s'adapter de la façon la plus convenable à la satisfaction des besoins de communication de la communauté qui s'en sert.

Dans ce cas-là, comment diffuser les concepts et comment traiter et présenter les dénominations correspondantes au sein de ce discours changeant ? Il est vrai, comme le note Daniel Jacobi (1994 : 89) que "vulgariser est une entreprise qui se situe au cœur d'une contradiction : comme le scripteur se propose de faire connaître le sens des notions et des concepts spécialisés construits par les sciences, il est contraint d'utiliser les termes et les lexies des langues de spécialité ; mais, en employant dans son texte des termes spécialisés, il redoute – à juste titre – que les locuteurs ne puissent en comprendre le sens ; pour prévenir les difficultés d'accès au sens des destinataires, le scripteur recourt à une série de mécanismes, de type métalinguistique le plus souvent, qui lui permettent de mettre en relation les termes scientifiques avec des mots de la langue commune". Comment ces mécanismes donnent-ils accès au sens ? Et par quels moyens la transmission du savoir et l'acquisition des connaissances se font-elles au sein de ce chantier de vulgarisation ?

Parmi ces séries de mécanismes qui caractérisent les discours de vulgarisation techno-scientifique, nous présentons dans cette étude une stragégie "souple" et "labile" qu'est la **reformulation des termes arabes d'Internet** à travers un corpus de six ouvrages. En matière de typologie des discours, il est important de noter que ce phénomène est commun surtout à deux types de textes : les textes de vulgarisation et les textes didactiques. Toujours est-il que les textes très spécialisés ne sont pas exempts de termes reformulés. Le spécialiste se sent parfois obligé dans certaines situations d'expliquer, de schématiser, de reformuler certains concepts à ses collègues.

Les exemples cités dans cette étude sont tirés des textes suivants[4] :

- cours de Houssam Abed : *استثمار الإنترنت (Exploitation d'Internet)*, **ouvrage didactique ou pédagogique** ;

4 Notre corpus couvre les trois zones géographiques dans lesquelles se situent les pays arabes, à savoir la Péninsule arabique, le Moyen-Orient et l'Afrique. Les pays concernés par la présente étude sont : les Émirats arabes unis, la Syrie, l'Égypte, la Tunisie et l'Algérie. Nous donnons une traduction en français de tous les contextes arabes de reformulation. Chaque contexte est également suivi d'une référence codée comprenant les éléments suivants : COR (= corpus) + les 2 premières lettres attestant de l'origine géographique du corpus + la 1ère lettre du nom et la 1ère lettre du prénom de l'auteur + l'année de publication du livre + le numéro de la page dans laquelle figure le contexte tiré.

- ouvrage de Abdel-Sattar Ghammouri : *شبكة الاتّصال الدّولي الإنترنيت INTERNET (Le réseau de communication internationale Internet)*, **ouvrage de type mémoire / thèse** ;

- ouvrage de Mohamed Ben Abdallah Zayed : *مَدخل إلى عالم الأنترنت (Introduction à Internet)*, **ouvrage de semi-vulgarisation technique** ;

- ouvrages de Adel Abdel-Mawla : *تَعلّم الإنترنت في ثلاث ساعات (Apprenez Internet en 3 heures)*, M. Béchir : *الإنترنت للمبتدئين (Internet pour débutants)*, et Ahmed Rayyan : *خَدَمات الإنترنت (Les services offerts par Internet)*, ouvrages de vulgarisation technique.

Une remarque mérite d'être retenue à propos de ces ouvrages : nous savons déjà que trois d'entre eux sont rédigés par des spécialistes (Houssam Abed, Mohamed Ben Abdallah Zayed et Ahmed Rayyan). Pour les deux autres, aucune indication n'a été donnée sur leurs auteurs Adel Abdel-Mawla et M. Béchir. Quant à l'ouvrage de type mémoire, il s'agit d'un projet de maîtrise présenté par un étudiant en sciences politiques (Abdel-Sattar Ghammouri).

2. Délimitation du concept de reformulation

Pour qu'une communication s'établisse correctement entre deux êtres humains ou plus, il est nécessaire que le discours qui la véhicule soit bien structuré. La structuration se fait grâce à des procédés d'enchaînement bien déterminés qui doivent refléter la cohésion interne du produit final. Ces procédés, comme l'affirme Manuel Célio Conceição (2005 : 73), ont généralement "un caractère anaphorique ou cataphorique, puisqu'ils répètent et/ou reprennent les affirmations antérieures pour faire avancer le discours". La reformulation fait partie de ces procédés puisqu'elle consiste à revenir ou retourner sur le contenu linguistique et cognitif d'une formulation en la présentant autrement afin de l'élucider. Elle est ancrée dans un acte psychologique qui veut que l'émetteur ait toujours peur que son discours se heurte à l'incompréhension du récepteur, c'est pourquoi il sent parfois le besoin de le reformuler, c'est-à-dire de le dire autrement.

Philippe Thoiron et Uzoma Chukwu (1989 : 25) délimitent schématiquement la reformulation. Sur un axe orienté, simplifié et ramené à deux positions P_1 et P_2 qui représente linéairement la dimension chronologique du discours, ils considèrent que "dans le procès de la reformulation un élément X [qui est dans notre étude le terme ou le reformulé] est mis en relation, au moyen d'un relateur (Rel), avec un élément Y [qui représente la reformulation de X] afin d'améliorer la

perception du contenu de X, en lui attribuant une forme différente. Cette mise en relation de X et Y doit être le résultat d'une action délibérée, d'un arrêt sur le terme X dans le cadre d'un acte explicatif".

Nous suivons dans notre étude cette représentation qui sera, comme nous le verrons ci-après, d'une grande utilité afin de distinguer la reformulation de certains concepts qui lui sont contigus et présenter également une typologie des reformulations des termes arabes d'Internet que viennent illustrer plusieurs exemples de contexte.

En revanche, nous mettons en relief - à travers nos divers exemples - l'aspect socio-communicatif qui gouverne la reformulation ainsi que la fonction didactique et interactive qui inscrit la reformulation dans un processus dynamique d'acquisition des connaissances.

3. Rapport reformulé / reformulation

Il est important de traiter la nature du lien sémantique entre les deux formulations X et Y, et de jeter la lumière sur ce qu'on appelle "connecteurs de reformulation" ou "relateurs". Le résultat de la reformulation est la création de liens d'équivalence entre X et Y. Mais de quelle équivalence s'agit-il ? D'une quasi-équivalence, d'une équivalence totale et parfaite, etc. ? Manuel Célio Conceição (2005 : 82) a imaginé quatre représentations possibles pour relier X à Y :

a) $X = Y$

b) $X > Y$

c) $X < Y$

d) $X \neq Y$

L'alinéa a) représente le cas des équivalences totales, ce qui veut dire que tous les traits conceptuels du terme X se trouvent exprimés dans la reformulation Y. Or, dans ce cas là nous aurons affaire à des définitions de type encyclopédique où il y a une description et une énumération exhaustives des propriétés d'un concept donné.

L'alinéa b) décrit parfaitement le mécanisme de reformulation. D'ailleurs, les exemples de notre corpus l'attestent clairement. "La reformulation ne reprend qu'une partie de l'information sémantique et conceptuelle du reformulé. Ceci se justifie pour des raisons d'économie de la structuration discursive" (Manuel Célio Conceição, 2005 : 82).

Dans la possibilité c), la reformulation donne plus de détails sur le concept que le reformulé. C'est un cas qui est beaucoup moins fréquent que le précédent.

La possibilité d) nous éloigne de la reformulation proprement dite, car moins il y a des traits conceptuels communs à X et Y, plus on se rapproche de l'introduction d'un nouveau concept différent du premier. Cette possibilité représente "le cas des reformulations faites pour corriger une dénomination ou un usage terminologique, ou pour, à partir d'un concept connu, en introduire un nouveau" (Manuel Célio Conceição 2005 : 82).

3.1. Les relateurs

Les expressions qui font la liaison entre l'élément X et sa reformulation Y sont appelées "relateurs". Leur rôle est de déclencher le processus de reformulation et de l'introduire dans le texte. Ces relateurs ont été décrits dans les textes relevant du vocabulaire général. Corinne Rossari[5], par exemple, cité par Manuel Célio Conceição (2005 : 92), établit une distinction entre ces relateurs ou connecteurs de reformulation et l'ensemble des connecteurs argumentatifs en disant que "le locuteur ne les utilise pas [les argumentatifs] pour présenter dans le point de vue q une nouvelle interprétation de p, mais pour assigner au point de vue introduit un statut par rapport à celui auquel ils renvoient ; q pouvant être suivant le choix du connecteur, soit un argument pour p, soit un contre-argument pour p, soit encore la conclusion de p".

Pour ce qui est des discours techno-scientifiques, le sujet n'a été traité de façon détaillée que dans les travaux de Philippe Thoiron et Uzoma Chukwu. Les exemples que nous citons ci-dessous présentent aussi en détail des relateurs en arabe présents dans notre corpus. Il existe deux types de relateurs, les uns linguistiques, les autres non linguistiques d'ordre typographique, tels les deux-points, les points-virgules, les guillemets, etc. Comme nous le verrons plus loin, les relateurs introduisent plusieurs types de reformulation et sont polyfonctionnels. Compte tenu de leur polyfonctionnalité et de leur dépendance contextuelle, il est difficile, voire impossible d'établir une liste close des relateurs.

5 Rossari C. Les opérations de reformulation, analyse du processus et des marques dans une perspective contrastive français-italien, Berne, Peter Lang, 1997, p. 9.

4. Reformulation et concepts contigus

Pour bien circonscrire le concept de reformulation, il est important de le différencier de certains concepts voisins comme la paraphrase, la définition, l'anaphore ou encore la synonymie. Pour le faire, nous suivons la méthode adoptée par Philippe Thoiron et Uzoma Chukwu (1989 : 25). Ces deux auteurs ont fait appel à deux types d'entités, les premières (notées X, X', X"…) sont des termes et les secondes (notées Y, Y', Y",…) sont des syntagmes libres, et ont dressé quatre critères concernant le mécanisme de la reformulation :

1. "appartenance au système de nomination d'au moins une des deux entités mises en relation (X satisfait à ce critère, Y non) ;

2. position relative de X et Y dans la chaîne du discours : X avant Y ;

3. position relative de X et Y sur un continuum "opacité ---transparence" : X avant Y ;

4. caractère délibéré de la fonction explicative de X par Y".

Les critères 1 et 2 sont binaires, et en les combinant ils donnent lieu aux quatre représentations théoriques suivantes :

	P_1	P_2
A	{X, X', X" ,…}	{Y, Y', Y" ,…}
B	{Y, Y', Y" ,…}	{X, X', X" ,…}
C	{X, X', X" ,…}	{X, X', X" ,…}
D	{Y, Y', Y" ,…}	{Y, Y', Y" ,…}

Figure 1. empruntée à Philippe Thoiron et Uzoma Chukwu (1989 : 26)

Parmi ces quatre cas, nous pouvons d'emblée éliminer le cas D puisqu'il s'agit de mettre en relation deux associations libres et non pas un terme avec une association libre. Pour les trois autres cas, les deux auteurs ont fait aussi appel au critère numéro 3. Etant donné que le cas A satisfait aux trois premiers critères, il constituera le prototype de toute reformulation. Le terme est en P_1 et l'association libre en P_2. On part tout d'abord de l'élément inconnu et on cherche à l'expliquer. Dans le cas B, où les critères 1 et 3 sont satisfaits, le processus est inversé, c'est-à-dire on part de ce qui est supposé être connu pour aboutir au terme inconnu. A et B forment ce qu'on appelle reformulation. Le cas C répond au critère 1 mais non pas au critère 2. Dans ce cas, nous avons affaire à deux termes commutables ; ainsi on parlerait de synonymie. Cependant, peut-on considérer la synonymie comme un moyen de reformulation ? Philippe Thoiron et Uzoma Chukwu (1989 : 27) apportent la réponse suivante : "le critère 3

donne un élément de décision pour un problème qui repose notamment sur les relations entre morphologie et syntaxe et entre opacité et transparence. Si l'un des deux X est, morphologiquement, plus transparent que l'autre, il pourra servir à le reformuler. On peut concevoir qu'un affixe, resté productif, joue, au sein d'un lexème, le même rôle qu'une structure syntaxique donnée au sein d'une association libre […] On admettra donc que C constitue bien un cas de reformulation si le critère 3 est satisfait".

Le critère 4 sert à faire la distinction entre reformulation et anaphore. L'anaphore lexicale a pour but de fuir la répétition et de modifier une visée. Or, s'agissant de la reformulation, la fonction explicative est délibérée. Le terme est introduit en discours et appelle immédiatement une explication pour faciliter sa compréhension. "D'un point de vue formel, ceci se manifeste par la proximité dans la chaîne du discours et par la présence de formes lexicales qui peuvent être considérées comme relateurs. Ces deux manifestations permettent d'apprécier le caractère délibéré de l'explication de X par Y, étant entendu que là aussi il s'agit d'un continuum" (Philippe Thoiron et Uzoma Chukwu, 1989 : 28).

Reformulation et définition appartiennent à deux plans distincts. À la suite de Philippe Thoiron et Uzoma Chukwu, nous considérons la reformulation comme un mécanisme intra-discursive propre aux écrivains ou orateurs alors que la définition est un procédé lexicographique propre aux lexicographes. Au sein des discours techno-scientifiques, nous préférons parler de reformulations définitoires.

Certains auteurs ne font pas la différence entre paraphrase et reformulation. Il s'agit surtout de Marie-Françoise Mortureux qui les prend pour deux dénominations interchangeables[6]. Nous considérons la paraphrase, comme une relation entre "phrases", une sorte de description différente d'une même situation référentielle, alors que la reformulation traitée dans une optique terminologique comme la nôtre, se situe au niveau inférieur de la phrase.

5. Typologie des reformulations à travers l'étude du corpus

Après avoir défini le cadre théorique de la reformulation et montré son importance au sein des discours techno-scientifiques, nous passons à

6 Marie-Françoise Mortureux (1982 : 51 sq.) étudie le fonctionnement sémiotique du discours métalinguistique en identifiant et reliant deux entités ou séquences X et Z par une relation d'équivalence uniquement où Z = X, Z étant le terme scientifique et X une paraphrase non marquée scientifiquement.

la description de ce phénomène tel qu'il se présente dans notre corpus. Dans les extraits que nous présentons ci-après, les relateurs sont soulignés, le reformulé ou le terme est mis en **gras**, et la reformulation ou la formulation du terme en *italique*. Nous nous sommes basée, comme nous l'avons déjà indiqué, sur la typologie tri-catégorielle de Philippe Thoiron et Uzoma Chukwu (1989 : 29 *sq.*) qui comprend les **reformulations construites (DU TYPE Rel (X, Y))**, les **reformulations synonymiques (DU TYPE Rel (X, X'))**, et les **reformulations énumératives (TYPE Rel (X, X', X'', …))**[7] et que nous avons, à notre tour, largement développée en y ajoutant les catégories suivantes : **reformulations définitoires (DU TYPE Ø (X, Y))**, **reformulations comparatives (DU TYPE Rel_{cp} (X, Y))**, **reformulations par fonction (DU TYPE Rel_{fc} (X, Y))**, **reformulations antonymiques (DU TYPE Rel_{at} (X, Y))** et **reformulations à deux ou plusieurs étages**.

5.1. Reformulations construites (DU TYPE Rel (X, Y))

Elles ont une fréquence très élevée dans notre corpus. Elles comprennent les sous-catégories suivantes :

1 Les reformulations copulatives

Elles consistent essentiellement en des reformulations où X et Y sont mis en relation par le pronom هو pour le masculin et هي pour le féminin[8]. Elles fonctionnent le plus souvent au plan référentiel.

A Les reformulations copulatives (DU TYPE Rel (X, Y))

Ce genre de reformulation fait partie du groupe de reformulations prototypiques ou X précède Y ; "la reformulation étant une aide à la présentation de l'inconnu, il est logique, dans un premier temps, de penser qu'elle interviendra surtout à la fin des textes" (Philippe Thoiron et Henri Béjoint 1991 : 101). **Exemple :**

الإنترنت (Internet) هي *شبكة معلومات عالمية عبارة عن مجموعة من شبكات الحاسب موصولة مع بعضها البعض* (COREA/RA, 2001 : 19).

(**Internet** est un réseau d'information mondial constitué d'un ensemble de réseaux d'ordinateurs reliés entre eux).

---

7 À noter que nous excluons de notre étude deux sous-catégories de reformulations présentées par Philippe Thoiron et Uzoma Chukwu puisque nous n'avons trouvé aucun exemple dans notre corpus qui puisse être casé sous ces rubriques : il s'agit des reformulations métalinguistiques explicatives inverses et des reformulations énumératives par liste exhaustive discontinue.

8 À ce pronom correspond en français le verbe copule "être" ainsi que le verbe "to be" en anglais. En arabe, il faut savoir que le pronom /huwa/ n'est pas une copule.

B Les reformulations copulatives (DU TYPE Rel (Y, X))

Ce schéma est l'inverse du précédent ; Y précède X. Notons que ce genre de reformulation est très rare dans notre corpus, ce qui conforte et appuie les propos précédents de Philippe Thoiron et de Henri Béjoint. "Le mode d'introduction du terme X dans le discours […] se rapproche du processus de la dénomination" (Philippe Thoiron et Uzoma Chukwu, 1989 : 31).
Exemple :

إن لم يكن هناك عنوان، يجب على المُستعمِل تشغيل *برمجيّة اضافيّة مختصّة في البحث عن مواقع الواب* وهي **محرّكات البحث** [...] (CORTU/BAZM, 2005 : 82).

(S'il n'y a pas d'adresse, l'utilisateur doit lancer *des logiciels supplémentaires ayant pour fonction de rechercher les sites Web* et qui sont appelés **moteurs de recherche**).

2 Les reformulations métalinguistiques

Ce genre de reformulations fait intervenir des relateurs qui rendent évident, dans le texte même, l'aspect métalinguistique du procès de reformulation. Ces relateurs sont, le plus souvent, des verbes appartenant au champ sémantique de la dénomination :

يُدعى / تُدعى، يُعرَف بـ / تُعرَف بـ ، يُسمّى / تُسمّى ...[9]

Ce genre de reformulations est appelé reformulations appellatives. D'autre part, on trouve souvent dans les textes des verbes comme يَعني / تَعني (vouloir dire) ou نَعني بِ (entendre par) et des particules comme أي (c'est-à-dire) qui marquent la présence d'une reformulation explicative.

A Les reformulations métalinguistiques appellatives

Elle se divise en deux catégories : directes et inverses.

a) Les reformulations appellatives directes

Ces reformulations se présentent sous la forme prototypique Rel (X, Y).
Exemple :

تُعرَف **الدودة** في مجال المعلوماتيّة بأنها *برنامج ينشر نفسه عبر الشبكة مستعملا مصادر حاسوبه الحالي لإصابة حواسيب أخرى* (CORSY/GAS, 1996-1997 : 92).

(On appelle **worm**[10] en informatique le programme qui se reproduit par lui-même et se propage à travers le réseau [ici Internet][11] pour s'attaquer à d'autres systèmes informatiques).

9 Leurs équivalents dans les textes français sont : appeler, nommer, baptiser, etc. Et dans les textes anglais : to call, to name, to term, to be referred to as, etc.

b) Les reformulations appellatives inverses

Elles ont la forme Rel (Y, X). Nous avons relevé **l'exemple** suivant :

يوجد عادة عدد من *الشَّركات أو المؤسَّسات التي تؤمِّن خدمة ربط المستثمرين بشبكة الإنترنت*. تُدعى مثل هذه المؤسسة **"مُزوِّد خدمة الإنترنت"** (Internet Service Provider) [...] (CORSY/AH, 2003 : 3).

(Il existe généralement un certain nombre d'*entreprises qui proposent aux clients des prestations de services Internet* ; on les appelle "**fournisseurs d'accès à Internet**" (Internet Service Provider) […]).

B Les reformulations métalinguistiques explicatives

Il n'y a qu'une seule catégorie représentée dans notre corpus. Il s'agit des reformulations métalinguistiques explicatives directes. Comme nous l'avons déjà signalé, elles sont représentées à l'aide de la formule "c'est-à-dire" ou bien à l'aide d'un verbe faisant explicitement référence à des processus cognitifs ("entendre" au sens de "comprendre").

a) Les reformulations explicatives directes

تعمل خدمة الوب **بالنموذج زبون/مُخَدِّم** (Client/Serveur)، أي *أنّ برنامجا زبونا يعمل عادة على حاسوب المستثمِر (المستعرض) يُرسِل طلباً بالبيانات المرغوبة إلى البرنامج المُخَدِّم الذي يعمل على حاسوب أخر في مكان ما على شبكة الإنترنت. عندما يتلقّى المُخَدِّم الطلب فإنه يُرسِل البيانات إلى برنامج المستعرض عبر الشبكة*. (CORSY/AH, 2003 : 11)

(Le Web fonctionne grâce au **modèle Client/Serveur**, c'est-à-dire qu'un logiciel client se trouvant généralement dans l'ordinateur de l'utilisateur envoie une demande de ressources souhaitées à un logiciel serveur situé quelque part dans un autre ordinateur relié à Internet. Lorsque le serveur reçoit la demande, il envoie alors les données voulues au logiciel client).

3 Reformulations appositives

---

10 Nous avons eu recours à l'équivalent anglais car il est très utilisé par les informaticiens et dans les textes français relevant du domaine de l'informatique. De plus, en posant la question à une doctorante en informatique à l'INSA de Lyon, elle nous a affirmé que l'équivalent français "ver informatique" que nous avons trouvé sur le site du Grand dictionnaire terminologique (http://www.granddictionnaire.com) dans une fiche qui date de 2005 est peu ou même pas connu par la grande majorité des informaticiens et que c'est bien la forme empruntée à l'anglais qui figure dans le discours français de l'informatique.

11 Il s'agit bien ici du réseau des réseaux "Internet" et non pas d'un autre réseau car ce contexte est tiré d'un chapitre qui traite des problèmes de sécurité dans Internet. À noter aussi que dans notre corpus /aš-šabaka/ (le réseau) utilisé seul est une variation contextuelle des termes /al-?Intirnit/ (Internet) et /šabakat al-?Intirnit/ (réseau Internet).

Par opposition à toutes les reformulations que nous avons citées jusqu'ici, les reformulations appositives sont celles où les deux éléments X et Y sont mis en relation au moyen d'une marque formelle appelée indice de reformulation[12]. Les indices de reformulation sont représentés par des signes typographiques, comme les virgules dans cette reformulation appositive directe (Rel (X, Y)) :

من أخطر ما يواجهُكَ وأنتَ على الشبكة **(الهاكرز)**[13] أو «قراصنة» الشبكة، *أولئك الذين يستطيعون اختراق جهازكَ أو التأثير فيه بشكل أو بآخر*، فقد يستطيع أحد المفسدين أن يخترق جهازكَ ويطّلع على كل ما فيه، ويفعل به ما يريد. (COREG/AMA, 2001 : 43)

(La chose la plus dangereuse à laquelle tu auras affaire sur Internet, est la présence des **(hacker)** ou "pirates", *ceux qui peuvent sonder ton système informatique afin de pouvoir éventuellement s'y immiscer*, un gâcheur peut donc pénétrer ton ordinateur, voir tout son contenu et en faire tout ce qu'il veut).

5.2. Reformulations synonymiques

Le seul critère qui permet de différencier les reformulations synonymiques des synonymes est celui de la position relative de X et X' sur l'axe "opacité - transparence". Nous considérons qu'une unité est la reformulation d'une autre si elle est un mot du vocabulaire général ou si elle fait partie de la terminologie du domaine en question[14]. Parmi les relateurs qui lient les deux éléments X et X', nous trouvons la conjonction أو (ou). **Exemple :**

إنّ التطوّر المنطقي لمفهوم لوائح البريد هو *المؤتمرات التخاطبيّة* أو **مجموعات الأخبار** newsgroup. (CORSY/AH, 2003 : 8)

(Les listes de diffusion se sont développées et ont par conséquent abouti à la création de *forums de discussion*[15] ou **groupe de nouvelles** newsgroup).

---

12 Par commodité et pour maintenir la même formule tout au long de notre démonstration, à savoir Rel (X, Y) ou l'inverse, ces indices sont aussi appelés "relateurs".

13 On assimile parfois le bidouilleur (ou hacker) au pirate informatique - comme c'est le cas dans ce contexte arabe – parce que les deux notions sont souvent désignées par le même terme anglais "hacker". Cependant, l'utilisation de l'anglais "hacker" dans le sens de "pirate informatique" prête à confusion. Le premier n'a pas d'intention malveillante, c'est avant tout un programmeur informatique qui a pour seul but de faire évoluer ses connaissances et celles des autres. Alors que le second, "pirate" ou "cracker" en anglais est un criminel informatique.

14 À noter que d'autres critères ont été mentionnés et développés par Philippe Thoiron et Uzoma Chukwu concernant des textes techno-scientifiques anglais et français.

15 Ce qui justifie le choix de cet exemple, bien que "forum de discussion" soit "un terme" en français, est le fait que son calque arabe /al-muctamarât at-takâÔubiyya/ ne l'est pas puisqu'il n'est cité qu'une seule fois dans toutes les sources de notre corpus, une fréquence généralement largement insuffisante pour accéder au rang d'un terme du domaine.

5.3. Reformulations énumératives (TYPE Rel (X, X', X'',…))

Comme leur nom l'indique, ces reformulations énumèrent des membres de la classe à laquelle renvoie le terme.

1 Reformulations énumératives par liste exhaustive

À la suite de Philippe Thoiron et Uzoma Chukwu (1989 : 45), nous considérons qu'une reformulation appartient à cette classe, "s'il n'y a, dans le texte, aucune mention explicite indiquant qu'il s'agit simplement de quelques membres de la classe".

A Reformulations énumératives par liste exhaustive continue

يَستخدم **الوب** ثلاثة تقنيّات جديدة : *HTML أو Hyper Text Markup Language التي تُستخدَم لكتابة صفحات الوب؛ ومُخدِّم الوب الذي يَستخدم البروتوكول HTTP أو Hyper Text Transfer Protocol لإرسال صفحات الوب إلى الزبائن، ومُستعرض الوب الذي يتلقّى البيانات من المُخدِّم ثم يفسِّرُها ثم يعرض النتيجة* (CORSY/AH, 2003 : 10) .

(Le **Web** fonctionne selon trois nouvelles techniques : le langage HTML (Hyper Text Markup language) qui sert à rédiger les pages Web ; le serveur Web qui utilise le protocole HTTP (Hyper Text Transfer Protocol) pour l'envoi des pages Web aux clients, et le navigateur Web qui reçoit les données du serveur, les interprète et les diffuse).

Cet exemple comporte un indice de reformulation, les deux-points, qui est le relateur de reformulations énumératives par excellence. Ce sont des membres de la classe caractérisant le fonctionnement du Web qui sont cités.

2 Reformulations énumératives par exemplification

Ce type de reformulations consiste à donner quelques exemples des membres de la classe. Les relateurs qui mettent en relief cette reformulation sont : مثلاً، نذكر منها، على سبيل المثال، مثل، إلخ[16]

وإذا كان لديكَ برنامج **مُستعرض للإنترنت Internet Browser** مثل *إنترنت إكسبلورر Internet Explorer* أو *نيتسكيب نافيغيتر*، فإنه يمكنكَ الإنتقال إلى جزء (استعراض الإنترنت) [...] (CORAL/BM, 2002 : 19-20) .

(Si tu possèdes un **navigateur** Internet Browser comme *Internet Explorer ou Netscape Navigator*, tu pourras ainsi commencer (la navigation sur Internet) […]).

16 Les équivalents français de ces expressions sont "par exemple", "comme", "tel", etc. ; en anglais nous trouvons dans les textes "such as", "for example", "for instance", "e.g.", etc.

5.4. Reformulations définitoires

La présentation de la vedette séparée de l'article ou de l'explication ainsi que l'absence de relateurs lexicaux rapprochent énormément ces reformulations des définitions telles qu'elles se présentent dans les dictionnaires. Elles sont toutes des reformulations directes auxquelles nous attribuons la formule suivante : Ø (X, Y) où l'ensemble vide renvoie à l'absence d'un relateur lexical. **L'exemple** archétype de cette représentation est tiré de la partie tunisienne de notre corpus :

البريد الإلكتروني "المايل" **(e_mail)** : *تُتيح هذه الخدمة قراءة وكتابة الخطابات وارسالها واستقبالها أوتوماتيكيا* (CORTU/BAZM, 2005 : 81).

(**Le courrier électronique (e_mail)** : ce service permet de lire, d'écrire, d'envoyer et de recevoir électroniquement des courriers).

5.5. Reformulations comparatives

L'introduction de nouveaux concepts est donc un processus qui fait appel à différentes sortes de reformulations qui sont des moyens d'aide à l'appropriation des traits conceptuels. L'un des moyens est de mettre ces concepts en relation de comparaison avec ce qui est déjà connu. C'est ce que fait d'une manière intéressante l'ingénieur Ahmed Rayyan dans plusieurs parties de son livre. Ces reformulations peuvent être schématisées à l'aide de la formule Rel_{cp} (X, Y) contenant une expression qui marque la comparaison. **Exemple :**

وتَعتمد (WWW) أساساً على ما يُسمّى **بهايبرتكست** (Hypertext) والهايبرتكست هي طريقة لإدارة المعلومات. ويمكن تمثيلها *بكتاب يحتوي على صفحات، والصفحات تحتوي على مقطوعات وكلمات. وعند قراءتكَ للكتاب تمرّ على صفحة صفحة، كما يمكن القفز من صفحة إلى أخرى بعيدة، ثم الرّجوع مرّة أخرى إلى الصفحة الحالية* (COREA/RA, 2001 : 155-156).

(Le (WWW) est essentiellement basé sur **l'hypertexte** (Hypertext) qui est une méthode permettant la gestion des données et qui peut être comparée à *un livre qui contient des pages, les pages contenant à leur tour des paragraphes et des mots. Tu peux lire le livre page par page, comme tu peux sauter d'une page à une autre qui est plus loin, puis revenir à nouveau à la page actuelle*).

5.6. Reformulations par fonction

Compte tenu de la nature même des discours techniques axés sur deux points la compréhension et l'action, et qui à la différence des textes scientifiques dans lesquels il s'agit d'une formation purement intellectuelle, ces discours présentent des reformulations de termes qui tiennent compte

de la fonction des concepts afin que le lecteur-utilisateur puisse saisir le contenu de l'information pour la mettre en pratique. Nous représentons ce genre de reformulations de la manière suivante : Rel_{fc} (X, Y). **Exemple :**

تتمثل وظيفة **(Archie)** في *عملية البحث عن الملفّات في موزّعات (FTP) في جميع أنحاء العالم* [...] (CORTU/BAZM, 2005 : 77).

(**(Archie)** a pour fonction de *rechercher les fichiers sur les serveurs (FTP) dans tous les coins du monde*).

L'utilisateur peut savoir, d'après les informations contenues dans cet extrait de contexte, qu'il faut passer par Archie afin de chercher les fichiers sur les serveurs FTP.

5.7. Reformulations antonymiques

Comme son nom l'indique, ce genre de reformulations met en relation deux concepts antonymes. Nous le formulons par : Rel_{at} (X, X'…), suivis de leurs éléments (Y, Y'…) qui ont pour fonction de mettre en relief les points de divergence annoncés par le relateur. **Exemple :**

ويجب التمييز أو التفريق بين **اليوزنت والانترنيت، فالانترنيت** هي *تلك الشبكة الواسعة التي تنقل بيانات من أنواع وفئات متعدّدة* أمّا **اليوزنت** *فهي احدى هذه الفئات من البيانات التي يتمّ نقلها عبر الانترنيت* (CORSY/GAS, 1996-1997 : 19).

(Il faut distinguer **Internet** de **Usenet** ; **Internet** *est ce vaste réseau qui opère un transfert de données de divers genres et catégories,* quant à **Usenet,** *elle constitue l'une de ces catégories de données transférées via Internet*).

5.8. Reformulations à deux ou plusieurs étages

Ce cas peut être expliqué par le fait que l'auteur s'adresse tout d'abord aux néophytes et cherche dans un deuxième temps à décrire un nouveau domaine, d'où il sent le besoin de revenir plus d'une fois sur le terme ou le concept (la plupart des ouvrages cités datent de l'année 2001 ou 2002, peu après l'introduction d'Internet dans les pays arabes et sa mise à la disposition du grand public). Les exemples sont nombreux, nous en citons un comprenant une reformulation appellative inverse complétée par une reformulation copulative (DU TYPE Rel (X, Y)) :

تتضمّن أغلب شبكات الحاسب *تقنيّة أمنيّة* تُدعى **الجدار النّاري (Firewall)**. **والجدار الناري** هو *نظام برمجي (يترافق غالبا مع أجهزة خاصة) يشكّل حاجزا يمنع المستثمرين غير المرخّص لهم من خارج الشبكة من الوصول إلى الموقع*. (CORSY/AH, 2003 : 11)

(La plupart des réseaux informatiques renferment *un dispositif de sécurité* appelé **pare-feu (Firewall)**. **Le pare-feu** est *un système logiciel (accompagnant*

souvent des systèmes spécifiques) qui constitue une barrière empêchant les utilisateurs qui n'ont pas d'autorisation d'accéder au réseau interne).

6. Conclusion

Les textes ont été pour nous un outil nécessaire et fondamental pour examiner comment les spécialistes ont présenté l'information ou les informations sur un domaine. Ils nous ont aidée à réfléchir sur la manière dont les connaissances doivent être transmises selon le niveau des récepteurs, car la conceptualisation des faits scientifiques ne semble pas être la même selon les milieux, selon les domaines, selon les langues et selon les communautés linguistiques. À travers les nombreuses reformulations recueillies de notre corpus arabe d'Internet, nous avons pu constater et relever la richesse et l'utilité du processus de reformulation quant à la diffusion du savoir parmi le grand public. Ainsi, et grâce à ce processus, l'information techno-scientifique devient à la disposition de tout le monde et l'acquisition des connaissances se fera par un moyen riche et diversifié et ce au sein du discours et à travers une terminologie dynamique qui décrit, qui propose et qui revêt trois dimensions : communicative, sociale et cognitive. Il ne faudrait jamais oublier qu'une terminologie en dehors de l'usage est une terminologie morte et insignifiante. La terminologie possède à son compte - le compte du passé et celui du futur - tous les atouts nécessaires pour innover et progresser. À partir de différents types de textes elle pourrait aider à améliorer la manière d'introduire les termes dans ces textes pour qu'ils s'adaptent le mieux au public visé.

Bibliographie

Ouvrages et revues

Conceição M. C. (2005) : *Concepts, termes et reformulations*, Travaux du C.R.T.T., Lyon, Presses universitaires de Lyon, 279 p

Elqasem F. (2003) : "Le rôle de la reformulation dans la traduction des textes de spécialité vers l'arabe", in Hamzé H. et Ougammadan M., dir., *La terminologie au service du traducteur,* Turjumān, Maroc, Ecole Supérieure Roi Fahd de Traduction – Tanger, vol. 12, n° 1, pp. 39-54

Grand dictionnaire terminologique, Office québécois de la langue française (http://www.granddictionnaire.com)

Jacobi D. (1984) : "Du discours scientifique, de sa reformulation et de quelques usages sociaux de la science", *Français technique et scientifique : reformulation, enseignement,* Langue française, Paris, Larousse, n° 64, pp. 38-52

Jacobi D. (1994) : "Lexique et reformulation intradiscursive dans les documents de vulgarisation scientifique", *in* CANDEL, Danielle, dir., *Français scientifique et technique et dictionnaire de langue*, Paris, Didier Érudition, Coll. Études de sémantique lexicale, pp. 77-91

Martinet A. (1967) : *Éléments de linguistique générale*, Paris, Armand Colin, 221 p

Mortureux M.-F. (1982) : "Paraphrase et métalangage dans le dialogue de vulgarisation", *La vulgarisation*, Langue Française, Paris, Larousse, n° 53, pp. 48-61

Thoiron Ph. et Chukwu U. (1989) : "Reformulation et repérage des termes", *La banque des mots*, Paris, Conseil International de la Langue Française, n° spécial, pp. 23-50.

Thoiron Ph. et Béjoint H. (1991) : "La place des reformulations dans les textes scientifiques", *Meta*, Montréal, Presses de l'université de Montréal, vol. 36, n° 1, pp. 101-110

Corpus

Abdel-Mawla A. (2001) : *Taʿallam al-?intirnit fî talât sâʿât*, Le Caire, ad-Dâr ad-Dahabiyya, 54 p

Abed H. (2003) : *?Istiṯmâr al-?intirnit*, Damas, Cours à l'Université de Damas, 31 p

Béchir M. (2002) : *al-?Intirnit lil-mubtadi?în*, Algérie, Dâr al-Maᶜrifa, 75 p. + glossaire des termes d'Internet

Ben Abdallah Zayed M. (2005) : *Madkal ?ilâ ʿâlam al-?intirnit*, Tunis-Carthage, Phéni Éditions, 99 p

Ghammouri A.-S.(1996-1997) : *Šabakat al-?ittiÒâl ad-dawlî al-?intirnît INTERNET*, Mémoire de maîtrise en sciences politiques, Syrie, Institut Supérieur des Sciences Politiques, 115 p.

Rayyan A. (2001) : Kadamât al-?intirnit, Abu Dhabi, al-Maµmaᶜ at-taqâfî, 242 p.

A propos des auteurs

Andrée Affeich
Centre de Recherche en Terminologie et Traduction (CRTT)
Université Lumière Lyon 2
86, rue Pasteur
69365 LYON Cedex 07
Andree.Affeich@univ-lyon2.fr
http://recherche.univ-lyon2.fr/crtt/

SESSION 3

Le projet NucSTML Structuration d'un dictionnaire de spécialité en vue de sa publication sur internet Bénéfices du langage XML.

Jacques JOSEPH

Résumé : Le langage XML est particulièrement bien adapté à la représentation d'un dictionnaire de spécialité destiné à être diffusé sur le Web. La simplicité du langage assure l'interopérabilité et la portabilité des documents et rend possible leur diffusion sur une grande variété de systèmes d'exploitation. Le format est facile à mettre en œuvre, à installer sur un serveur, il est économe en ressources. Il sépare le fonds de la forme, facilitant ainsi la gestion du contenu et assurant la pérennité du document. Les feuilles de styles et les langages associés permettent de formuler des requêtes puissantes et des fouilles profondes et rapides dans l'arborescence du fichier source, il peut être une aide à l'élaboration de réseaux conceptuels. L'application présentée dans cet article fait suite aux travaux de Marie Calberg pour l'élaboration d'un dictionnaire de l'ingénierie nucléaire (Calberg-Challot *et al.* 2008a) et décrit sa mise en œuvre au sein d'une grande entreprise du secteur de l'énergie. Le projet NucSTML (Nuclear Science and Technology Markup Language) a pour objet de réaliser un modèle de document simple, adapté à la publication et à l'échange de données linguistiques dans le domaine des sciences et techniques de l'énergie nucléaire.

Mots-clés : Modèle, Dictionnaire, Nucléaire, Internet, XML

1. Introduction

Pour faire face à l'évolution des techniques et du vocabulaire associé, à l'élargissement de leur périmètre et à l'arrivée d'une nouvelle génération de femmes et d'hommes dans les domaines techniques, commercial ou administratif, ainsi qu'aux besoins de la société, certains acteurs de l'industrie et de la recherche nucléaire française ont entrepris de mettre à jour leurs principaux dictionnaires techniques voire d'en créer de toute pièce. Le Commissariat à l'Énergie Atomique (Cea) a publié en 2008 sous la conduite de M. le Haut Commissaire et avec la collaboration active de ses principaux partenaires - Autorités de Sûreté Nucléaire (ASN), Institut de Recherche sur la Sûreté Nucléaire (IRSN), Areva, Edf, Andra - notamment, le *Dictionnaire des sciences et techniques nucléaires* (Bigot *et al.* 2008), regroupant près de quatre mille entrées qui couvrent l'ensemble des secteurs : scientifique et industriel, civil ou de défense, ce dictionnaire constituant la mise à jour d'un ouvrage plus très ancien (Dictionnaire des sciences et techniques nucléaires). Areva-Np a de son côté entrepris de réaliser son propre dictionnaire de spécialité (Calberg-Challot *et al.* 2008a), (Calberg-Challot *et al.* 2008b) dédié à l'ingénierie nucléaire. Notons que le Cea, répondant à sa mission d'information, a diffusé l'ouvrage de façon publique, tandis que dans le cas des industriels, le contenu reste prioritairement réservé à un usage interne. La caution active des plus hautes autorités prouve l'importance qu'elles attachent à la maîtrise de la terminologie technique et à sa diffusion.

Cet article décrit une expérience industrielle transposable, présente une méthode pour réaliser de façon économique et efficace un dictionnaire de spécialité, facilement intégrable dans un site internet ou intranet, dans des applications de gestion documentaire, portable et gérable de façon simple et pérenne.

2. Analyse de la diffusion sur internet de la terminologie nucléaire

De nombreux organismes français ou étrangers, (lexique nucléaire de A à Z), des sociétés savantes telles que la Société Française de l'Énergie Nucléaire (SFEN), des électriciens francophones (Hydroquébec), des organismes de régulation (Glosary) ou des organismes internationaux tels que l'Agence Internationale pour l'Énergie Atomique (AIEA 2006), des laboratoires de protection (Lexique Cea), (Santé Canada : glossaire), proposent dans leur site internet, sous la dénomination *lexique* ou *glossaire*, des listes de termes techniques et scientifiques en rapport avec leur activité

et leurs métiers (production d'énergie, physique des réacteurs, protection des populations, biologie, médecine, contrôle, instrumentation...).

Ces listes sont généralement construites sur un modèle similaire : un contenu de termes communs tels que *actinides, ALARA*, *Becquerel, uranium* (physique fondamentale) ou de termes spécifiques *pressuriseur, réacteur* (domaine des réacteurs), *évacuation* (domaine de la protection), etc... On note deux méthodes de diffusion électronique : documents PDF (Postscript Document Format) de manipulation peu aisée, pages HTML (Hyper Text Markup Language) d'utilisation plus souple et mieux adaptée à une diffusion électronique, mais avec les limitations propres à ce langage (nombre de balises limité, pauvre d'un point de vue sémantique, mélange de la forme et du fonds...)

3. Historique et justification du choix du langage XML

3.1. Rappel des principales caractéristiques du langage

XML est un métalangage de balises, permettant de représenter et de stocker des documents structurés dans un format texte, de les partager, de les transformer en tout format compatible les rendant facilement publiables sur le Web. Les applications justifiant l'utilisation de XML sont variées et multiples : oeuvres poétiques ou théâtrales (Gooss 1999), (Jung 2000), mise en place de systèmes documentaires à caractère administratif par le gouvernement du Québec (Marcoux 2001), ontologie (Gandon 2002). Les DTD (**D**ocument **T**ype **D**efinition) permettent de créer des modèles de document auquel peuvent se référer différents groupes travaillant sur une thématique similaire. Par exemple l'initiative libre Text Encoding Initiative (TEI) préconise un standard de représentation des types de textes relevant d'une représentation XML (TEI).

3.2. Les dictionnaires et XML justification des choix

La publication en ligne de dictionnaires généraux ou techniques invoquant la technologie XML se développe dans de nombreux secteurs industriels ou de recherche. Dans la banque ou le bâtiment par exemple où l'interopérabilité, la standardisation des échanges, l'indépendance vis-à-vis des systèmes d'exploitation sont les arguments principaux (Henriet *et al.* 2006), (Méta modèle de Dictionnaire Standard de la Construction) en faveur de son usage. Pour d'autres, XML est : un moyen puissant de hiérarchisation et de structuration des données textuelles (Bernard *et al.*),

une méthode de balisage de dictionnaires anciens (Rey 2004) ou bien encore un moyen utilisé dans l'élaboration de dictionnaires multilingues (Mangeot *et al.* 2003), d'autres encore soulignent que l'accès à n'importe quelle partie d'un dictionnaire, grâce à XML amplifie la capacité de diffusion des versions électroniques d'un dictionnaire (Elchacar 2008). Notons enfin la standardisation de dictionnaires multilingues grâce à XML (Maks *et al.* 2008).

4. Construction du document XML

La méthode de collecte et de sélection des entrées, ainsi que la formulation des définitions adoptées par Areva-Np a été décrite par Marie Calberg (Calberg-Challot *et al.*), le fonds terminologique repose sur des ouvrages techniques, cf. par exemple (Coppolani *et al.* 2004). Un nombre limité de descripteurs, a été sélectionné, certains sont obligatoires d'autres facultatifs. Les données fondamentales sont considérées comme **élément :** *dictionnaire* (élément racine), *entry* (élément courant) contenant les éléments fils : *expression, équivalent(s) anglais, définition*, d'autres facultatifs : *note(s), exemples, renvoi(s), forme abrégée, synonyme(s)*, *illustration.* Les données complémentaires sont considérées comme **attribut** (ex : *langue*, *date* -de validation de l'entrée-, *domaine*, *auteur, statut* -approuvé-) : Chaque élément *entry* possède un identificateur unique et obligatoire sous forme d'attribut, *id,* un nom de *domaine,* chaque élément référence *(ref)* possède un attribut *idref* qui assure la relation avec la cible repérée par son identificateur. L'arborescence correspondante du dictionnaire est représentée sur la Figure 1.

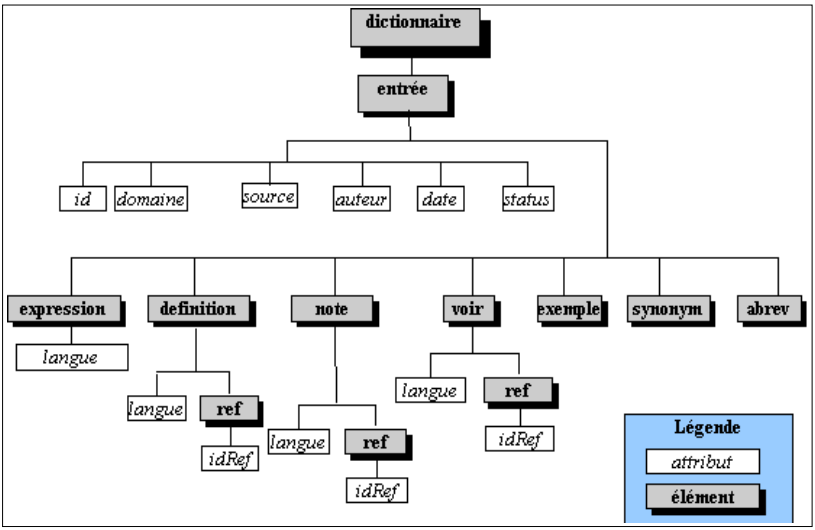


Figure 1. Arborescence général du dictionnaire

Le Tableau 1 représente le contenu de l'élément de base du document, la DTD est représentée dans le Tableau 2.

Le fichier XML est créé à partir du tableur source, à l'aide de langages de programmation classiques (Visual basic ou Php), qui permettent de transformer le contenu et de l'intégrer entre les balises. Simultanément la liste de l'ensemble des termes présents dans le dictionnaire est insérée dans un module *javascript*, servant de support à un formulaire de requête.

```
<entry id='dinId5' langue='fr' domain='P' resp='RCC-P' date='2009-02-15' status='GPA'>
<expression langue='fr'>accident</expression>
<expression langue='en'>accident</expression>
<synonym>transitoire accidentel</synonym>
<definition langue='fr'>Évènement hautement improbable, pouvant entraîner
l&apos;endommagement d&apos;une ou plusieurs <ref idref='dinId112'>barrières de confinement</ref>
avec éventuel relâchement de produits radioactifs et demandant
la mise en service des <ref idref='dinId1015'>systèmes de sauvegarde</ref> par
le <ref idref='dinId1005'>système de protection</ref>.</definition>
<note langue='fr'>Deux catégories d&apos;accidents sont prises en compte :
les accidents de fréquence très faible dits de condition 3,
et les accidents hypothétiques dits de condition 4.</note>
<voir><ref idref="dinId949">situation de troisième catégorie</ref>,
<ref idref="dinId948">situation de quatrième catégorie</ref></voir>
</entry>
```

Tableau 1. Exemple d'élément du dictionnaire

```
<!ELEMENT entry (illustration?|expressions+)>
<!ELEMENT expression (sub|sup)*>
<!ELEMENT definition (ref|sub|sup)*>
<!ELEMENT note (ref|sub|sup)*>
<!ELEMENT abrev (ref|sub|sup)*>
<!ELEMENT voir (ref|sub|sup)*>
<!ELEMENT exemple  (ref|sub|sup)*>
<!ELEMENT ref (sub|sup)*>
<!ATTLIST entry id ID #REQUIRED>
<!ATTLIST entry langue CDATA #IMPLIED>
<!ATTLIST entry categorie CDATA #IMPLIED>
<!ATTLIST entry resp CDATA #IMPLIED>
<!ATTLIST entry status CDATA #IMPLIED>
<!ATTLIST entry date CDATA #REQUIRED>
<!ATTLIST illustration source CDATA #REQUIRED>
<!ATTLIST illustration alt CDATA #IMPLIED>
<!ATTLIST expression language (fr|en) #REQUIRED>
<!ATTLIST definition language (fr|en) #REQUIRED>
<!ATTLIST note language (fr|en) #REQUIRED>
<!ATTLIST ref idref IDREF #REQUIRED>
<!ENTITY din SYSTEM "din.xml">
```

Tableau 2. DTD du document XML

4.1. Balisage des références.

XML permet d'organiser les renvois en marquant les termes cible, les instructions XSL se chargeant d'en assurer l'affichage de façon automatique, dans notre cas un renvoi est marqué par l'élément <ref>..</ref>, l'attribut *idref* a la valeur de l'identificateur du terme cible :

```
<ref idref='idValue'>mot cible</ref>
```

Cette méthode permet d'afficher de façon automatique les termes associés au moment de la requête sur le poste client, à l'aide de "templates" placés dans les feuilles de style XSL dont un exemple est représentée dans le Tableau 3. Le marquage peut être réalisé de façon automatique à l'aide d'algorithmes, soit au début du processus à l'aide de langages de programmation (Vb, Php), soit en fin de processus en utilisant les instructions XSL.

```
<xsl:template match='ref'>
<xsl:if test="//entry[@id=$idreference]">
<a><xsl:attribute name="href">
javascript:definition(&apos;
<xsl:value-of select="@idref"/>&apos)</xsl:attribute>
<xsl:apply-templates/>
</a>
```

<xsl:if> <xsl:template>

Tableau 3. Type de modèle XSL utilisé pour relier une référence à sa cible

4.2. Requêtes

L'association des langages, *HTML, javascript, XSL, XPATH,* permet de recueillir les arguments de la requête formulée à partir du poste client par l'intermédiaire d'un formulaire et de la transmettre par l'entremise d'un script *java script* sous forme de paramètres (Tableau 4) à la feuille de style *XSL* (Tableau 5), les instructions de la feuille de style organisent la requête à l'aide d'instructions *XPATH,* qui pointent les éléments recherchés et les retournent au poste client au format *HTML*. La page d'interface *HTML*, propose à l'utilisateur plusieurs manières de parcourir le dictionnaire pour afficher les définitions : soit directement à partir de la liste dynamique de l'ensemble des entrées, soit à partir d'une chaîne de caractères figurant dans le contenu de l'entrée (*expression, définition, notes, exemples*).

```
var xslDoc = new ActiveXObject("Msxml2.FreeThreadedDOMDocument")
var xmlDoc = new ActiveXObject("Msxml2.DOMDocument");
xslDoc.async= false;
xslDoc.load("dictionnaire.xsl");
xmlDoc.validateOnParse = false;
xmlDoc.async = false;
xmlDoc.load("dictionnaire.xml");
var xslt = new ActiveXObject("Msxml2.XSLTemplate");
xslt.stylesheet = xslDoc;
var xslProc = xslt.createProcessor();
xslProc.input = xmlDoc;

function defirition(arg,lang,option) {
  | xslProc.addParameter("langue",lang);
  | xslProc.transform();
  | xslProc.addParameter("terme",arg);
  | xslProc.transform();
  | |   xslProc.addParameter("ident","");
  | xslProc.transform();
  | |     |
 zoneDefinition.innerHTM_=xslProc.output;

}
```

Tableau 4. Script d'interface transmettant la valeur des paramètres de la requête à la feuille de style

Les instructions de traitement des chaînes de caractères de XPATH permettent d'identifier l'élément concerné, un exemple est indiqué ci-dessous. L'expression présentée, applique un modèle (templates) à toutes les expressions possédant l'attribut langue (@langue) = $langue) et qui sont égales au terme ($terme).

```
<xsl:apply-templates
select="//expressions[expression[@langue=$langue]=$terme]"/>
```

Le contenu de l'entrée est affiché et les termes associés (renvois) sont repérés et marqués, grâce à la correspondance *idref* = *id,* leur définition complète est affichée sur demande de l'utilisateur, l'image illustrant le terme s'affiche lorsqu'elle existe, la demande étant traitée comme une nouvelle requête suivant la même méthode. La Figure 2 présente le processus général de création et d'exploitation du dictionnaire électronique.

```
xmlns:xdt="http://www.w3.org/2005/xpath-datatypes">
<xsl:output method="html" version="1.0"
 encoding="ISO-8859-1" indent="yes"/>
<!--Liste des paramètres transmis par javasscript-->
<xsl:param name="terme"/>
<xsl:param name="langue"/>
<xsl:param name="iopt"/>
<!--==================================-->
<xsl:template match="/">
<xsl:apply-templates/>
<!--Intégration des compléments de styles-->
</xsl:template>
<xsl:include href="dictionnaire3.xsl"/>
</xsl:stylesheet>
```

Tableau 5. Feuille de style. Recueil de la valeur des paramètres transmise par la fonction definition ()

4.3. Diffusion du dictionnaire

La méthode est utilisée pour représenter plusieurs dictionnaires de spécialité. La Figure 2 présente le processus général de création et d'exploitation du dictionnaire électronique.

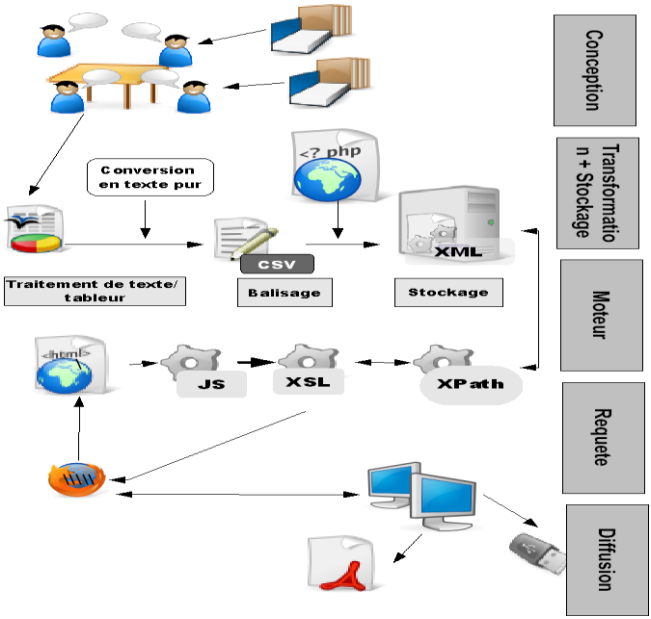


Figure 2. Processus général de production et de diffusion du dictionnaire

4.4. Facilité de partage et d'association, le dictionnaire virtuel

Une des bénéfices les plus fructueux de XML réside dans la possibilité de constituer un document unique à partir de plusieurs documents épars obéissant à la même DTD. Ceci permet par exemple de créer un dictionnaire virtuel multilingue par assemblage de dictionnaires ou de lexiques monolingues. Pour cela il faut créer un document XML, qui devient l'élément père et dont chaque fils est l'un des dictionnaires, ceci est matérialisé par une nouvelle DTD (Tableau 6) invoquée dans le document XML virtuel (Tableau 7). Le schéma d'ensemble de la méthode est représenté sur la Figure 3.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<?xml-stylesheet type="text/xsl" href="dicitionnaire.xsl"?>
<!DOCTYPE  dictionnaireVirtuel SYSTEM  "dictionnaireVirtuelDt
<dictionnaireVirtuel>
&dictionnaire_1;
&dictionnaire_2;
&dictionnaire_3;
</dictionnaireVirtuel>
```

Tableau 6. DTD du dictionnaire virtuel.

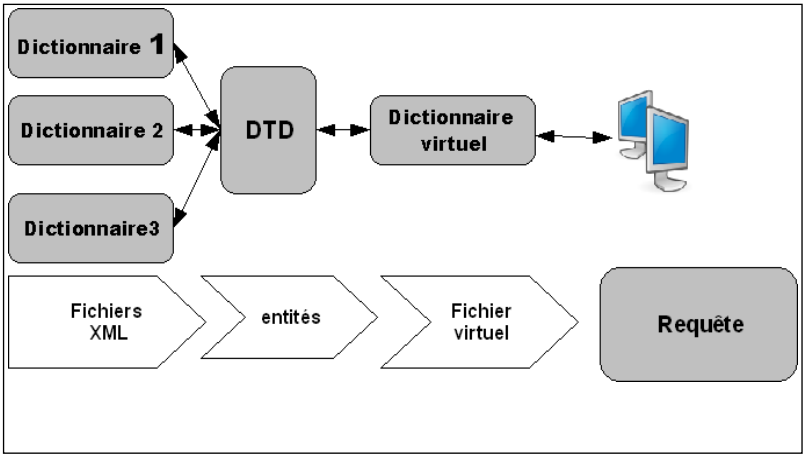


Figure 3. Processus de création d'un dictionnaire virtuel à partir de dictionnaires élémentaires

5. Extension des applications

5.1. Recherche d'associations de termes

La structuration du dictionnaire suivant un schéma XML, en dehors de la publication du contenu sous la forme que nous venons de présenter, permet d'en exploiter le contenu pour rechercher des associations et des groupements de termes, soit définis a priori (renvois), soit implicitement. Pour cela nous avons rédigé une feuille de style XSL, qui par la simple utilisation de *templates* permet d'extraire des informations souhaitées. Nous avons réalisé un test qui a porté sur l'entrée *système de sauvegarde* et sur le terme isolé *sauvegarde* considéré comme une chaîne de caractères, la notion de *système de sauvegarde* ayant été analysée par Marie Calberg-Challot (Calberg-Challot *et al.* 2008a) pour en dégager l'aspect conceptuel.

5.2. Termes explicitement associés

Nous avons explicité plusieurs couches de termes associés en cascade (renvois dans les renvois...), en exploitant les éléments <*ref*> par le biais de l'attribut *idref* . Ceci permet par exemple d'observer le maillage formé autour de l'expression *système de sauvegarde* et de vérifier si les termes associés peuvent relever du concept. Nous nous sommes limités à trois couches successives, le résultat est indiqué dans le Tableau 8. Ceci révèle par exemple les termes *source froide*, *chaleur résiduelle, sous critique* qui ne sont pas des objets relevant des *systèmes de sauvegarde* mais qui cependant en conditionnent fortement la conception et le dimensionnement, le terme

chaleur résiduelle par exemple est mentionné dans le graphe du réseau conceptuel de «système».

1 - accident hors dimensionnement
2 conditions de fonctionnement complémentaires
3 arrêt d'urgence
3 générateur de vapeur
3 source froide
1 - fonction de sauvegarde
2 arrêt d'urgence
3 système d'arrêt d'urgence
3 système de protection
2 arrêt sûr
3 chaleur résiduelle
3 sous-critique
3 tranche nucléaire
2 bâtiment réacteur,
2 chaudière nucléaire

Tableau 8. Liste des 3 couches de renvois successifs dans l'expression système de sauvegarde

5.3. Associations implicites

En effectuant une recherche à partir de la chaîne de caractères "*sauvegarde*" nous étendons le maillage du document, le résultat est illustré dans le Tableau 9. Les entrées mentionnées contiennent au moins une occurrence de la chaîne de caractères dans l'un quelconque des éléments fils. Nous retrouvons dans ce tableau des termes déjà mentionnés dans la première liste, ainsi que certains termes relevant du concept *système de sauvegarde* mentionnés dans la publication en référence tels que : *aspersion d'enceinte*, *système d'injection*, *système de refroidissement, alimentation de secours*. Une analyse plus détaillée est cependant nécessaire pour valider les relations qui existent avec le concept *système de sauvegarde* et les expliciter, le terme *période longue* par exemple fait bien référence à la *fonction de sauvegarde***.** La méthode permet de clarifier a posteriori la cohérence du contenu du dictionnaire, le terme *sauvegarde* est lié aux termes *accident*, *protection, refroidissement*, *secours*, *enceinte de confinement*, *sécurité*, *auxiliaires, injection*. Les *systèmes* qui apparaissent dans la liste concordent avec ceux qui apparaissent dans le graphe de Marie Calberg-Challot. L'analyse du terme *système de traitement des effluents primaires* montre une concordance partielle avec le réseau conceptuel, dans lequel le

terme *bore* par exemple n'est pas mentionné, bien que cet élément utilisé dans le circuit primaire comme absorbant de neutrons, contribue fortement à produire des effluents liquides.

Les deux méthodes s'apparentent plus à une extraction de termes dans un document qu'à la recherche de sens et de liens, elles montrent cependant le parti que l'on peut tirer d'une structuration avec XML qui permet la fouille très fine du contenu et par exemple de dégager un réseau de mots dont certains peuvent après analyse, se référer à un concept identique.

cident - accident de dimensionnement - accident enveloppe de perte de réfrigérant primaire - actionneur de sauvegarde - automate de protection - auxiliaires de sauvegarde - auxiliaires électriques - auxiliaires secourus - bâtiment des auxiliaires de sauvegarde - bâtiment des auxiliaires nucléaires - circuit d'eau brute secourue - classe de sûreté 1E - fonction de sauvegarde - période longue	tème d'alimentation de secours des générateurs de vapeur tème d'aspersion de l'enceinte de confinement - système d'eau brute secourue - système d'injection de sécurité - système de contrôle volumétrique et chimique - système de protection - système de protection intégré numérique -système de refroidissement intermédiaire - système de sauvegarde - système supports - train de sauvegarde - unités logiques de sauvegarde - voies de sauvegarde redondantes

Tableau 9. Liste des entrées extraites dans l'ensemble du dictionnaire contenant au moins une fois la chaîne de caractère "sauvegarde"

6. Bénéfices d'XML

Nous pouvons souligner quelques-uns des avantages majeurs qui à nos yeux procurent une grande facilité d'installation dans un système industriel. La construction d'un dictionnaire fondée sur XML peut être réalisée au fur et à mesure de l'avancement des travaux d'élaboration du dictionnaire, directement dans les groupes de travail.

Les avantages procurés par cette méthode sont donc nombreux :

- développements informatiques légers et autonomes ;
- portabilité ;
    - le dictionnaire peut être transporté par des moyens légers (ordinateurs portables, clés usb, Cd Roms),
    - possibilité d'extension et d'intégration dans des systèmes d'information plus étendus,

- flexibilité et évolutivité,
    - les modifications ultérieures ne nécessitent pas de revoir la forme,
    - facilité de maintenance,
- pérennité des supports ;
    - le format texte est universel,
    - la forme peut être adaptée et évoluer suivant les besoins sans avoir à modifier le contenu,
- mise en commun de dictionnaires complémentaires par exemple dans un univers professionnel éclaté ou multilingue ;
- exploitation ultérieure en utilisant la structure du document ;
    - recherche de similitudes,
    - aide à la recherche de concepts.

7. Conclusions

Le langage XML et ses langages associés permettent de représenter et de partager simplement et sans outils sophistiqués, le contenu d'un dictionnaire. Le modèle de document NucSTML reproduit la structure d'un dictionnaire de spécialité utilisable par tous les protagonistes de l'industrie nucléaire. Un bloc note et un navigateur permettent d'en stocker et d'en diffuser le contenu sur tout système portable notamment, de plus XML autorise la mutualisation de ressources linguistiques éparses, sans obligation de centralisation, permettant ainsi à différents groupes linguistiques appartenant à la même communauté technique de construire séparément un dictionnaire de spécialité monolingue et de créer un dictionnaire virtuel regroupant chacun des dictionnaires au moment de la requête sur le poste client. Un dictionnaire conçu suivant les principes XML rend possible son intégration dans des systèmes de GED ou de KM. Le choix initial de la structure du document et la construction de groupes reconnus par des attributs spécifiques devraient permettre d'aller au delà d'une simple représentation séquentielle de mots et faciliter des regroupements ou des associations de mots.

Bibliographie

Barraclough I. (2006) : *IAEA safety glossary. Terminology used in nuclear, Radiation, radioactive waste and transport safety*, Version 2.0., Departement of nuclear safety and security, IAEA, Vienna, Austria

[*Lexique nucléaire de A à Z. http://www.areva-np.com/scripts/info/publigen/content/templates/show.asp?P=882&L=FR*

Bernard Pascale, Dendien Jacques, Pierrel Jean-Marie *A computerized dictionary : Le Trésor de la langue française informatisé (TLFI).* ATILF. UMR 7118-CNRS/Université Nancy

Calberg M. Dumont X. (2004) : *Projet du dictionnaire de l'ingénierie nucléaire*, Note interne, Areva-NP

Calberg-Challot Marie, Candel Danielle, Roche Christophe (2008a) : *De la variation des usages au consensus terminologique : vers un dictionnaire de l'ingénierie nucléaire*, Conférences Toth. Annecy, 5-6 juin 2008

Calberg-Challot Marie, Danielle Candel, Didier Bourigault, Xavier Dumont, et Jacques Joseph (2008b) : *Une analyse méthodique pour l'extraction terminologique dans le domaine du nucléaire*, Terminology 14 :2, pp. 183-203

Dictionnaire des sciences et techniques nucléaires (1975) : *3ème édition.* Eyrolles. Paris

http://www-cadarache.cea.fr/fr/accueil/lexique.php

Santé canada : Glossaire. http://www.hc-sc.gc.ca/ed-ud/event-incident/radiolog/info/glossary-glossaire-fra.php

Coppolani Pierre, Hassenboehler Nathalie, Joseph Jacques, Pétetrot Jean-François, Py Jean-Pierre, Zampa Jean-Sébastien (2004) : *La chaudière des réacteurs à eau sous pression*, Collection Génie Atomique, Instn, EDP Sciences

Méta modèle de Dictionnaire Standard de la Construction. http://www.edibatec.org/Accueil/SDC/ProjetSDC.htm

Elchacar Mireille *Le projet "FRANQUS" : la langue française vue par le Québec*, Projet de Thèse de doctorat, Université de Sherbrooke, Québec, Ca

Fabien Gandon (2002) : *Ontology Engineering a Survey and a Return on Experience. Rapport de Recherche*, Ch 4. Pg 61. ACACIA team. N° 4396

Gossens Michel (1999) : *XML et XSL : un nouveau départ pour le web*. Cahiers GUTenberg n° 33-34 _ Congrés GUT99-Journée XML

Henriet Laurent, Fagnent Sylvain (2006) : S*tandardisation des échanges : Mise en œuvre d'un dictionnaire bancaire d'interopérabilité*, Revue Banque, N° 679

Hydroquébec, Glossaire des termes nucléaires, *http://www.hydroquebec.com/fr/*

XML, Backgrounder. *Technology and applications.*

Software AG (2000) : The XML Company, Darmstatt, De

Maks Isa, Tiberius Carole, van Veenendaal Remco. *Standardising bilingual lexical resources according to the Lexicon Markup Framework*, Dutch HLT agency (TST-centrale), Institute for Dutch Lexicology (INL), Leiden, Nederlands

Mangeot-Nagata Mathieu, Bilac Slaven (2004) : *Construction collaborative d'un dictionnaire multilingue*, *Le projet Papillon.* Actes de JSF 2003, National Olympic Memorial Youth Center, Tokyo, Japon

Manuélian Hélène et Schang Emmanuel (2007) : *XML, DTD et TEI pour un dictionnaire étymologique des créoles*, Universités de Cergy Pontoise et Orléans

Marcoux *et al.* (2001) : *XML en route au gouvernement du Québec*, Groupe de Recherche sur les documents structurés, École de Bibliothéconomie et des Sciences de l'Information, Université de Montréal, Rapport Final

Glossary http://www.nrc.gov/reading-rm/basic-ref/glossary.HTML

Rey Christophe, Zaouy Corinne (2004) : *Balisage XML "ciblé " : une nouvelle approche dans l'informatisation des corpus*, Université de Provence, Colloque international sur la fouille des textes, La Rochelle

Bigot B., Santarini G. *et al.* (2008) : *Dictionnaire des sciences et techniques nucléaires, 4e édition*, OmniSciences

Text Encoding Initiative, *The XML version of the TEI Guideline*, Chapitre 13, The terminological database

A propos des auteurs

Jacques Joseph

45 rue de Saint Cyr
69009 Lyon
jacquespjoseph@aol.com

Mémoire du Club informatique des grandes entreprises françaises (CIGREF)[1] Nouveau plan de classement

M.-P. Lacroix, J.-Y. Gresser

Résumé : Mettre l'utilisateur au cœur des systèmes d'information est un des grands défis de notre époque. Dans le cadre d'un double exercice historique et prospectif, le CIGREF revoit les modes de diffusion de ses documents. Sa longue histoire posait le problème d'accès à un fonds dont les thématiques ont évolué, tout comme les concepts de l'informatique. L'objectif était d'obtenir un plan de classement qui puisse :

- faciliter cet accès sur l'ensemble du fonds disponible ;
- évoluer en fonction des thèmes abordés et de leur dénomination.

Nous disposions en préalable d'une analyse linguistique dont l'objet avait été de dégager l'évolution de la thématique des travaux sur une longue période. À partir de tableaux chronologiques et/ou thématiques détaillés, nous avons recherché une expression des thèmes plus schématique, mais non simpliste. Cette expression, nous l'avons trouvée à partir d'une modélisation sémantique. Ce résultat illustre l'intérêt de ce type de modélisation.

Un autre intérêt, confirmé par d'autres cas, est la valeur pédagogique ou communicationnelle des diagrammes produits.

Le souci de garder l'utilisateur au centre de la démarche était essentiel. La compréhension des concepts sous-jacents de l'outil utilisé pour le nouveau système d'information et l'acceptation du classement proposé ont été déterminants. La démarche se prolonge actuellement par une analyse plus fine de l'usage effectif du fonds pour en améliorer la diffusion.

Mots-clés : modélisation, classement, diffusion, circulation, usage, information, informatique, profil utilisateur, sémantique, ontologie

1 Le Club informatique des grandes entreprises françaises a été fondé en 1970. Sa mission est de « promouvoir l'usage des sytèmes d'information comme facteur de création de valeur et source d'innovation dans l'entreprise ». Environ 140 entreprises « utilisatrices de l'informatique » en sont membres. Elles sont représentées par leur directeur des systèmes d'information (DSI) ou équivalent.

1. Introduction

Un plan de classement est "une organisation structurée et hiérarchique[2] d'un ensemble de concepts ou d'objets. C'est un outil intellectuel qui permet aux documents et dossiers de trouver logiquement leur place les uns par rapport aux autres" (MoReq2 2008). Le plan de classement est le point de départ de nombreuses problématiques d'entreprise : documentation, archivage, logistique, nomenclature commerciale etc.

En recherche documentaire, un plan de classement peut être concurrencé par les outils de recherche en texte intégral mais il reste indispensable à la visibilité des thématiques de travail sur le court, moyen et long terme.

Le fonds propre du CIGREF comprend un nombre de documents plutôt réduit mais ce nombre même n'était pas précisément connu, faute d'une bonne classification.

Le premier objectif était d'en rendre l'accès et l'alimentation simples et intutitifs pour les permanents du CIGREF comme pour les membres, les partenaires et le grand public, à savoir :

1. disposer d'un accès facile à l'ensemble du fonds disponible (papier, électronique et certains états du site internet ou de l'intranet) ;
2. de pouvoir évoluer en fonction des thèmes abordés et de leur dénomination.

La problématique du CIGREF participait à la fois de la documentation et de l'archivage.

Il existe pour la documentation ou la gestion de l'information des nomenclatures détaillées. Les plus connues sont celles de grandes bibliothèques (DEWEY, JITA). Ces dernières sont adaptée au classement d'un savoir encyclopédique. Leurs rubriques gardent un certain niveau de généralité et d'universalisme, vite dépassé dans un contexte plus spécialisé.

L'examen de plans propres à l'informatique, comme celui du MISQ[3] montre que ceux-ci correspondent d'abord à la recherche, la conception ou la production d'artefacts matériels ou immatériels. Ces plans n'abordent

2 L'aspect hiérarchique est jugé souhaitable pour l'archivage mais il n'est pas toujours vérifié pour d'autres usages.

3 Managemnt of Information Systems Quartely : "a peer reviewed scholarly journal. The editorial objective ... is the enhancement and communication of knowledge concerning the development of IT-based services, the management of IT resources, and the use, impact, and economics of IT with managerial, organizational, and societal implications".

pas la problématique des usages, alors que celle-ci est aujourd'hui au cœur des préoccupations de tous les acteurs de l'informatique.

Dans le domaine de l'archivage, des exigences ont été formulées sur les caractéristiques formelles ou fonctionnelles d'un plan de classement (MoReq2 2008) mais ces exigences ne disent pas si un plan de classement orienté activités est meilleur qu'un plan de classement orienté métiers ou savoir-faire ou même nature de document.

En bref, il n'existait pas de plan tout prêt dont nous pouvions réutiliser les rubriques. Par contre, nous pouvions continuer à nous inspirer des plans existants et surtout de principes formels existants.

Dernier aspect et non des moindres, les travaux du CIGREF sont autant prospectifs qu'historiques. Il fallait pouvoir imaginer les rubriques de sujets encore peu ou pas traités.

2. Les "hiérarchies" existantes

Nous incluons dans ces hiérarchies :

- un comptage des documents selon leur "type", fait sur le fonds disponible;

- un recensement des activités du CIGREF depuis 1998 ;

- un document d'analyse linguistique des rapports d'activités du CIGREF, effectuée par la société Anacom et couvrant la période 1974 à 2001.

2.1. Typologie des documents

Le tableau 1 liste les catégories existant dans l'ancien système.

	Nombre de documents
Documents de référence (incluant les catégories suivantes)	9556
Actualité presse- (liens)	19
Compte-rendu (de réunion)	7
Contact ?	0 ?
Discours (texte ou vidéo)	?
Document de travail	51
Intervention à l'extérieur	45
Notes	7
Présentation (support de)	12
Rapports (publication) CIGREF dont rapport d'activité.	Non pertinet
Ressource de l'internet (URL)	1?
Autre	0

Tableau 1. "Type" des documents du Cigref & comptage

En vigueur au début notre exercice, cette catégorisation a été rarement appliquée.

Le caractère non-discriminant des catégories est frappant. Un plan de classement thématique par activité ou par sujet traité apporte davantage de pertinence.

2.2. Organisation des travaux des membres ou autres contributeurs

Le tableau 2 est construit à partir :

1. d'une typologie organisationnelle en deux catégories- comité de pilotage devenu comité d'orientation et groupe d'activités ;
2. de la mise en correspondance des groupes d'activités avec les regroupements thématiques de l'intranet du CIGREF (accessible aux seuls membres).

Assez curieusement, la seule structure pérenne - le cercle des DSI - ne correspondait à aucune rubrique thématique de l'intranet. Il fallait en tenir compte pour le futur.

Le tableau correspond à la période 2004 - 2008, donc à l'histoire récente. Les flèches montrent comment le regroupement des thèmes peut varier au fil du temps, en fonction de l'organisation des activités et de la communication. Sur la période considérée, quelques thèmes ont perdu de l'importance mais l'ensemble montre une grande continuité.

Certains thèmes sont horizontaux. D'autres apparaissent "hors classement", d'autres encore comme les partenariats sont hors périmètre

de l'intranet. Il semblait important de les faire émerger dans le nouveau plan.

Comité de pilotage (avant)	**Comité d'Orientation (auj)**	**Groupes d'activités**	**Regroupement thématique de l'intranet**
1 DSI, stratégie, métiers de l'Entreprise.	1 Le SI au service des métiers et de la Direction génerale	Contrôle interne et SI Dialogue stratégique SI Pilotage de la stratégie SI	Gouvernance du SI Capital immatériel Intelligence économique et stratégique (voir 6)
2 Management de la Dsi et des métiers	2 La performance durable des SI	Poste de travail – évolution Open source Web 2.0 en entreprise Telecom et infrastructures Protection de l'information Club achats	Urbanisme et architecture
3 Urbanisme, architecture technologies	3 Le management de la fonction SI	RH – scénarios d'évolution des besoins en compétences, indicateurs Le SI pour la DSI	Finance, contrôle de gestion, coûts de l'informatique RH - Compétences
4 Télécommunications			
5 Relations avec les fournisseurs et autorités régulatrices	Écosystème informatique	(associations, fournisseurs et leurs syndicats etc.)	
6 Entreprise et société de l'information			
	Cercles (ouverts à non CIGREF)		
			La recherche au CIGREF
	…	Partenaires	Travail direct des permanents dans l'optique d'être utile aux membres, souci d'influence

Tableau 2. Comités, groupes de travail & thématiques de l'intranet du CIGREF

2.3. Analyse des rapports d'actvité

En 2002, la société Anacom a fourni au CIGREF une analyse sur les sujets abordés dans les rapports d'activités ainsi que sur les descripteurs de ces activités pour la période allant de 1975 à 2000 (Anacom 2002).

Le tableau 3 est tiré de l'annexe du rapport d'Anacom. Anacom y regroupait l'ensemble des thèmes abordés en 25 thèmes majeurs. Ont été rajoutées au tableau original les périodes correspondant aux thèmes abordés.

L'histoire "ancienne" du CIGREF montre l'accent mis sur les thèmes techniques. Cette histoire devait être prise en compte même si les thèmes dominants aujourd'hui sont la gouvernance et les usages.

Thèmes (grandes familles)	**Sujets abordés (exemples)**	**Années**[4]
Sécurité	Sécurité des réseaux, responsables de sécurité informatiques.	1987- 2001
Relations institutionnelles	Relations avec le Ministère de l'industrie, avec l'AFNOR, avec les instances européennes.	1975 - 2001
Qualité	Club des rsponsables Qualité, certification ISO 9000.	1983 – 1998
Réseaux	Architecture de réseaux, Internet, Intranet, réseaux haut débit *(déjà !)*.	1975 – 2001
Normalisation / Réglementation	Politique nationale, projet ISOTOP.	1977 puis 1984 – 2001
Exploitation	Gestion du parc informatique	1975 – 1991 puis 1994
Bureautique	Micro-informatique, logiciels.	1975 - 1993
Contrôle de gestion	Evaluation et maîtrise des coûts liés (?) au servie informatique.	1979 – 1991 puis 1993 – 1997, 1999
La fonction DSI	Club des DSI, marketing de l'informatique.	1975 – 1981, 1987 – 1988, 1992 – 1994, 1998 - 2001
Management de l'entreprise	Processus de décision, travail en groupe, accompagnement du changement, gestion de la relation client (de la DSI ?).	1978 – 1981, 1984 -1987
Commerce / e-commerce	Transactions automatisées, transactions à distance, bases de données commerciales, commerce international.	Pratiquement continu de 1975 à 2001
Veille / prospective		1981 – 1984, 1987, 1991 – 1994, 1999, 2001
Passage à l'an 2000		1997 - 2001
Technologies spécifiques	Vidéotex, traduction assistée par ordinateur, fibres optiques, faisceaux hertziens.	1983 - 1998
Problèmes juridiques	Signature électronique, droit de la preuve.	1977 – 1983, 1993

Tableau 3. Thématiques des rapports d'activités

Dans le rapport précité, Anacom se livre aussi à une analyse des activités du CIGREF sur la période considérée. L'apparition, le renommage, l'évolution des thèmes nous ont paru pertinents pour notre démarche. Notre lecture "historique" est résumée dans les tableaux 4 et 5. Les mots ou expressions sont celles tirées par Anacom des rubriques des rapports d'activités.

On notera l'évolution des désignations comme :

1. *Relations extérieures - Relations avec les fournisseurs - Relation avec l'offre-*

4 L'année du Cigref est à cheval sur deux années calendaires. Par convention nous retenons la deuxième année. NB Ce tableau est un extrait du tableau complet qui liste les 25 thèmes.

Relation avec l'offre en informatique, ou

2. *Préoccupations de gestion - Problèmes de gestion - Gestion du SI de l'entreprise - Stratégie et économie du SI.*

Les **aspect relationnels** sont plus explicites à partir de 1985, soit dix ans après la création du Club. Ils deviennent de plus en plus importants au fil du temps de même que les **considérations de gestion puis d'usage de l'informatique**.

Période	**Activité 1**	**Activité 2**	**Activité 3**	**Activité 4**	**Activité 5**
1975-1982			Relations extérieures	Préoccupations de gestion	
1983 - 1984			Relations extérieures	Problèmes de gestion	
1985			Relations extérieures		
1986 – 1988			Relations extérieures	Gestion du SI de l'entreprise	
1989 – 1990			Relations avec les fournisseurs	Stratégie et économie du SI	
1991 – 1993			Relation avec l'offre		
1994 – 1995					Contrôle de gestion et maîtrise des coûts
1996			Relation avec l'offre		Métrique de la rentabilité, contribution de la DSI à la performance de l'entreprise
1997	Opération An 2000	Passage à la monnaie unique	Relation avec l'offre en informatique		
1998	Opération An 2000	Passage à la monnaie unique	Relation avec l'offre en informatique		Club Benchmarcking
1999	Opération An 2000	Passage à la monnaie unique			Coût de possession informatique
2000	Opération An 2000	Passage à la monnaie unique			

Tableau 4. Chronologie des activités du CIGREF (activités 1 à 5)

Période	**Activité 6**	**Activité 7**	**Activité 8**	**Activité 9**			
1985	Groupes des responsables	*Évolution de l'informatique dans l'entreprise (Gestion du parc, formation, exploitation…)*		Groupes des responsables			
1986 – 1988		Évolution de la fonction informatique					
1989 – 1990			GRH et formation				
1991 – 1993*		Évolution de la fonction informatique	GRH et formation				
1994 – 1995	Eclairage européen,CI-group		Evolution de la fonction RH	Eclairage européen,CI-group			
1996	Groupes Clubs,	CI-group	RH, mobilité et formation des professionnels I&T	Groupes Clubs,	CI-group		
1997			RH	Groupes Clubs,	CI-group		
1998			RH	Dialogue avec les acteurs	CI-group		
1999			Observatoire des RH	Marketing de l'informatique auprès des décideurs	Un dialogue permanent avec les acteurs	CI-group	
2000			Formation de base à l'informatique	*Observatoire des RH[4]*	Dialogue permanent avec les acteurs	CI-group	

Tableau 5. Chronologie des activités du CIGREF (activités 6 à 9)

2.4. Apport des analyses antérieures

Ces tableaux confirment l'inadéquation pour le CIGREF, et finalement pour le positionnement de l'informatique dans les entreprises et les administrations, des classifications exclusivement fondées sur les techniques.

Quant aux rubriques elles-mêmes, outre les aspects lexicaux, les questions essentielles avaient trait à la genèse ou à la pérennité des thématiques, ainsi

qu'aux recouvrements ou convergences éventuelles. C'est sur ces derniers aspects que nous attendions beaucoup d'une approche "ontologique".

3. Modèles sémantiques

Parallèlement au recensement des classifications existantes, nous avons imaginé plusieurs plans de classement en allant du général au particulier, et en nous limitant à 4 niveaux.

La pertinence des rubriques était validée au fur et à mesure de la construction des modèles sur un échantillon d'une 100e de document.

La première esquisse de classement s'appuyait sur les quatre catégories d'un métamodèle général : acteurs, usages, outils, ressources.

La deuxième esquisse avait été imaginée à partir de l'application récursive d'un métamodèle classique des grandes fonctions de l'entreprise, développé pour :

1. la DSI en tant qu'entreprise, et
2. l'entreprise elle-même.

La simplicité du principe était attirante mais la multiplication mécanique des rubriques aboutit à un schéma peu "parlant" et parfois ambigu. La première esquisse se situant davantage dans la continuité des analyses antérieures nous l'avons retenue comme hypothèse de travail.

3.1. Modélisation de l'outil du système d'information & premier essai de modélisation thématique.

Le premier modèle[5] est une transposition de l'option "navigation" de l'outil du système d'information (SI). La figure 1 est une "sémasiologie" rapide et partielle du site d'accueil de Nëlis.

5 Tous les modèles présentés ont été construits avec SNCW (Semantic Network Craft Workbenchmarch) d'Ontologos corp.

Elle est fondée sur les concepts d'espace, de catégorie, de thème et de document, complétés de quelques unes des instances de documents.

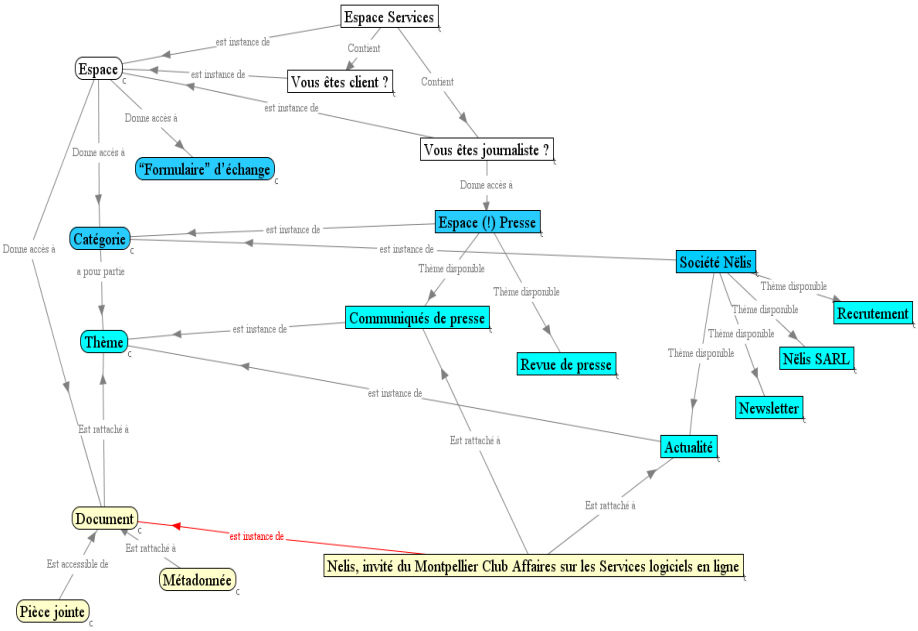


Figure 1. Concepts de haut niveau de l'outil du SI

Il semble que la notion d'espace soit "récursive". L'emploi du mot "espace" pour une catégorie nous a d'abord paru qualifier une "vue", plutôt qu'un concept particulier dans un schéma.

Après la première mise en place du nouveau plan, la notion est ressurgie. Elle nous a permis d'étendre le plan dans le sens d'une prise en compte différenciée des besoins des destinataires des documents.

Telle qu'apparaissant sur le site, la relation entre catégories et thèmes est hiérarchique.

Plusieurs espaces peuvent accéder à une même catégorie. Ceci rend possible un plan de classement hiérarchique à deux niveaux.

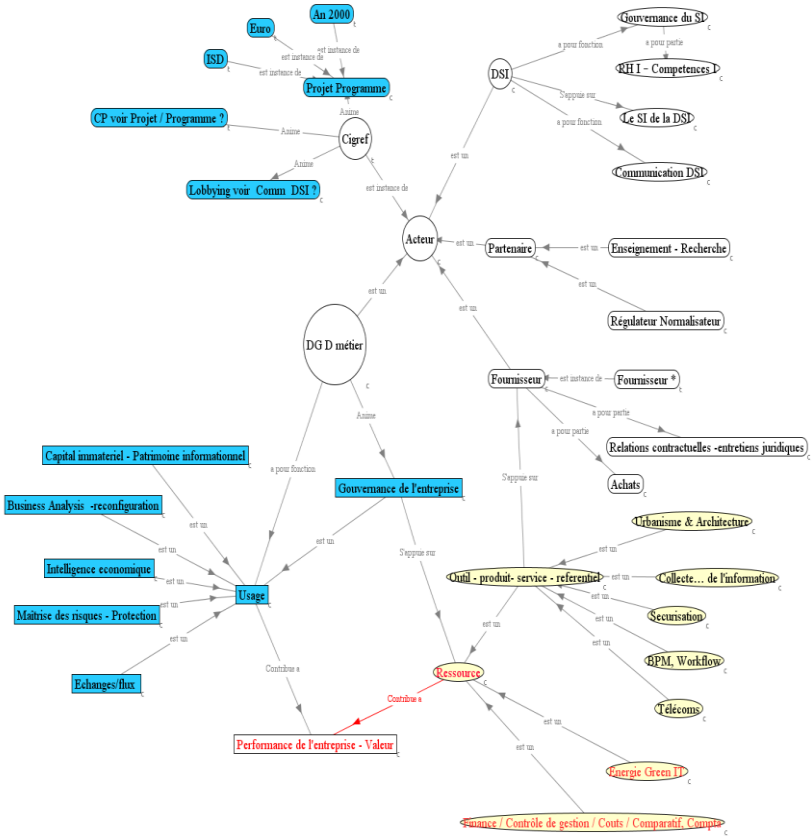


Figure 2. Modélisation des thématiques (premier essai)

Notre premier essai de modélisation des thématiques (Figure 2) a été construit dans un formalisme qui permettait d'exprimer les relations entre les "thèmes"ou "catégories". Nous n'avons pas tenu compte des "espaces" et nous avons procédé à quelques simplifications.

Certaines expressions des rapports d'activités se retrouvent dans ce modèle. C'est normal. Nous avons essayé de garder l'expression la plus actuelle tout en évitant les effets de mode trop voyants.

L'aspect essentiel de ce modèle est la symétrie qui se dégage entre les grandes catégories liées aux usages (Figure 3) et celles liées aux technologies ou ressources (Figure 4) se dégage à travers un examen plus attentif, ainsi :

1. Capital immmatiel – patrimoine informationnel face à Finance/ Contrôle de gestion/ Coûts/ Comparatif/ Comptabilité,

2. "Business analysis" - reconfigutaion face à BPM (Business Process Management), "Workflow",
3. Échanges/flux face à "Télécoms" etc.
4. Nous en avons tenu compte pour simplifier le modèle. Cette simplification était d'ailleurs indispensable pour la mémorisation des thématiques par les contributeurs et les utilisateurs.

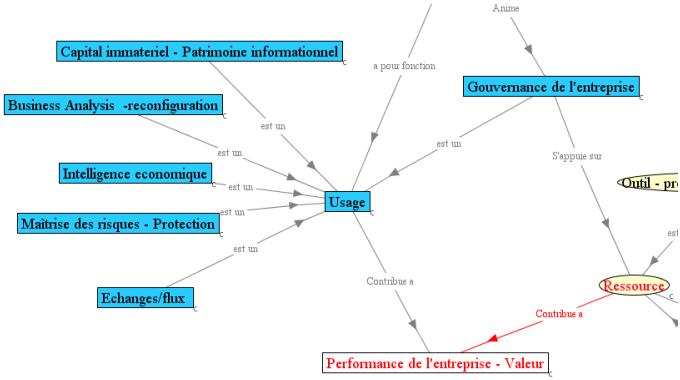


Figure 3. Usages

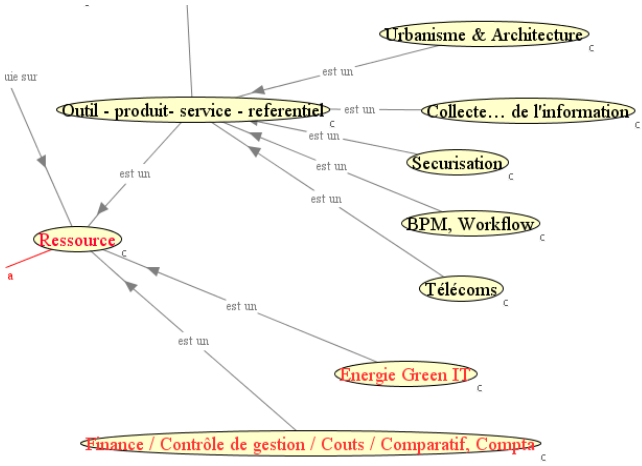


Figure 4. Outils (au sens large) – Ressources

3.2. Deuxième essai de modélisation thématique

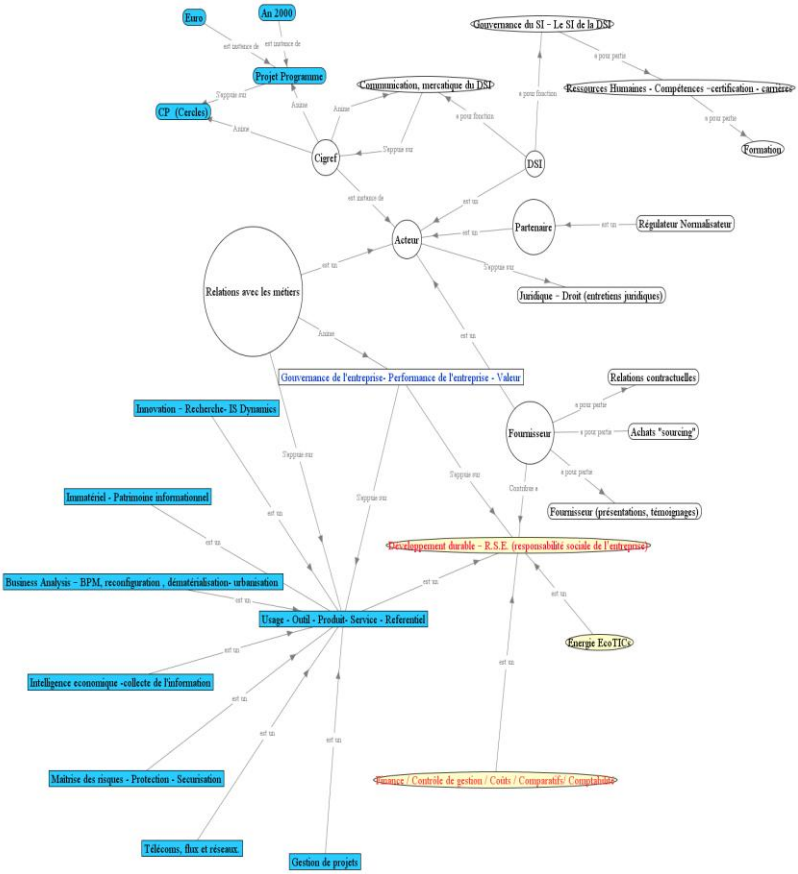


Figure 5. Deuxième modèle sémantique

Ce modèle (Figure 5) a le mérite de regrouper les documents traitant de situations, de problèmes et de leurs solutions dans un périmètre commun.

Le regroupement entre "usage etc." et "technologie etc." impliquait :

- la définition d'un concept de niveau supérieur. Nous l'avons fait figuré dans le diagramme sans vraiment le nommer,

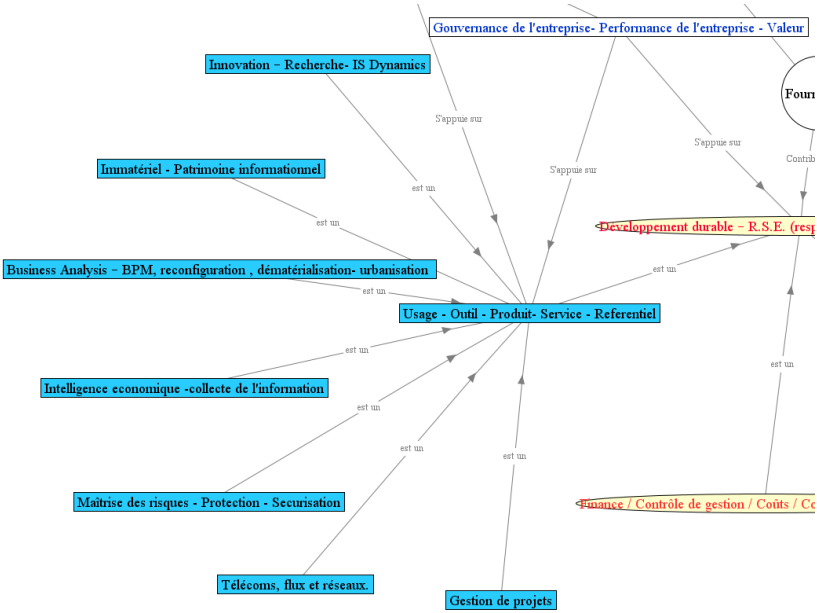


Fig. 3.6 Deuxième modèle sémantique (détail)

- le tri entre deux types de notions, "nature" et "type" (voir section Mise en œuvre). Attributs du concept document ou caractéristiques des documents en tant qu'objets. Ces notions étant largement solubles dans l'outil du SI, nous n'avons pas tranché.

4. Mise en œuvre

Le plan retenu s'écarte légèrement du schéma "théorique" : certains thèmes y sont regroupés, d'autres éclatés. La distinction entre les concepts initiaux et certaines caractéristiques ou attributs perd sa signification au niveau de l'interface du système d'information, fondé sur des menus déroulants ou des listes.

Un document peut être rattaché à plusieurs thèmes, ce qui a un sens d'un point de vue documentaire mais guère au niveau des pratiques actuelles de l'archivage. Mais l'archivage n'était pas la préoccupation essentielle, ne serait qu'au vu de la taille du fonds.

Thématiques selon le 2e schéma	**Thématiques retenues & mises en oeuvre**
Business Analysis – BPM, reconfiguration - urbanisation - dématérialisation	Accompagnement du changement
Business Analysis – BPM, reconfiguration - urbanisation - dématérialisation	Architecture d'entreprise - EA
Régulateur Normalisateur	Autorités régulatrices
Finance / Contrôle de gestion / Coûts / Comparatif/ Comptabilité	Contrôle de gestion
Gouvernance de l'entreprise- Performance de l'entreprise - Valeur	Décisionnel
	Déontologie - Ethique
Développement durable – R.S.E. (responsabilité sociale de l'entreprise	Développement durable – R.S.E.
Fournisseur (présentations, témoignages)	Fournisseur (relatif ~a un fournisseur particulier)
Immatériel - Patrimoine informationnel	Gestion des connaissances
Gouvernance du SI – Le SI de la DSI	Gouvernance et stratégie IT
Immatériel - Patrimoine informationnel	Immatériel
Intelligence économique et stratégique	Intelligence économique et stratégique
Communication, mercatique du DSI	Marketing du SI
Gouvernance de l'entreprise- Performance de l'entreprise - Valeur	Performance – Création de valeur
Innovation – Recherche- IS Dynamics	Programme de Recherche (du CIGREF)
Relations avec les métiers (anc.nt DG D métier)	Relations avec les métiers
Ressources Humaines - Compétences – certification - carrières	Ressources Humaines - RH
Maîtrise des risques - Protection - Securisation	Sécurité

Tableau 6. Correspondances entre thématiques et entités.

Indicateurs (en tant qu'information)	Usage - Outil - Produit- Service – Referentiel ou Nomenclature
Normes et référentiels	Usage - Outil - Produit- Service – Referentiel ou Nomenclature
Outil - Méthode	Usage - Outil - Produit- Service – Referentiel ou Nomenclature
Technologies	Usage - Outil - Produit- Service – Referentiel ou Nomenclature
Usages	Usage - Outil - Produit- Service – Referentiel ou Nomenclature

Tableau 7. "Nature" de document > thèmes (de travail)

Les catégories du tableau 7 ont été intégrées à celle du tableau 8.

Nous avions pensé traiter la liste suivante en tant qu'attributs ou caractéristiques des documents.

Accord (Charte…)
Actualité (annonces)
Agenda (CIGREF)
Astuce (à savoir)
Communiqué de Presse
Compte-rendu
Demande DSI
Document associatif
Données CIGREF
Evénement
Informations du bureau
Informations du conseil d'administration
Intervention
Nomination
Partenariat
Parrainage
Présentation
Publication (non CIGREF
Rapport CIGREF
Rencontre – Audition
REX Membre

Video

Tableau 8. "Type" de document ou d'information

C'est à ce stade qu'est réapparue la notion "d'espace". Ces "types" sont propres aux documents à destination de l'extérieur : DSI, collaborateurs des DSI, autres métiers de l'entreprise, partenaires, grand public.

Par ailleurs, le CIGREF comme toute association manipule des documents qui sont propres à sa gestion ou à sa gouvernance et qui n'ont aucune vocation à une large diffusion. Une première liste en a été élaborée et doit être intégrée au schéma initial.

Le travail d'indexation de ce fonds selon les nouveaux thèmes est terminé. En dehors de la probématique exposée dans le paragraphe précédent, la seule modification notable est celle d'une nouvelle catégorie : relations (avec les) opérateurs.

5. Conclusion

Il reste maintenant à tester plus avant le fonctionnement du mode recherche. Au moment du séminaire nous avions trois mois d'expérience.

Autre aspect : dans un domaine comme l'informatique il est normal qu'un plan de classement évolue. Il est toutefois souhaitable qu'il puisse rester stable sur plusieurs années (typiquement 4 à 5 ans). Ceci ne pourra se constater qu'en 2012 ou plus tard.

Avons-nous innové sur le plan scientifique ou méthodologique ? Probablement pas. L'essentiel du résultat est dans le plan final. Ce plan constitue une novation importante par rapport à tous les plans antérieurs dont nous avons eu connaissance.

Certains ont l'habitude de dire que tout est possible en informatique et beaucoup d'approches applicatives ont été fondées sur ce postulat. Notre exemple montre le contraire, mais nous avons su transformer certains contraintes en avantages.

Le choix de ne pas modifier l'outil du SI a été maintenu. Il était structurant, sans doute plus que nous l'avions imaginé au départ. Cet outil est d'abord un outil de gestion relationnelle (CRM) et non de gestion de contenu, ni de production de documents. Au contraire des outils classiques de gestion documentaire ou d'archivage, il nous a forcé, à replacer les « récepteurs » au centre de notre démarche. C'était l'objectif bien compris de notre travail…En cohérence totale avec une stratégie que le CIGREF a commencé à mettre en place il y a vingtaine d'année.

Ce n'est pas fini. Dans le prolongement de ce travail, le CIGREF a lancé une recherche méthodologique pour favoriser l'émergence et la circulation de l'information. L'approche (Petit 2009) porte sur une qualification de l'information en fonction des destinataires éventuels, et des usages (valorisation, partage). Mais ceci mériterait d'autres développements.

Bibliographie

J.-Y. Gresser, *Terminology & Information Science(s)*, ICTTA '08, avril 2008

Dewey, *Dewey®Services at a glance*

"The Dewey Decimal Classification (DDC) system, devised by library pioneer Melvil Dewey in the 1870s and owned by OCLC since 1988, provides a dynamic structure for the organization of library collections. Now in its 22nd edition, and available in print and Web versions, the DDC is the world's most widely used library classification system."

JITA Classification Schema of Library and Information Science, *The JITA Classification Schema*

http://eprints.rclis.org/jita.html

"The JITA Classification Schema has been developed starting from a merger of NewsAgentTopic Classification Scheme (maintained by MikeKeen at Aberystwyth, UK, until 31st March 1998) and the RIS classification scheme of the (now defunct) Review of Information Science originally conceived by Donald Soergel (University of Maryland)."

Gerald kembellec, *Ontologie recherche en informatique,* TOTh 2008

http://geka.ec-10.eu/

Divers, MISQ C -- Information Technology

Anacom, *Analyse linguistique des rapports d'activité du Cigref sur la période 1975-2000*, Rapport d'étude privé, 2002

Évelyne Alliaume et al. IALTA, *Apprivoiser MoReq, Pour archiver et conserver l'information*, FNTC 2007, CR2PA 2008

Commission Européenne, *Model Requirements for the Management of Electronic Records*, 2002, traduit en français en 2004 sous le titre *Modèle d'exigences pour l'organisation de l'archivage électronique* mis à jour en avril 2008 sous le nom de MoReq2.

Henri Barki, Suzanne Rivard, Jean talbot (École des Hautes Études Commerciales, Montréal), A Keyword Classification Scheme for IS Research Literature, MISQ Novembre 1997[1]

Joan M. Reitz, ODLIS (Online Dictionary of Library and Information Science), WCSU Librairies, 2004

Loïc Depecker, Violette Dubois (editors), Terminologie et ontologies, Le savoir des mots, Sociéte Francaise de Terminologie, 2008.

Loïc Depecker, Violette Dubois (editors), Terminologie et sciences de l'information, Le savoir des mots, CIGREF, Sociéte Francaise de Terminologie, Paris 2006, 111 p.

Christophe Roche, Ontology- A survey, University of Savoie, Equipe Condillac - Campus Scientifique, 2003.

Christophe Roche, Terminologie et ontologie. Revue Langages n°157 mars 2005. pp. 48-62. Editions Larousse

Christophe Roche, Dire n'est pas concevoir. 18èmes journées francophones d'Ingénierie des Connaissances. Grenoble 4-6 juillet 2007. pp.157-168

Christophe Roche, Le terme et le concept : fondements d'une ontoterminologie. Conférence TOTh 2007« Terminologie & Ontologie : Théories et Applications ». Annecy 1er juin 2007. pp. 1-22.

Kim H. Veltman, Towards a Semantic Web for Culture, Maastricht McLuhan Institute, 2007

1 http://www.misq.org/roadmap/classscheme/classification.htm#authors

Christophe Roche, Le terme et le concept : fondements d'une ontoterminologie, TOTh 2007 : « Terminologie & Ontologie : Théories et Applications » - Annecy 1er juin 2007

Antonio De Nicola, Michele Missikoff, Roberto Navigli, A software engineering approach to ontology building, Information Systems- Databases : Creation, Management and Utilization, Elsevier, Vol 34, Issue 2, Avril 2009

Bertrand Petit[2], Guide utilisateur pour favoriser l'émergence de l'information pertinente (et favoriser) son routage et son classement, V3 aussi Mise en application du prototype, Essai d'un modèle prototypique d'émergence de l'information pertinente au sein d'un SI et à destination d'un acteur, Documents internes du CIGREF, Avril 2009

A propos des auteurs

M.-P. Lacroix, J.-Y. Gresser

Club informatique des grandes entreprises françaises CIGREF

21, avenue de Messine, Paris 8e

www.cigref.fr

ychause@alum.mit.edu

marie-pierre.lacroix@cigref.fr

2 Bertrand Petit est doctorant en Sciences de gestion, en terrain d'étude au CIGREF – bertrand.petitlouage@laposte.net

Les secteurs d'activité à l'épreuve des discours

Frédéric Erlos

Résumé : Le découpage de l'activité d'un groupe bancaire en secteurs d'activité constitue un point d'accès privilégié pour l'organisation de l'information sur le portail d'un intranet. On propose une méthode d'identification des noms de secteurs d'activité dans les textes d'un corpus de rapports d'activité. Cette méthode s'appuie sur l'observation du fonctionnement discursif de certaines classes de dénominations propres. On accède ainsi aux manières de classer l'activité bancaire telles qu'elles sont offertes à des publics aussi bien internes qu'extérieurs à l'entreprise.

Mots-clés : secteur d'activité, nom propre, référentiel terminologique, terminologie textuelle, linguistique de corpus, textométrie, situation de communication, organisation de l'information

1. Introduction

Depuis leur développement au milieu des années 1990, les intranets se sont progressivement rendus indispensables pour la réalisation de la plupart des activités à l'intérieur d'une organisation. C'est le cas, plus particulièrement, des activités liées au partage et à la diffusion des informations. De ce point de vue, les intranets possèdent des objectifs similaires à ceux des systèmes d'organisation de l'information. Il s'agit, d'une part, de permettre la localisation d'informations à partir de critères caractérisant les supports et leur contenu, et d'autre part, de rendre possible la consultation de ces informations et la navigation au sein du fonds documentaire auquel elles appartiennent. Cependant, on constate que dans le cadre des intranets il est rarement fait usage des outils documentaires habituellement utilisés dans les systèmes d'organisation de l'information, tels que les classifications ou les thésaurus. Par ailleurs, les activités liées au traitement de l'information semblent s'être dissoutes dans les diverses tâches qui incombent quotidiennement aux salariés. La réalisation de tâches telles que l'indexation d'un contenu, la construction de l'arborescence d'un site, ou encore, le regroupement thématique de contenus, s'effectue sans les repères fournis par un usage normalisé et contrôlé du langage. Par ailleurs, les contraintes imposées par l'organisation du travail ne favorisent pas le développement de nouveaux comportements que l'on rencontre sur le Web "gratuit", comme la prise en charge de ce type de tâches par les consommateurs de l'information (folksonomie, indexation sociale). Il résulte de cet état de fait que le vocabulaire et les habitudes de classement des différents publics d'un site sont rarement pris en compte, ce qui constitue un obstacle à la diffusion des informations, et partant, à la bonne réalisation de nombreuses activités au sein des organisations.

Dans un tel contexte, on propose de guider les tâches liées à l'organisation et à la diffusion de l'information en restituant les manières de dire et de classer propres aux publics des sites d'un intranet. Ces "images linguistiques" doivent être organisées rationnellement sous la forme de référentiels terminologiques, de manière à pouvoir faire l'objet d'une exploitation directe par les webmestres ou les contributeurs d'un site. Mais surtout, elles sont destinées à suivre l'évolution des usages linguistiques à l'intérieur d'une organisation. On a développé dans un autre travail les différentes questions soulevées par la mise en place de référentiels terminologiques adaptables aux situations de diffusion de l'information. C'est pourquoi, on se limitera ici à exposer les principales réponses que l'on a proposées. En revanche, on présentera de façon plus détaillée un

aspect du travail relatif à la collecte de noms de secteurs d'activité. Ces ensembles, qui désignent un regroupement d'activités et d'agents économiques, fournissent une entrée couramment utilisée pour la présentation des informations relatives à une entreprise sur un portail intranet.

Après avoir présenté la démarche retenue pour la constitution de ressources terminologiques dédiées à la documentation de tâches d'information dans un environnement professionnel contraint, on exposera les résultats obtenus en ce qui concerne la collecte et l'utilisation des noms de secteurs d'activité. On évaluera également la capacité de cette démarche à répondre aux besoins qui ont été évoqués précédemment : d'une part, rendre compte des usages linguistiques d'un public de site intranet pour une situation de communication donnée, et d'autre part, apporter des indications opérationnelles afin d'orienter les regroupements thématiques de contenus et l'organisation de l'arborescence des sites.

2. Une approche communicationnelle pour la construction de ressources terminologiques

2.1. La prise en compte d'un sociolecte particulier

Dans la mesure où la construction de ressources terminologiques dédiées à l'organisation de l'information sont développées dans un cadre professionnel précis, il est tout d'abord nécessaire de caractériser le sociolecte propre à l'organisation concernée. En effet, dans l'optique retenue, il s'agit moins d'élaborer la terminologie d'une science, d'un secteur d'activité ou d'un métier, que de parvenir à caractériser les échanges linguistiques ayant cours au sein d'une entreprise dans laquelle chaque activité constitue un foyer énonciatif particulier. C'est dire qu'un tel sociolecte agrège non seulement différentes terminologies liées aux activités professionnelles, mais aussi des usages linguistiques différents liés à chaque situation. De ce point de vue, la notion de "parler d'entreprise" proposée par D. de Vecchi constitue un modèle adapté pour rendre compte de cette diversité. Elle permet d'englober "(…) *l'ensemble des processus linguistiques qui actualisent les répertoires linguistiques des membres d'une communauté, définie en fonction de l'appartenance à une entreprise. Autrement dit, la cristallisation linguistique de tout moyen de communication mis à la disposition d'une entreprise, pour des conceptualisations ayant des origines diverses*[1]". On voit que le périmètre à prendre en compte est immense, et qu'il peut s'avérer

1 (Vecchi de 1999 : 316)

contradictoire avec les objectifs opérationnels qui sont poursuivis, tant pour ce qui est de la phase de construction que pour les mises à jour. Il faut donc identifier les critères permettant de resserrer la collecte sur les éléments nécessaires et suffisants.

2.2. Situations de communication et traces discursives

En d'autres termes, il s'agit de procéder à un découpage au sein du sociolecte qui soit adapté au besoin. On s'appuie pour cela sur les caractéristiques de la situation d'échange d'informations réelle qu'il s'agit de documenter. On pose que celle-ci correspond à une situation de communication particulière dont peut rendre compte le modèle proposé par C. Kerbrat-Orecchioni[2]. Même si ce modèle a pour référence une situation simple d'interlocution, un tête-à-tête, il possède des caractéristiques suffisamment génériques pour lui permettre de situer la plupart des échanges verbaux réels. Dans la mesure où les représentations de la situation de communication et le référent du discours sont convertis en contenu du message, toute situation de communication laisse dans les discours des traces qu'il est possible d'analyser et d'interpréter. Dès lors, la documentation d'une situation de communication où sont impliquées des activités relatives à la diffusion et à l'organisation des informations sur un site peut s'appuyer sur les traces discursives laissées par une situation de communication analogue.

On dispose ainsi de critères permettant de sélectionner les discours susceptibles d'être utilisés comme sources pour la construction d'un référentiel terminologique adapté à la situation qu'il s'agit de documenter. Sont comparés principalement les finalités, le propos, le statut des partenaires légitimes, les lieux et moments légitimes, les supports matériels et l'organisation textuelle[3]. Chaque situation de communication ayant ses caractéristiques propres, il est inévitable de travailler par valeur approchée. Par ailleurs, si un site possède plusieurs publics, il est nécessaire d'utiliser plusieurs sources en accord avec les manières de dire et de classer propres à ces publics. Lorsque les traces discursives laissées par l'un de ces publics sont difficiles à identifier, comme c'est le cas pour des non-initiés à l'intérieur d'une entreprise, une solution de contournement peut consister à identifier des discours spécialement produits par l'entreprise à destination de populations ne partageant pas le même référentiel.

Afin de documenter l'organisation de l'information sur un portail donnant accès à une centaine de sites destinés à un public interne de 150 000

2 (Kerbrat-Orecchioni 1999 : p. 22)
3 (Charaudeau et al. 2002)

personnes environ, on a constitué un corpus de rapports d'activité. Ce genre de discours est utilisé dans le cadre de la communication institutionnelle et financière afin de présenter annuellement une vitrine des activités de l'entreprise à destination de publics divers, aussi bien externes qu'internes. En considérant que la situation de diffusion d'information à documenter se situe en 2004 sur l'intranet de l'organe central du Crédit agricole, on a constitué un corpus de rapports d'activité des années 1995 à 2003. Destiné à être exploité à l'aide des techniques textométriques, le corpus a été converti dans un format *machine readable* et segmenté en formes graphiques[4]. Les principales partitions utilisées correspondent aux documents "rapports d'activité" de chaque année, aux rubriques découpant le texte de ces documents, et aux paragraphes organisant les textes des rubriques. Le texte original a été repris, y compris lorsqu'il comportait des graphiques, organigrammes et autres histogrammes. On a restitué cette différence de présentation de l'information en opérant une distinction entre les rubriques à dominante syntactique (texte suivi) et celles qui sont à dominante non syntactique (organigrammes, etc.).

2.3. Référentiel et noms propres

Lorsqu'une source a été identifiée et constituée en corpus, son exploitation soulève de nouvelles questions d'ordre théorique et pratique. Tout d'abord, la collecte d'unités destinées à constituer le référentiel terminologique doit permettre de capter les verbalisations réalisées à propos d'un référentiel spécifique, et non tout le vocabulaire du corpus de textes utilisé comme source. La notion de référentiel, avancée par F. Gonseth[5], et reprise par J. Rey-Debove permet d'aborder l'univers des propos liés à une situation de communication concrète. Pour la documentation d'une situation de partage d'information sur un intranet, on peut restreindre le périmètre à "*l'ensemble des objets (concrets ou abstraits, réels ou imaginaires) dont un locuteur peut parler dans une langue donnée* [et qui ont un rapport direct ou indirect avec l'exercice de ses activités dans une entreprise][6]". Il reste que les objets à prendre en compte peuvent s'avérer très nombreux. On a donc recherché un point de départ pour la collecte qui garantisse la sélection dans les discours des unités les plus caractéristiques du référentiel d'une organisation.

---

4 Les principales caractéristiques textométriques du corpus sont les suivantes : plus de 200 000 occurrences pour 11 000 formes graphiques différentes, un découpage en 9 parties correspondant chacune à un rapport d'activité.

5 (Gonseth 1975 : p.22)

6 (Rey-Debove 1998 : p.289). La définition de J. Rey-Debove est complétée par la partie entre crochets dans (Erlos 2009 : 121 et 766).

Les noms propres, unités un peu négligées autant en linguistique, qu'en terminologie[7] ont semblé constituer un point de départ adapté. En effet, ceux-ci possèdent des propriétés pragmatiques pertinentes pour la démarche adoptée, car ils facilitent l'identification de certains objets caractéristiques d'un référentiel. Par ailleurs, ils établissent un lien dénominatif stable entre un référent et une dénomination, ce qui leur permet d'être présents et donc repérables dans de nombreux discours reflétant divers usages d'un même sociolecte. Enfin, ils constituent un bon indicateur des changements affectant un référentiel, ce qui explique, entre autres, leur utilisation dans des problématiques voisines telle que la veille technologique ou concurrentielle. Cependant, les noms propres forment aussi une catégorie d'unités hétérogène, aux contours mal définis, hormis les toponymes, les patronymes et les prénoms. De plus, leur fonctionnement discursif est relativement peu étudié en dehors des problématiques de l'antonomase et de la référence dans les discours. De même, leur intégration dans les référentiels terminologiques reste marginale. Enfin, un parler d'entreprise ne peut pas être réduit à ses noms propres. Il a donc été nécessaire d'identifier les moyens permettant de conduire une collecte qui prenne les noms propres pour point de départ d'une exploration des données textuelles destinée à capter, entre autres, les noms de secteurs d'activité véhiculés par un parler d'entreprise dans une situation de communication donnée.

3. Le cas des secteurs d'activité

3.1. Unités pilotes et explorations textométriques

Dès lors, il s'agit de procéder à un recensement des manières dont l'information est structurée dans les rapports d'activité. On a vu que pour cela on propose de partir des dénominations propres caractéristiques d'un référentiel. Celles-ci sont utilisées comme unités-pilotes afin d'explorer leur voisinage dans les textes du corpus[8]. Le repérage des secteurs d'activité présents dans le corpus revient alors à rechercher les modes d'articulation discursifs entre dénominations propres et noms de secteurs. Un certain nombre de questions se posent alors : comment repérer et caractériser ces

7 Pour un état de la question en terminologie présenté par un auteur favorable à l'intégration des noms propres dans les terminologies, on renvoie à (Kocourek 1991). J. Humbley partage le constat de R. Kocourek sur l'ostracisme qui frappe cette catégorie d'unités et propose une piste d'intégration possible des noms propres aux référentiels terminologiques (Humbley 2006). Pour un point à jour sur la question des noms propres en linguistique, voir (Vaxelaire 2005).

8 Sur les 3000 dénominations propres et variantes recensées dans le corpus, on a utilisé un échantillon composé d'une centaine de dénominations propres de produits et de personnes morales les plus fréquentes dans le corpus.

relations ? Celles-ci permettent elles de recenser tous les secteurs d'activités évoqués dans les textes ?

Une première étape consiste repérer les classes de dénominations présentes dans les textes du corpus puis à étudier leur fonctionnement discursif. Pour cela, on utilise les principaux représentants de chaque classe (ceux qui possèdent les fréquences les plus élevées), et on recherche les relations qu'ils entretiennent avec les autres unités présentes dans les textes. Parmi ces relations, on a distingué celles qui relèvent d'un type et celles qui réalisent le type dans une instance particulière. En cherchant à identifier des relations types, on vise à établir l'existence d'une structure de contenu qui constituerait, pour une classe de noms propres (ou certains de ses représentants) et pour un genre de discours donné, le principe organisateur des sortes d'informations associées de façon récurrente aux entités nommées. L'expérience montre que le programme constitué par cette structure topique (que doit-on dire de telle entité dans tel genre de discours, compte tenu des circonstances ?), n'est pas réalisé de façon complète pour tous les membres d'une même classe. Elle nous semble néanmoins de nature à permettre une intégration des noms propres dans un référentiel terminologique, dans la mesure où cette structure est un gage de stabilité relative attestée par les usages. À l'issue de cette étape, les classes de noms propres les mieux représentées dans le corpus sont identifiées, et parmi celles-ci, on retient celles qui tissent avec les autres composantes du vocabulaire des relations pertinentes pour la collecte de données envisagée. Dans le corpus utilisé, les classes des noms propres de personnes morales et de produits sont parmi les mieux représentées, mais surtout, elles possèdent des structures types de contenu établissant une relation avec des noms de secteurs d'activité, comme le montre schéma ci-dessous.

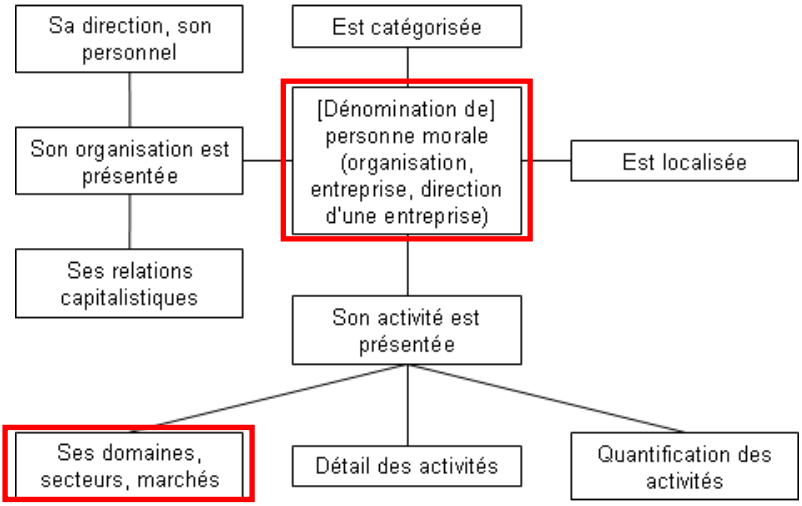


Figure 1. structure type de la classe des dénominations propres des personnes morales dans un corpus de rapports d'activité

Une seconde étape consiste à sélectionner des unités appartenant aux classes les plus pertinentes, compte tenu des buts assignés à la collecte, et à procéder à l'exploration de leur voisinage textuel afin d'identifier la présence de noms de secteurs d'activité. Le repérage de ces relations établies dans les textes a été réalisé à l'aide de techniques textométriques classiques. Pour les unités de fréquences faible à moyenne (de 3 à 20 occurrences dans le corpus), un repérage manuel à l'aide des concordances, des segments répétés[9] et de la carte des sections des textes (celles-ci correspondent ici aux phrases et aux paragraphes) permet un dépouillement complet. Lorsque les contextes sont plus nombreux, pour les dénominations de 20 à plusieurs centaines d'occurrences, on s'appuie sur le calcul des co-occurrents. Celui-ci est obtenu à l'aide de la méthode des spécificités[10] qui opère une comparaison entre les sections des textes comportant une occurrence au moins de la dénomination propre, et celles qui en sont dépourvues. De cette confrontation entre sous-ensembles du vocabulaire du corpus résulte une liste de formes ou de segments répétés particulièrement présents ou peu fréquents dans les contextes d'apparition de la dénomination propre. Les co-occurrents sont assimilés aux unités

9 Un segment répété est "une suite de formes dont la fréquence est supérieure ou égale à 2 dans le corpus." (Lebart et al. 1994).

10 "Pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique positive pour la partie j (ou forme caractéristique de cette partie) si sa sous-fréquence est « anormalement élevée" dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou supérieures à la sous-fréquence constatée est inférieure au seuil fixé au départ." Ibidem.

plus particulièrement présentes. Ce calcul fournit ainsi une sorte de résumé du contenu des contextes qui doivent être explorés. Mais rien n'empêche de procéder à des vérifications sous la forme de sondages réalisés directement dans les textes.

En raison de son orientation avant tout communicationnelle, l'approche proposée ne cible pas a priori un groupe d'unités terminologiques "bien formées" selon les patrons syntagmatiques les plus fréquemment utilisés dans les langues spécialisées, et repris dans les programmes d'extraction automatique de candidats termes[11]. Par ailleurs, le repérage et l'extraction de noms propres, voire la catégorisation des entités nommées à partir de catégories prédéfinies[12], ne constituent qu'une série d'étapes destinée à permettre la collecte d'autres unités utilisées dans les textes d'un corpus. Enfin, les corpus servant à documenter des situations de communication différentes (publics experts / publics néophytes ; publics appartenant à l'entité émettrice de l'information / publics appartenant à des entités différentes accédant à l'intranet d'une holding, etc.), l'outillage informatique retenu doit être portable d'un corpus de textes à l'autre, quel que soit le genre de discours.

Enfin, la collecte de ces données nécessite que le terminologue puisse naviguer sur tous les paliers textuels s'étageant du corpus pris dans sa globalité à l'occurrence d'une forme graphique en passant par le syntagme, la phrase, le paragraphe, la rubrique (encadrée par deux intertitres) et la partie (correspondant à un rapport d'activité). La présence des unités doit être quantifiée et l'usage doit être documenté au moins pour ce qui concerne la récurrence observée dans l'emploi de telle ou telle expression. Cela suppose que l'on dispose du recul suffisant pour l'étude des variations de fréquence. C'est pourquoi, l'approche retenue préconise la constitution de corpus organisés sous la forme de séries textuelles chronologiques homogènes, c'est-à-dire restreintes aux éléments d'une série de discours produits dans des conditions d'énonciation similaires. Ce type de corpus doit être ouvert (on parle aussi de corpus de suivi, ou *monitoring corpus*), afin d'accueillir de nouvelles parties destinées à suivre l'évolution des usages linguistiques. La plupart des logiciels de textométrie[13] offrent l'outillage

11 On renvoie aux synthèses de (Bourigault et al. 2000), (Poibeau 2003), (L'Homme 2004).

12 On renvoie sur ce point à (Maurel et al. 2001).

13 Ces programmes font l'objet d'une présentation et peuvent être utilisés en ligne ou téléchargés aux adresses suivantes :
http://www.ling.uqam.ca/ato/sato/
http://ancilla.unice.fr/~brunet/pub/hyperbase.html
http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW
http://weblex.ens-lsh.fr/wlx/
http://www.xaira.org/
Dans ce travail on a plus particulièrement utilisé le logiciel Lexico.

nécessaire pour une exploitation des données textuelles en rapport avec la démarche proposée. En effet, outre leur portabilité et leur robustesse, ils permettent, d'une part, de disposer de données statistiques sur le vocabulaire d'un corpus, et d'autre part, de porter des jugements en termes de probabilité sur les fréquences des formes attestées.

3.2. Secteurs d'activité et normes du corpus

La notion de domaine constitue un principe organisateur essentiel pour les terminologies, mais il peut-être conçu comme leur étant interne ou externe. Dans le premier cas, il correspond à la reconstruction d'un système de concepts appartenant à un champ notionnel[14]. Dans le second, il est assimilable à un champ d'action[15] regroupant activités, procédés, acteurs, produits, bref, tout ce qui relève d'une *praxis* plus ou moins institutionnalisée à une époque donnée. On parlera alors tantôt de domaine de connaissance ou de domaine d'activité. C'est cette deuxième acception que l'on a retenue dans ce travail, et pour la distinguer de la première, on parlera plutôt de secteurs d'activité. Ceux-ci constituent un moyen utilisé couramment afin d'évoquer une forme d'activité humaine marchande ou non marchande. Généralement, chaque classement s'insère lui-même dans un processus finalisé, comme, par exemple, la *Nomenclature d'activités française* (*NAF*), dont le but essentiel est d'organiser les données statistiques relevant de "*l'information économique et sociale*[16]". Dans l'approche que l'on a retenue, on vise la réutilisation du découpage de l'activité proposé par l'énonciateur collectif des rapports d'activité.

En l'absence de définitions satisfaisantes ou d'une liste normalisée de dénominations, il est nécessaire d'identifier la manière dont certaines expressions sont spécialement utilisées dans les textes afin de désigner les secteurs d'activité. Tout d'abord, la phrase fournit un premier cadre lorsque le secteur d'activité est introduit sous la forme d'un circonstant. Dans cet espace, les deux dénominations d'entité nommée et de secteur d'activité peuvent se rencontrer selon le schéma type suivant : "l'entité X exerce son activité dans le secteur Y". Le patron type dans lequel sont associés ces deux ingrédients comporte une préposition, un nom plus ou moins général dont le sens comporte au moins un trait relatif à l'idée d'ensemble d'éléments, et le nom d'un secteur d'activité. Un exemple prototypique de cette construction serait "dans le domaine du crédit à la

14 (Depecker 2003 : 145)
15 (Bessé 2000 : 184)
16 Nomenclature des activités et guide sont disponibles à l'adresse : http://www.insee.fr/fr/methodes/default.asp?page=nomenclatures/naf2008/naf2008.htm, consultée en janvier 2009.

consommation". Le mot "domaine "peut être remplacé dans ce patron par "secteur", "marché" ou, de façon métonymique, par "métiers", "équipes", ou encore "activités". On rencontre ainsi avec la préposition "dans" : "dans l'assurance-vie" (7 occ.), "dans l'épargne bancaire" (6 occ.), "dans la banque privée" (10 occ.), etc. Avec la préposition "en" : "en assurance-vie" (13 occ.), "en crédit à la consommation" (3 occ.), "en épargne salariale" (6 occ.), etc.

En second lieu, on remarque que le secteur d'activité peut introduire un paragraphe et être ainsi désigné comme thème principal. Ce statut est en quelque sorte garanti dans une telle configuration par la reprise du nom du secteur d'activité dans l'intertitre. Lorsque c'est l'activité du secteur qui constitue l'objet principal restitué dans les textes du corpus, la mention de l'entité peut apparaître dans le même paragraphe, à une ou deux phrases de distance : "*L'assurance-vie a continué de se développer à un rythme rapide. Le chiffre d'affaires de Predica, qui s'est établi à 51,3 milliards de francs, a enregistré une hausse de 11 %, supérieure à celle du marché. L'encours a augmenté, quant à lui, de 24% pour atteindre 211,2 milliards de francs. La part de marché de Predica a, ainsi, progressé de 0,7% pour atteindre 9,7% des encours*". Cette position ouvrante en tête de paragraphe est également occupée, par exemple, par les syntagmes : "Le crédit-bail" (4 occ.), "L'agriculture" (6 occ.), "L'assurance-vie" (4 occ.), "Les collectivités locales" (9 occ.), "Les entreprises" (10 occ.), "Les métiers de gestion d'actifs" (1 occ.), "Crédit-bail:" (2 occ ;), "Assurance-vie :" (3 occ.), etc. Le contexte fourni par les rubriques possède des propriétés similaires, dans la mesure où celles-ci sont définies comme un ensemble de paragraphes compris entre deux intertitres.

Ces observations permettent de tirer une première série de conclusions. En effet, la norme relative au traitement des noms de secteurs d'activité dans les textes du corpus a pu être dégagée[17]. On a ainsi distingué les critères permettant de caractériser directement un syntagme comme étant un nom de secteur d'activité, et ceux pour lesquels cette caractérisation se fait de manière indirecte. La première sorte de critères a déjà été présentée : il s'agit des unités introduisant des univers de discours que l'on rencontre comme circonstants ou comme thèmes désignés en début de paragraphe. On peut ajouter à ces deux configurations principales le cas des séries. Ces dernières sont constituées lorsqu'un classificateur de secteur d'activité s'applique de façon distributive à une série d'unités, ou bien lorsqu'une unité candidate est insérée dans une série de dénominations de secteurs déjà identifiés. Les critères indirects correspondent à des utilisations de l'unité candidate qui suggèrent l'existence d'un secteur d'activité, sans pour

17 Cela permet de contrôler a posteriori l'exhaustivité de la collecte.

autant permettre de l'appréhender directement. On relève, par exemple, la situation dans laquelle la dénomination de secteur d'activité est le complément du nom d'un classificateur (société, banque, filiale, partenaire, leader, produit, offre, etc.), appliqué à une personne morale ou à une autre sorte d'entité nommée. Un autre indice est fourni par la reprise d'un tel nom dans un intitulé d'unité (direction ou département) appartenant à l'une des principales entreprises du groupe bancaire, comme "Marché des entreprises et des collectivités locales", où deux noms de marchés sont coordonnés pour former le nom d'une direction. Enfin, l'attribution d'une majuscule à un terme désignant une sorte d'opération ou d'objet constitue une indication à prendre en compte, en particulier dans les contextes à dominante non syntactique. Établies à partir de tels critères directs et indirects, les dénominations candidates de secteurs d'activité peuvent ensuite être évaluées à l'aune d'une seconde norme fournie par le corpus.

Lorsqu'une unité candidate a été repérée à l'aide d'au moins un critère direct, on peut décider de la retenir pour le référentiel terminologique en fonction de sa fréquence et de sa récurrence dans les textes du corpus. Pour cela, on vérifie que l'unité candidate possède une fréquence supérieure ou égale à trois, et une récurrence constatée sur deux parties au moins du corpus. Cela permet d'éliminer des unités dont l'apparition peut revêtir un aspect trop conjoncturel qu'il n'a pas paru nécessaire d'introduire dans un référentiel. En revanche, le traitement est différent pour les unités présentes dans la dernière partie du corpus, puisqu'elles ne peuvent pas être soumises au test de récurrence. Cependant, elles doivent valider les autres critères (au moins un critère direct de qualification attesté, et une fréquence supérieure à trois). Ce n'est que lorsqu'une unité candidate a rempli ces différents critères qu'elle est intégrée au référentiel terminologique. En définitive, ce sont plus de 200 noms de secteurs d'activité qui ont été collectés, pour 400 relations établies avec une centaine de dénominations propres, soit en moyenne 4 relations par dénomination propre[18].

3.3. Secteurs d'activité et diversité des usages linguistiques

Une première forme de cette diversité résulte de l'insertion des noms de secteurs d'activité dans une sorte de *continuum* homonymique. Ainsi, la "conservation de titres" ou l' "affacturage" renvoient à des opérations financières mais aussi aux secteurs correspondants. L'étude de ce phénomène montre qu'une même unité est susceptible d'endosser

18 Cette moyenne cache une répartition inégale entre les deux classes de dénominations propres. La classe des noms de personnes morales permet d'établir les ¾ des relations et de collecter les ¾ des noms de secteurs d'activité recensés.

différents statuts dans un même corpus de textes. En tant qu'homonyme, elle peut être utilisée d'une façon non spécialisée ; elle est également susceptible de désigner un concept ou une notion ; en tant qu'unité terminologique, elle peut en outre désigner par métonymie un secteur d'activité ou un marché ; dès lors, il n'est pas rare de la rencontrer comme dénomination propre d'une subdivision de l'organisation (service, département ou direction), ou comme dénomination d'un regroupement de subdivisions de l'organisation et/ou de secteurs d'activité.

Cette diversité des usages révèle également l'existence de points de vue différents qui se rencontrent dans le contexte des rapports d'activité. À côté des secteurs d'activité les plus fréquents (secteurs dédiés à une forme d'activité bancaire ou financière, marchés définis en termes de clientèles ou de zones géographiques), on note la présence de micro- et de macro-secteurs dont l'apparition est liée à des changements intervenant dans le référentiel évoqué par les rapports d'activité. En ce qui concerne les premiers, ils apparaissent lorsqu'un secteur de niveau intermédiaire est détaillé. Ainsi, les "moyens de paiement" peuvent être découpés en "cartes bancaires", en "monétique" ou en "gestion des flux" selon l'actualité. De même, on peut trouver en plus de "crédit-bail", des dénominations de sous-secteurs tels que "crédit-bail mobilier" et "crédit-bail immobilier", mais aussi "crédit-bail matériel", "location de longue durée" ou encore, "location de longue durée automobile". Ce recours aux micro-secteurs varie en fonction de l'actualité que représentent, par exemple, le lancement d'un produit ou la conclusion d'un accord commercial. L'acquisition d'entreprises des secteurs bancaire et financier s'accompagne en revanche de l'apparition de nouveaux secteurs, mais surtout, de macro-secteurs. Ces derniers ont pour fonction de procéder à des regroupements de secteurs existants de manière à produire une image plus harmonieuse du développement du groupe bancaire. C'est, par exemple, le pôle "assurances" qui chapeaute "assurance-vie" et "assurance IARD", ou le pôle "services financiers spécialisés" qui est placé au-dessus de "crédit à la consommation", "crédit-bail", et "affacturage". Ces macro-secteurs correspondent à des créations redondantes et tardives, en ce sens qu'ils s'ajoutent généralement aux secteurs d'activité déjà en place.

La collecte des noms de secteurs d'activité associés à une dénomination propre de personne morale ou de produit dans les textes du corpus est ainsi susceptible de livrer une photographie de la manière dont une forme d'activité est traitée dans une situation de communication spécifique. On donne ci-dessous la collecte de noms de secteurs d'activité réalisée à partir de la dénomination "Predica", qui est le nom d'une filiale d'assurance-vie du Crédit agricole.

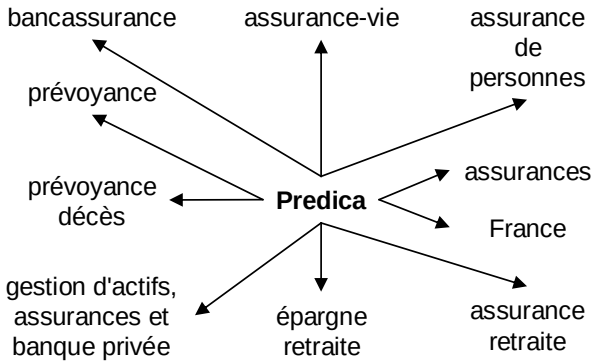


Figure 2. Secteurs d'activité associés à la dénomination "Predica" dans le corpus de rapports d'activité du Crédit agricole

Les différents rattachements dont la dénomination propre "Predica" fait l'objet mettent en évidence l'utilisation de plusieurs découpages hétérogènes utilisés pour la présentation d'un secteur d'activité pris au sens large. En effet, on constate l'existence de points de vue différents ("assurance-vie" pour le secteur d'activité au sens étroit, "bancassurance" précisant une position dans l'environnement bancaire, "France" qui renvoie au marché domestique de l'assurance-vie), mais aussi de macro-secteurs ("assurances" et "gestion d'actifs, assurances et banque privée") et de micro-secteurs ("prévoyance", "prévoyance décès"). On note également la mention de secteurs parallèles ("assurance retraite", "épargne retraite"), qui relèvent du même flottement terminologique que celui qui concerne "assurance-vie" et "assurance de personnes".

Dans cet ensemble, le nom de secteur correspondant au niveau intermédiaire ("assurance-vie") apparaît non plus comme le seul moyen de caractériser l'activité de l'entreprise Predica, mais comme une possibilité parmi d'autres. La collecte réalisée à partir d'une dénomination pilote offre ainsi une restitution schématisée des différentes facettes d'un secteur d'activité tel qu'il est évoqué dans les discours du corpus. Par ailleurs, le recours à une série textuelle chronologique met en évidence le fait que ces dénominations sont concurrentes sur la durée. C'est ce que montre la graphique ci-dessous, qui présente la ventilation des occurrences des dénominations de cinq secteurs d'activité associés à la dénomination "Predica".

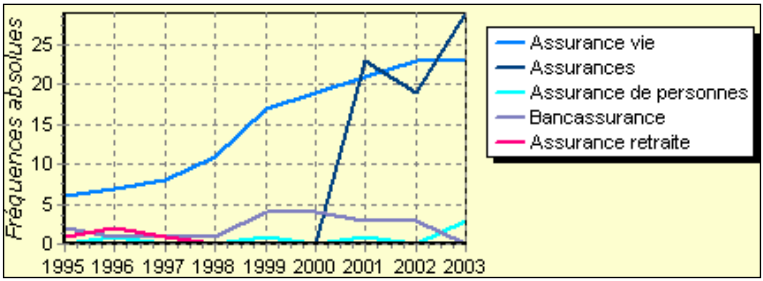


Figure 3. Ventilation en fréquences absolues de cinq noms de secteurs d'activité sur les neuf parties du corpus de rapports d'activité

La formalisation des rattachements d'une entité nommée à différents secteurs d'activité permet ensuite de documenter l'organisation de l'information sur des sites, en proposant les principaux découpages utilisés relativement à un pan de l'activité du groupe bancaire.

3.4. Intégration des données collectées dans un référentiel terminologique

La formalisation dans un référentiel terminologique du lien établi en discours entre une dénomination propre et un nom de secteur d'activité ne va pas de soi. On pourrait être tenté de traduire ce lien par une relation de type "partie-tout", étant donné que les secteurs d'activité fournissent les cadres dans lesquels prend place le fonctionnement économique de certaines entités. Néanmoins, la norme *ISO 704* rappelle qu'une telle relation ne peut être établie que si elle repose, non pas sur une fonction discursive consistant à ouvrir des domaines de discours, mais sur les traits caractéristiques des concepts que l'on cherche à relier par une telle relation hiérarchique : "*On considère qu'il existe une relation partitive lorsque le concept superordonné représente un tout, et que les concepts subordonnés représentent des parties de ce tout. Les parties s'assemblent pour former le tout* (…)[19]". La question se pose, même lorsque l'on cherche à établir cette relation à partir de marqueurs présents dans les textes : il s'agit de retrouver certains traits définitionnels, qui sont ensuite validés dans un cadre terminologique[20]. Or, l'absence de définition stable associée aux dénominations propres, constitue un obstacle important à leur intégration dans un référentiel terminologique. Le contenu de la dénomination ne peut pas se substituer à une telle définition dans la mesure où il repose sur les seules indications fournies par les discours.

19 (AFNOR 2001 : 5.4.2.3. p.10)
20 (Otman 1996 : 82) et (Condamines et al. 2000)

Par conséquent, on propose de relier les deux sortes d'unités par une relation associative, reposant non plus sur un emboîtement conceptuel hiérarchique, mais sur les données de l'expérience. Parmi les exemples cités par la norme[21], la relation "contenant – contenu" semble convenir au type de relation que l'on souhaite formaliser dans un référentiel terminologique. Les secteurs d'activité jouent alors le rôle de contenants possibles pour certains objets et personnes morales. Cette souplesse a néanmoins pour inconvénient de bloquer l'héritage dans les gestions automatisées de réseaux sémantiques, ce dernier étant réservé aux relations hiérarchiques. Lorsque cette règle est observée, la relation associative reste dotée d'un potentiel informationnel aussi important que la relation hiérarchique, car elle permet de multiplier les passerelles entre dénominations propres et termes.

4. Conclusion

Le recensement des noms de secteurs d'activité et leur intégration dans un référentiel terminologique permet de mettre au jour plusieurs phénomènes qu'il est nécessaire de prendre en compte afin d'adapter l'organisation des sites à des publics qui ne sont pas initiés à toutes les facettes de l'activité d'un groupe bancaire. En effet, un secteur d'activité est rarement représenté par une dénomination unique dans le cadre d'un même sociolecte. L'observation des usages montre qu'en dehors des quasi-synonymes, il est également nécessaire de collecter des dénominations différentes qui mettent l'accent non seulement sur des aspects plus précis au sein d'un même secteur d'activité, mais aussi sur l'expression de points de vue complémentaires. En second lieu, l'utilisation d'un corpus organisé en série textuelle chronologique permet de mettre en évidence que ces dénominations se font concurrence dans la durée. Enfin, le développement de l'activité prenant souvent la forme d'acquisitions de nouvelles entreprises, il engendre un renouvellent continu du stock des dénominations en place, et provoque l'apparition de noms de secteurs d'activité "chapeaux" destinés à habiller d'un effet de cohérence le développement économique.

Les expressions collectées, ici des noms de secteurs d'activité, doivent permettre de procéder à des regroupements thématiques concurrents sur les pages d'un portail, de manière à proposer aux visiteurs des navigations mieux adaptées à leur vocabulaire ou à leurs habitudes de classement. En prenant les noms propres présents dans les textes d'un corpus comme

21 (AFNOR 2001 : 5.4.3. p.13)

point de départ de la collecte, on s'expose au risque que celle-ci soit plus ou moins productive en fonction des genres de discours utilisés. Cette limitation est contrôlée par la méthode proposée, dans la mesure où l'étude du fonctionnement discursif des noms propres permet de l'anticiper. Il reste qu'à l'échelle d'un intranet comportant une centaine de sites, il semble que l'approche proposée gagnerait à être pilotée au sein d'un observatoire du parler d'entreprise. Celui-ci aurait pour tâche de coordonner la construction des référentiels terminologiques en fonction des principales situations de communication rencontrées sur un intranet.

Bibliographie

AFNOR *Travail terminologique* (avril 2001) : NF ISO 704, ISSN 0335-3931

Béjoint H., Thoiron P. (dir.) (2000) : *Le sens en terminologie*, Lyon, PUL

Bessé B. de (2000) : "Le domaine", in *Le sens en terminologie*, Lyon, PUL

Blampain D., Thoiron P., Van Campenhoudt M. (dir.) ([2005], 2007) : *Mots, termes et contextes – Actes des 7èmes journées scientifiques des chercheurs du réseau Lexicologie, Terminologie, Traduction*, Bruxelles, Paris, CPI

Charaudeau P., Maingueneau D. (dir.) (2002) : *Dictionnaire d'analyse du discours*, Paris, Seuil

Condamines A., Reyberolle J. (2000) : "Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode" in *Ingénierie des connaissances*, Paris, Eyrolles

Bourigault D., Jacquemin C. (2000) : "Construction de ressources terminologiques", in J.-M. Pierrel (dir.), *Ingénierie des langues*, Hermès, Paris

Depecker L. (dir.) (2005) : *Langages* n°157, Paris, Larousse

Depecker L. (2003) : *Entre signe et concept*, Paris, PSN

Erlos F. (2009) : *Discours d'entreprise et organisation de l'information*, thèse, Université de Paris 3

Gonseth F. (1975) : *Le référentiel univers obligé de médiatisation*, Lausanne, L'Âge d'Homme

Humbley J. (2006) : "Terminologie et noms propres" in *Des arbres et des mots*, Bruxelles, Éd. du Hasard

Habert B., Nazarenko A., Salem A. (1997) : *Les linguistiques de corpus*, Paris, Armand Colin

Kerbrat-Orecchioni C. ([1980] 1999) : *L'énonciation*, Paris, Armand Colin

Kocourek R. (1991) : *La langue française de la technique et de la science*, Wiesbaden, Brandstetter

Lebart L., Salem A. (1994) : *Statistique textuelle*, Paris, Dunod

L'Homme M.-C. (2004) : *La terminologie : principes et techniques*, Montréal, PUM

Maurel D. et Guenthner F. (dir.) (2001) : *TAL*, vol. 41 n°3, Paris, Hermès

Otman G. (1996) : *Les représentations sémantiques en terminologie*, Paris, Masson

Poibeau T. (2003) : *Extraction automatique d'information*, Paris, Hermès

Rey-Debove J. (1998) : *La linguistique du signe – Une approche sémiotique du langage*, Paris, Armand Colin

Slodzian M. (2000) : "L'émergence d'une terminologie textuelle et le retour du sens", in *Le sens en terminologie*, Lyon, PUL

Vaxelaire J.-L. (2005) : *Les noms propres – Une analyse lexicologique et historique*, (thèse publiée [2001]), Paris, Honoré Champion

Vecchi D. de (1999) : *La terminologie en entreprise – Formes d'une singularité lexicale*, thèse, Université de Paris 13

A propos des auteurs

Frédéric Erlos
Crédit Agricole S.A. (Intranet) – EA 2290 SYLED Paris 3
91, Bd Pasteur
75015 Paris
frederic.erlos@credit-agricole-sa.fr

De l'agriculture biologique aux espaces naturels : une étude des syntagmes terminologiques à l'intérieur des textes de spécialité

Lavagnino Elisa

Résumé : Notre projet concerne l'étude du comportement et de la linéarité des syntagmes terminologiques (ST) en perspective contrastive français/italien, à l'intérieur des textes de spécialité. Notamment, notre objectif est de comparer les documents disponibles à l'intérieur des différents sites Internet de plusieurs parcs italiens et français avec l'objectif de vérifier les résultats obtenus de nos analyses précédentes langues de spécialité. L'étude de la réduction dans différents domaines est fondamentale pour établir quelles sont les relations qui relient les syntagmes terminologiques aux typologies textuelles dans leur contexte d'apparition, en nous concentrant sur les raisons qui déterminent leur variation, sur le plan synchronique, et leur évolution formelle sur le plan diachronique.

Mots clés : Syntagme terminologique, réduction, typologies textuelles, langue de spécialité

1. Introduction

Notre recherche a commencé avec un projet de glossaire trilingue sur l'agriculture biologique[1]. Lors de la détection de la terminologie, nous avons enregistré une forte présence de termes complexes qui n'étaient pas figés. Le problème du choix des termes vedette nous a poussées vers une étude de la variation et de l'instabilité des syntagmes terminologiques (ST) à l'intérieur de différentes typologies textuelles.

2. Les syntagmes terminologiques et la réduction

Le syntagme terminologique est une « unité linguistique complexe, composée au minimum par deux lexèmes et dont la combinaison morphosyntaxique totale a pour fonction de dénommer un concept » (Silva, Costa, Ferreira 2004). Il représente une unité complexe relativement figée susceptible d'occuper dans une phrase une position de constituant minimal et autonome appartenant au système notionnel d'un domaine spécialisé. Un syntagme terminologique est composé par une structure binaire[2]. Les parties qui le composent sont la tête et les constituants. La tête est représentée par le déterminé et les constituants par le déterminant. Dans le ST riserva naturale, la tête riserva est suivie par le constituant naturale.

L'étude du comportement du ST à l'intérieur des différents textes est intéressante parce qu'il joue un rôle fondamental dans cohésion textuelle : son instabilité donne lieu à des variantes, et les relations qui s'instaurent entre celles-ci vont renforcer la cohésion du texte et l'économie du discours. Le mécanisme que nous avons étudié est la réduction.

Selon Collet (Collet 1997), la réduction représente l'expression de la dynamique discursive du syntagme terminologique, dynamique qui se manifeste par le biais d'un ensemble de mécanismes qui modifient la linéarité des syntagmes terminologiques et qui les rendent discontinus. La réduction, donc, peut être considérée comme un phénomène de production de *variation terminologique*, influencé par le contexte discursif. Elle représente un outil discursif utile au locuteur, que l'on doit observer et analyser pour éviter l'ambiguïté dans le discours. La réduction représente un élément important de cohésion textuelle, et elle satisfait le besoin d'économie de la langue. La répétition de la forme pleine du syntagme

1 Voir www.disclic.it/certem/glos_bio

2 Patrick Drouin, Acquisition automatique des termes: l'utilisation des pivots lexicaux spécialisés, Thèse de doctorat, Université de Montréal, 2002;

alourdit un texte et est contraire à la règle d'économie, alors que l'emploi d'une variante réduite conserve le sens notionnel et référentiel, sans donner lieu à ambigüité. La réduction est donc un procès de production de *variation terminologique*, influencée par le contexte discursif.

Vu que le ST n'est pas figé, étudier la réduction peut s'avérer important afin de définir les processus de collaboration entre les unités textuelles et les unités terminologiques pour garantir la transmission du sens. La relation existante entre les formes réduites et le contexte nous a poussées vers la distinction de deux catégories réductionnelles qui se caractérisent par un degré d'autonomie lexicale différente : la réduction anaphorique et la réduction à caractère lexical.

3. La réduction anaphorique et la réduction lexicale

La réduction anaphorique (RA) est un processus discursif et textuel qui a une valeur plutôt contextuelle, elle se réfère à la cohésion et à la cohérence textuelle. La réduction lexicale (RL) donne vie à de véritables variantes terminologiques. Voilà des exemples :

Exemple pour l'italien:	Exemple pour le français:
Metodo di produzione biologica **RL** = Metodo biologico (62 %) ; Produzione biologica (22%) **RA** = Metodo (16%)	Mode de production biologique **RL** = Production biologique (49%) ; Mode de production (35 %) **RA** = Mode (16%)

Sur le plan fonctionnel, Marie-Paule Jaques affirme que la reprise anaphorique a pour fonction d'établir des liens entre les phrases d'un texte, donc de contribuer à la cohésion et la cohérence textuelles. Dans une certaine mesure, la réduction à caractère lexical remplit cette même fonction cohésive, à laquelle s'adjoint la capacité de créer des variantes susceptibles de devenir des membres permanents de la terminologie du domaine. […] (Jacques 2000) Sur le plan formel, on peut dire que la réduction anaphorique permet seulement l'élision des constituants, alors que la réduction lexicale permet la suppression de la tête ainsi que des constituants, c'est-à-dire des éléments forts et faibles. En général, on peut affirmer qu'entre les variantes anaphoriques et les ST s'instaurent des relations de type hyperonymique, et entre les variantes lexicales et les ST

des relations synonymiques, bien que ces variantes ne soient pas de véritables synonymes.

4. Les corpus

Un corpus doit permettre de reconnaître tous les termes qui sont objet d'étude de façon homogène et significative, ainsi que la stabilisation de leur sens, leur comportement actuel et leur évolution. Nous avons choisis des textes ayant un degré de spécialisation différent afin d'obtenir des résultats les plus possible significatifs et fiables. Analyser des textes sur une échelle verticalement asymétrique permettra d'analyser la valeur de la réduction sur l'axe socio-pragmatique de production textuelle.

Les discours de spécialité étudiés présentent des caractéristiques différentes. Retrouver des critères de classement univoques nous à poussées à réfléchir sur les typologies textuelles que nous avons analysées. Etant donné que notre étude s'appuyait sur le figement de la terminologie, nous avons cherché à les classer sur la base de leur dimension *verticale*. (Cortelazzo 1994)

En ce qui concerne l'agriculture biologique, la recherche de textes est partie d'une enquête menée lors de la réalisation du glossaire mentionné ci-dessus. Nous avons contacté les organismes locaux qui travaillent dans le secteur du biologique, organismes qui nous ont aidées dans le choix des matériaux et dans leur classement. Le réseau Internet a été notre source principale. On a étudié environ 70-80 sites par langue qui représentent les différentes typologies textuelles de la façon la plus homogène et équilibrée; la période de référence de cette analyse est mars-juin 2007.

Les typologies principales que nous avons détectées sont résumées dans le tableau qui suit :

Textes normatifs	Textes explicatifs	Textes de vulgarisation
1. directives européennes ; 2. directives nationales ; 3. lois régionales ; 4. règlements concernant les certifications	5. revues en ligne concernant l'agriculture biologique ; 6. sites des agences de certification du secteur ; 7. glossaires attestés	8. sites des exploitations liées au biologique ; 9. sites relatifs aux produits biologiques

En ce qui concerne les textes sur les dessaleurs d'eau, nous avons étudié environ 70-80 sites qui, pour l'instant, ne représentent pas de façon homogène les typologies, à cause de la difficulté de repérage des textes. Les typologies textuelles identifiées sont les suivantes :

Textes techniques			Textes de vulgarisation
textes techniques pour les techniciens	Manuels	textes techniques adressés aux usagers finaux	textes à propos de l'osmose inverse

La langue de spécialité des espaces naturels se trouve à mi-chemin entre la langue de dessaleurs d'eau et celle de l'agriculture biologique. Nous avons enregistré une augmentation du degré de figement dès l'entrée en vigueur des normatives européennes à propos des espaces naturels : la terminologie dans le discours scientifique spécialisé est devenue moins instable et les termes ont commencé à être structurés de façon univoque à l'intérieur des différents champs notionnels. En ce qui concerne les typologies textuelles prises en examen, nous pouvons maintenir ici la subdivision présentée pour la langue de l'agriculture biologique. Jusqu'à maintenant, nous avons analyse les sites Internet *www.parks.it* et *www.espacenaturel.fr*. Ces deux portails contiennent des liens à d'autres espaces naturels italiens et français. Notamment, nous avons étudié les revues "Parchi" pour l'italien et "Espace naturel" pour le français, téléchargeables complètement de ces sites. Ces revues nous ont permis d'avoir une vue d'ensemble du comportement des syntagmes terminologiques. Le tableau qui suit décrit dans le détail les textes qui font partie du corpus :

Textes normatifs	Textes explicatifs	Textes de vulgarisation
1. directives européennes ; 2. directives nationales ; 3. lois régionales ; 4. règlements concernant les projets d'aménagement environnemental (ex. Natura 2000)	5. revues en ligne concernant les réseaux de parcs nationaux, régionaux et transfrontaliers ; 6. glossaires attestés	3. sites concernant la vie quotidienne des parcs nationaux, régionaux; 4. newsletters concernant les activités promues par les parcs

5. Pivotal: instrument de liaison entre le corpus et la réduction

Le logiciel d'extraction des termes employé pour la recherche a été créé par Simone Torsani, chercheur de l'Université de Gênes. L'emploi de ce programme justifie le choix d'utiliser de textes en version *.html*. Pivotal est disponible en réseau. Il présente des traits communs aux autres logiciels d'analyse quantitative des textes. L'application est réalisée grâce à la combinaison des technologies les plus communes sur l'Internet et elle permet de créer et de gérer des corpus textuels qui ont une dimension moyenne (5 millions de mots, environ). Il peut effectuer différents types d'extraction, collocations et listes de mots. Le système enregistre tous les mots du corpus, donnant des suggestions aux personnes qui utilisent les textes.

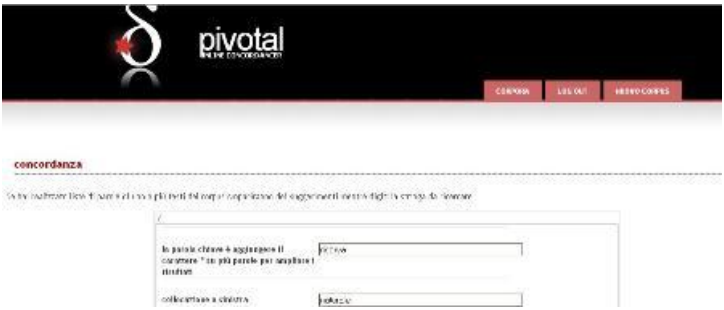


Figure 1. Pivotal et la concordance

Dans le cadre de notre projet, la possibilité d'analyser de nombreux textes de façon automatique nous a permis de disposer de résultats les plus fiables possibles. En particulier, le logiciel nous a aidés à repérer rapidement les ST et leurs variantes dans les différents contextes.

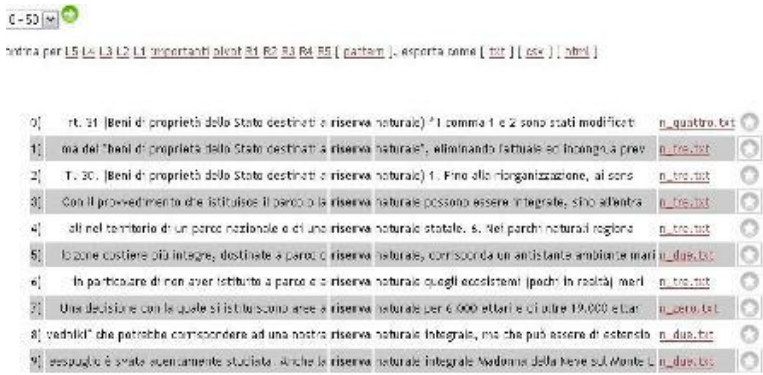


Figure 2. les repérages des ST et des variantes

6. Le projet en détail

Premièrement, nous nous sommes concentrées sur l'étude de la réduction des ST, en distinguant la réduction anaphorique de la réduction lexicale. Sur ce point, il faut souligner que le rapport qui relie l'emploi de ces types de variantes et le passage de l'information est inversement proportionnel : une variante anaphorique risque d'être ambiguë, en cas d'absence de son antécédent, alors qu'une variante lexicale, en général, réussit à transmettre la valeur informatique du ST plein.

Deuxièmement, nous avons focalisé notre attention sur la valeur des variantes terminologiques issues de ces deux types de réduction. Le rapport entre variation terminologique et néonymie ne se présente pas comme une opposition, ni comme une évolution nécessaire : entre les deux phénomènes on peut plutôt établir un continuum transformationnel de figement, des variantes lexicales à valeur stylistique pour arriver aux nouveaux termes entraînant de véritables variations notionnelles. (Giaufret e Rossi 2008) Dans notre étude, nous avons constaté qu'il existe un lien entre les variantes lexicales et l'évolution des termes auxquels elles se réfèrent. En particulier, certaines variantes prennent la place de termes vedette, c'est-à-dire qu'elles se transforment en termes à part entière. Cette évolution de forme a des conséquences importantes sur la terminologie d'une langue de spécialité, à l'intérieur de laquelle, l'évolution peut toucher la structure du terme et en modifier le signifié, le signifiant ou encore tous les deux. La réduction peut être considérée comme un mécanisme d'évolution des termes, si la forme réduite, notamment les variantes lexicales, s'implante dans l'usage, c'est-à-dire si les usagers commencent à l'utiliser au lieu du syntagme plein. Par exemple, les variantes réduites qui

suivent sont le résultat de la réduction du ST *prodotto ottenuto da produzione biologica* et *produit issu de l'agriculture biologique*, entrée officielle de la normative européenne. Ici, nous avons détecté également des variantes du terme déjà réduit :

FR - *Produit biologique*: [produit] tête + [biologique] constituant
Variantes : produit (33%), produit bio (67%)

IT - *Prodotto biologico*: [prodotto] tête + [biologico] constituant
Variantes : prodotto (47%), il biologico (53%)

Troisièmement, nous avons cherché à établir quelles sont les raisons qui poussent les syntagmes terminologiques à varier. Les résultats, jusqu'à présent, ont montré que les facteurs qui vont influencer cette variation sont :

1. le facteur diachronique: c'est-à-dire l'âge de la LSP. Si la terminologie est récente, elle enregistre un degré de figement plus faible, parce que la langue de spécialité n'est pas encore bien définie ;
2. le niveau de spécialité des textes, plus un texte est spécialisé et moins les ST présentent de réduction.

Maintenant, notre objectif est d'étudier quelles sont les relations qui s'instaurent entre la réduction et les typologies textuelles, notamment comment le cotexte peut influencer l'effacement d'une des parties du ST et sa charge sémantique d'une variante réduite.

7. La réduction en contexte

L'intérêt consacré à la variation en terminologie peut s'identifier avec l'étude de la terminologie textuelle et la possibilité d'exploiter de grands corpus pour des études systématiques du comportement des termes. Dans ce cadre, Sager affirme que :

Modern terminological theory accepts the occurrence of synonymic expressions and variants of terms and rejects the narrowly prescriptive attitude of the past which associated one concept with only one term (…). Terminology now adopts a corpus-based approach to lexical data collection. By being studied in the context of communicative situations, terms are no longer seen as separate items in dictionaries or part of a semi-artificial language deliberately devoid of the functions of other lexical items. (Sager 1990).

Notre objectif est en effet d'expliquer comment le cotexte peut influencer la réduction des syntagmes terminologiques. Le cotexte détermine le sémantisme du terme. En plus, les variantes s'insèrent dans un microcontexte linguistique, composé en général d'autres termes qui forment ensemble un domaine de référence, un champ conceptuel qui aide à situer du point de vue sémantique les formes réduites[3]. Selon ce microcontexte, les termes peuvent se voir réduits à leur tête ou à leur expansion. Nous présentons ici des exemples concernant ces types d'effacements pour analyser les relations entre les termes, les variantes réduites et les textes.

Voici les exemples que nous vous présenterons :

LANGUE ITALIENNE	*ST*
Agriculture Biologique	Agricoltura biologica; Metodo di produzione biologica; Prodotto biologico;
Espaces naturels	Riserva naturale Area protetta Parco nazionale Parco naturale regionale
LANGUE FRANÇAISE	*ST*
Agriculture Biologique	Agriculture biologique ; Mode de production biologique ; Produit biologique ;
Espaces naturels	Reserve naturelle ; Parc national ; Parc naturel régional.

7.1. L'agriculture biologique : exemples pour l'italien:

1) Agricoltura biologica

Variante lexicale : il biologico.

3 GAUDIN François et ALEXANDRU Cristina. Les contextes : à la source du terme ?; 7èmes Journées scientifiques AUF-LTT « Mots, termes et Contextes », 2005;

L'adjectif constituant s'est transformé en substantif, *il biologico.*

2) Metodo di produzione biologica

Variante anaphorique : metodo
Variantes lexicales : metodo biologico, produzione biologica

Ce syntagme enregistre un fort degré de variation : nous avons détecté des variantes lexicales et une variante anaphorique.

3) Prodotto biologico

Variante anaphorique : prodotto
Variante lexicale : il biologico

Nous avons déjà cité ce syntagme à cause de l'évolution qu'il a subi au fil des années. Ici, nous soulignons encore que cette variante représente désormais le terme principal.

Ce que nous voulons souligner ici est l'effacement de la tête du syntagme. L'adjectif *biologico* subit un procédé de substantivation qui lui permet d'acquérir un rôle principal dans les textes. Sa charge sémantique augmente. En plus, selon le contexte, cette variante peur représenter d'autres syntagmes : *metodo di produzione biologica* et *prodotto biologico*. Le contexte ou bien le cotexte est ici fondamental pour comprendre à quel terme se réfère la variante.

Voilà un exemple de texte :

L'incredibile, l'indicibile è stato approvato a larga maggioranza da burocrati che sembrano sempre più rappresentare se stessi. L'approvazione di questo nuovo Regolamento Europeo sull'agricoltura biologica è un passo indietro che il biologico europeo ora deve affrontare e che sicuramente risponde a logiche che non sono quelle di **chi lo produce e lo consuma**.[4]

Ici, le cotexte détermine le sémantisme de la variante. La phrase *chi lo produce e lo consuma* (en gras) se réfère à la variante *il biologico* c'est-à-dire le produit biologique.

Il biologico guarda avanti
Secondo Ferrante "**il biologico** torna a crescere e questo deve essere visto come un'iniezione di fiducia per tutta l'agricoltura italiana". Ferrante parla dell'**agricoltura biologica** come modello di riferimento produttivo per le nostre aree rurali, per quelle a rischio abbandono. I prodotti e le colture biologiche come elementi fondamentali per promuovere uno sviluppo

4 http://italy.peacelink.org/ecologia/articles/art_22044.html

locale "sostenibile e durevole", legato alla tutela del paesaggio, alla conservazione della natura e della biodiversità. Parla di dati incoraggianti anche il sottosegretario alle Politiche agricole alimentari e forestali, Stefano Boco, che " dimostrano una significativa attenzione degli agricoltori verso il **metodo biologico**"[5].

Ce dernier texte présente plusieurs exemples de réduction :

1. *il biologico* (en gras) renvoie au ST *agriculture biologique,* repris plus tard dans le texte ;
2. *metodo biologico* (en gras) remplace *metodo di produzione biologica*.

Grâce à la spécificité du discours le ST plein n'apparaît pas dans le texte.

7.2. L'agriculture biologique : exemples pour le français

1) Agriculture biologique :

Variante lexicale : l'agriculture bio, le bio

Nous avons détecté deux types de variantes: *la bio* et *l'agriculture bio*. On doit encore signaler qu'en français, le syntagme réduit a subi une suppression qui a donné vie à une forme tronquée : *bio*. Ce phénomène est très fréquent en français, mais on pourrait se demander si une variation de ce type peut être considérée comme une réduction. Toutefois, la variante *la bio* prend la même valeur que la variante italienne *il biologico*.

2) Mode de production biologique ;

Variante anaphorique: mode, mode de production ;
Variantes lexicales: production biologique ;

La structure de ce syntagme permet différents types de réductions. Les variantes anaphoriques doivent être insérées dans un cotexte pour éviter d'être ambigües. Par contre, *Production biologique* est fortement thématisée.

3) Produit biologique

Variante anaphorique **:** produit
Variante lexicale **:** produit bio, le bio

Ici, nous avons enregistré une variante anaphorique et deux variantes lexicales.

5http://www.agricolturaitalianaonline.gov.it/contenuti/agricoltura/tecnologie/biologica/il_biologico_guarda_avanti;

En ce qui concerne la relation avec le texte, ici, nous consacrons notre attention aux variantes *production biologique* et *bio* : en cas de contexte peu clair, le signifié pourrait être ambigu.

Voilà des exemples :

En 1994, le décret no 94-492 aboutit à la mise en place du label **Agriculture Biologique** (AB) et à la création d'organismes de contrôle et de certification des produits biologiques sous la tutelle du ministère de l'agriculture. La labellisation au moyen du logo (AB) peut être renforcée de logos émanant de certificateurs indépendants comme les labels NATURE&PROGRES (éthique et écologie) et DEMETER (biodynamie) qui sont des organismes privés qui ont contribué aux fondements de la **bio**.[6]

Ici, la présence du syntagme plein *agriculture biologique* (en gras) aide à retrouver le signifié de la variante *bio*.

Voilà un texte intéressant : ici, pour éviter l'ambigüité, l'acronyme bio (en gras) est rendu explicite par la phrase *pour biologique* (en gras).

Les produits **"bio", pour biologiques**, connaissent un véritable succès. La demande décolle à tel point qu'elle est supérieure à l'offre. La gamme de produits est de plus en plus large : des légumes à la viande en passant par les œufs, tous les produits végétaux et animaux ont leur label 100 % naturel [7].

Les exemples qui suivent soulignent l'importance du contexte et du cotexte dans la référenciation des variantes. Dans le texte A, *production biologique* (en gras) ne se réfère pas au mode de production, mais à la production en général. Par contre, dans le texte B, la variante substitue le ST *mode de production biologique*. Seulement une analyse attentive du cotexte et du microcontexte peut désambiguer les deux variantes.

a) Les acteurs de la ***production biologique****, et plus particulièrement les agriculteurs biologiques, appliquent des méthodes de travail fondées sur le recyclage des matières organiques naturelles et sur la rotation des cultures ; celles-ci visent à respecter l'équilibre des organismes vivants, qui peuplent le sol (bactéries, vers de terre, etc.).*[8]

b) Durant sa période d'adaptation à la ***production biologique****, la laiterie a choisi de produire ses propres préparations de fruits, par souci de cohérence et de qualité. Elle*

6 http://www.lesannuaires.com/annuaire-bio.html;
7 http://www.doctissimo.fr/html/nutrition/dossiers/produits-bio/produits-bio.htm
8 http://biogassendi.ifrance.com/reglesbio.htm

décide alors d'élaborer une nouvelle recette, nommée " fruits sur sucre", qui demandera plus d'une année d'investissement pour sa mise au point[9].

7.3. Les espaces protégés : exemples pour l'italien

1) riserva naturale

Variante anaphorique : riserva

2) area protetta

Le ST *area protetta* ne réduit pas : le composant *area* est très ambigu, même en contexte.

3) Parco Nazionale

Variante anaphorique : parco

4) Parco Naturale Regionale

Variante anaphorique : parco ;
Variante lexicale : parco regionale ;

Ces derniers ST sont fortement liés entre eux : ils partagent la même variante anaphorique. Dans ce cas, nous pouvons parler de co-hyponymie.

Voici quelques exemples :

Tra Etna e Vesuvio, tra le Cinque terre e Portofino, tra la Maremma e il Circeo dove passa la differenza, quel quid che ne fa in un caso un ***parco nazionale*** *e nell'altro un* ***parco regionale***[10]*?*

Dans ce cadre, il faut souligner ici l'usage du syntagme plein *parco nazionale* (en gras) et de la variante lexicale *parco regionale* (en gras). En cas de coprésence des ST, le cotexte ne réussit pas à garantir le passage de l'information puisque les syntagmes partagent la tête.

C'è il rischio, insomma, tutt'altro che remoto che ad una forte crescita delle **aree protette** si accompagni di fatto una sorta di 'normalizzazione', che renderà sempre meno chiara e netta l'identità, la fisionomia di un'**area protetta** rispetto ad altri organi e strumenti che operano sul territorio e sono preposti al suo governo[11].

Voilà un exemple de l'emploi du ST *area protetta* (en gras). *Area* peut se référer au concept d'espace en général. Dans ce domaine, donc, sa

9 http://www.la-cuisine-collective.fr/dossier/technologies/articles.asp?id=28;
10 http://www.parks.it/ilgiornaledeiparchi/eq1.pdf
11 http://www.parks.it/ilgiornaledeiparchi/eq1.pdf

référenciation est ambiguë, l'effacement de son déterminant cause une perte de sens, en risquant de rendre impossible le passage de l'information.

La ***Riserva Naturale Monte Rufeno****, area protetta facente parte del Sistema di Parchi e Riserve Naturali della Regione Lazio, si trova posizionata nell'estremo Nord della provincia di Viterbo al confine con Umbria e Toscana. […] Il territorio della* ***Riserva*** *é attraversato dal fiume Paglia, affluente del Tevere; […] Una seconda zona ad alta valenza naturalistica ancora non inserita nel contesto della* ***Riserva Naturale Monte Rufeno****, ma da tutti gli ambientalisti della zona considerata parte integrante e sostanziale di essa, é il Bosco del Sasseto. Sino all'istituzione della* ***Riserva*** *l'area era assai poco studiata sotto il profilo botanico e naturalistico in genere […]*[12].

Ici, la chaine anaphorique qui résulte de la réduction du ST plein nous clarifie la fonction du cotexte dans l'analyse de la variation terminologique. Il faut souligner que dans le texte, le ST réserve naturelle a été toujours détecté avant son nom propre, cet élément, apparemment banal, sera repris par la suite.

7.4. Les espaces protégés : exemples pour le français

1) réserve naturelle

Variante anaphorique : réserve

2) parc national

Variante anaphorique : parc

3) Parc Naturel Régional

Variante anaphorique : parc ;
Variante lexicale : parc régional ;

En ce qui concerne *parc national* et *parc naturel régional*, il existe une analogie entre l'italien et le français :

L'illustration sera donnée par la lutte contre l'ambroisie menée dans la **réserve** des Ramières. Par ailleurs, ce problème du non paiement des timbres-amendes par les étrangers étant également récurrent sur les espaces protégés du département, l'établissement a été autorisé à ouvrir une sous-section «**Réserves naturelles** catalanes » avec nomination d'un

12 http://www.parks.it/federparchi/rivista/P01/39.html;

préposé pour le dépôt des chèques concernant les timbres amendes dressés par les agents commissionnés des réserves naturelles[13].

Ici, la chaine anaphorique est inverse par rapport à l'italien. La variante réduite s'associe au nom de la réserve naturelle.

La spécificité d'un ***Parc Naturel Régional*** *par rapport à un* ***Parc National*** *est la complémentarité entre des objectifs de protection et de développement socio-économique. Le classement en* ***Parc Naturel Régional*** *n'induit pas de réglementation particulière mais un ensemble de mesures adoptées contractuellement par les collectivités et l'Etat. Sa gestion est confiée à un organisme regroupant au minimum le niveau régional et les communes du territoire.*[14]

Ici, nous proposons un cas similaire à l'italien. Le texte présente les parcs nationaux et régionaux. Les termes ne réduisent pas pour éviter l'ambigüité. Il faut toutefois souligner qu'en français, dans le domaine des espaces protégés, il y a une tendance à l'emploi des sigles au lieu des variantes réduites.

8. Conclusions

Les termes jouent un rôle dans les discours avec des valeurs sémantiques relatives, ils sont des organismes vivants qui bougent et se modifient (Faulstich 1999). Dans ce cadre, le rôle du contexte et du cotexte est fondamental, en ce qu'il peut déterminer le sémantisme d'une variante. En général, nous pouvons affirmer que la réduction oppose deux caractéristiques distinctives du texte : la clarté et l'économie. En effet, la clarté est assurée par l'exhaustivité d'un texte, par contre l'économie textuelle est obtenue à travers la non-répétition des informations déjà mentionnées : la réduction peut donc rendre un texte économique, mais elle risque de diminuer sa clarté. En ce qui concerne les deux langues comparées, elles présentent toutes les deux des phénomènes réductionnels. Le français tend à donner vie à des formes plus figées et collectivement acceptées, comme par exemples *la bio* ; par contre, l'italien ne se prête pas à ce type de suppression. En relation à la typologie textuelle, on peut affirmer que la terminologie dans les textes normatifs français est plus figée par rapport à l'italien, pour ce qui concerne les autres typologies textuelles, le degré de réduction devient de plus en plus élevé, selon le

13 Revue Espaces Naturels n°19, http://www.espaces-naturels.fr/a_la_une/la_revue_espaces_naturels/les_archives;
14 http://www.parc-volcans-auvergne.com/php/connaitre/parc/difference.php4

niveau de spécialité du texte. En général, il ressort de notre analyse que l'italien subit de plus le mécanisme de la réduction, alors que le français e une tendance majeure vers la siglaison. Par exemple, en italien le ST *parco naturale regionale* devient *parco regionale,* par contre, le français emploie le sigle PNR au lieu de *parc naturel régional.* Il faut souligner que même en italien nous avons détecté le sigle PNR, mais seulement dans les textes concernant les directives européennes, notamment des formulaires à remplir.

En ce qui concerne les relations entre texte et réduction, le texte doit contenir les moyens d'une interprétation correcte des variantes. Plus un concept est central à la terminologie et plus sa charge sémantique rendra autonomes ses variantes. Lorsque le terme réduit est employé sans mention antérieure du terme complexe, une telle opération de récupération via un antécédent est impossible. Le cotexte et le microcontexte du terme réduit jouent alors un rôle majeur. La mise en place de ce cadre interprétatif repose sur :

- le lexique environnant ;
- le thème du paragraphe ou de la section et le sujet du texte ;
- la centralité du référent du terme dans le domaine. (Jaques 2005)

Du moment que cette étude rentre dans le cadre d'un projet de thèse de doctorat concernant les relations de la variation terminologique, notamment la variation formelle de ST à l'intérieur des textes de spécialité, les résultats obtenus sont pour l'instant seulement partiels. Continuer l'analyse des ST en contexte permettra donc de nous concentrer sur les raisons externes aux syntagmes qui les poussent à varier.

Bibliographie

Adelstein Andreína (1998) : *Condiciones de reductibilidad léxica de los sintagmas terminológicos*, Área de Sistemas Léxicos, Universidad Nacional de General Sarmiento, Argentina - *www.riterm.net*, IV Simposio La Habana

Ahronian Céline et Béjoint Henri (2008) : *Les noms composés anglais et français du domaine d'Internet: une radiographie bilingue,* Meta vol. 53, n° 3, pag. 648-666

Cabre Maria Teresa (1998) : *La terminologie*, Colin-Presses Universitaires d'Ottawa, Paris-Ottawa

Cardero Garcìa Ana Marìa (2000) : *En torno a la frecuencia de algunas estructuras sintácticas en terminología*, Universidad Nacional Autónoma de México, Messico, *www.riterm.net*, VII Simposio Lisbona

Chrisment Claude, Hernandez Nathalie, Genova Françoise, Mothe Josiane, (2006), *D'un thésaurus vers une ontologie de domaine pour l'exploration d'un corpus*, *AMETIST*, INIST, Vol. 0, p. 59-92, septembre 2006

Collet Tanja (2004) : *Esquisse d'une nouvelle microstructure de dictionnaire spécialisé reflétant la variation en discours du terme syntagmatique*, Meta 49 n°1

Conicet et Ungs (2000) : Argentina - *www.riterm.net*, VII Simposio Lisbone 2000

Cortelazzo M. A. (1994) : *Lingue speciali. La dimensione verticale*. Padova, Unipress

Desmet Isabel, (2005), *Variabilité e variation en terminologie et langues spécialisées : Discours, textes et contextes*, 7èmes Journées scientifiques AUF-LTT "Mots, termes et Contextes"

Drouin Patrick (2002) : *Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés*. Thèse doctorale en linguistique, Université de Montréal

Faulstich Enilde (1998/1999) : *Principes formels et fonctionnels e la variation en terminologie*, Terminology, Vol. 5(1), p. 93-103

Freixa Aymerich, J. (2002) : *La variazio terminologica. Analisi de la variacio denominative et textos de diferent grau d'especialitzacio de l'area de medi ambient*; thèse doctorale, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra

Gaudin François et Alexandru Cristina (2005) : *Les contextes : à la source du terme ?* ; 7èmes Journées scientifiques AUF-LTT "Mots, termes et Contextes "

Giaufret Anna, Rossi Micaela (à paraître) : *Entre néologismes et variation terminologique dans le domaine des TICE : une analyse contrastive de trois langues romanes*, CINEO – I Congrés Internacional de Neologia en les llengües romàniques - Barcelona, 2008

Gouadec Daniel (1990) : *Terminologie : constitution des données*, Afnor

Jacques Marie Paule (2000) : *La réduction du syntagme terminologique au fil du discours*, Université de Toulouse - Cahiers de Grammaire 25

Jacques Marie-Paule (2005) : *De termes réduits comme révélateur de la centralité dans le domaine d'emploi,* 7èmes Journées scientifiques AUF-LTT "Mots, termes et Contextes"

Kageura Kyo (1999) : "Theories "of" terminology : A quest for a framework for the study of term formation", dans *Terminology*, vol. 5, no 1, p. 21-40

Kornfeld Laura, Resnik Gabriela. *Sintagmas terminológicos con adjetivos pasivos*

Kornfeld Laura, Resnik Gabriela (2002) : *Sintagmas terminológicos con adjetivos deverbales*, CONICET et UNGS, Argentina - *www.riterm.net*, VIII Simposio Cartagena

Kuguel Inés (1998) : *La reducción léxica de sintagmas terminológicos en el contexto discorsivo*, IDH – UNGS, Argentina - *www.riterm.net*, IV Simposio La Habana

Lino Maria Teresa, Rijo F. (2005) : *Contexte et néologie terminologique dans le domaine médical*, Centro de Linguística da Universidade Nova de Lisboa, U.I. "Lexicologie, lexicographie et terminologie"

Sager J. C. (1990) : *A practical course in terminology processing* John Benjamins : Amsterdam/Philadelphia (PA)

Temmerman Rita (2000) : *Une théorie réaliste de la terminologie: le sociocognitivisme*. Terminologie Nouvelles 21, 58-64

Torsani Simone (2007) : *DDL en réseau: un exemple d'utilisation des technologies Ajax pour améliorer les outils internet d'apprentissage des langues*, 2007 in ISDM n°29 - TICE MEDITERRANEE 2007

Vivaldi Jordi Màrquez, Lluís et Rodriguez Horacio (2001) : "Improving term extraction by combining different techniques". *Terminology* 7:1. 31-47

A propos des auteurs

Elisa Lavagnino

CeRTeM
Centre de Recherche en Terminologie Multilingue
DISCLIC - Université de Gênes
elisa.lavagnino@gmail.com
http://www.disclic.unige.it/certem/

Cette édition a été imprimée
en deux cent cinquante exemplaires
le 9 février deux mille dix