

Dans un monde où la communication et le partage d'information sont au cœur de nos activités, les besoins en terminologie se font de plus en plus pressants. Il est devenu impératif d'identifier les termes employés et de les définir de façon consensuelle et cohérente tout en préservant la diversité langagière.

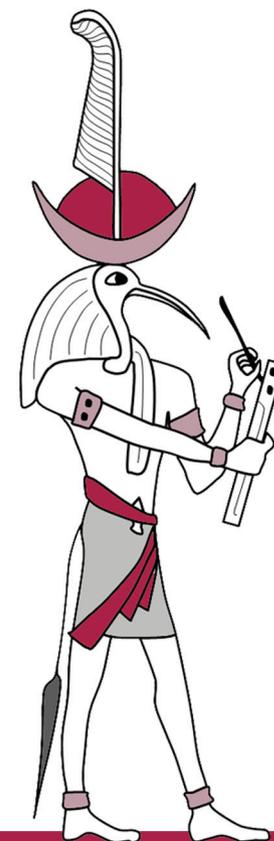
La terminologie, en tant que discipline scientifique, se fonde sur une conceptualisation d'un domaine et sur les mots pour en parler. Elle se doit donc de concilier un point de vue linguistique et un point de vue ontologique. Elle doit également, dans une société numérique où les connaissances constituent la principale richesse, pouvoir être opérationnalisée à des fins de traitement de l'information.

Les conférences TOTh se situent dans le prolongement des colloques annuels de la Société française de terminologie organisés en décembre à Paris (Ecole normale supérieure de la rue d'Ulm). Planifiées à mi-parcours, au mois de juin à Annecy (Polytech'Savoie), elles complètent l'offre et proposent des conférences avec appel à communications, comité de lecture et publication des actes.

Les conférences TOTh ont pour objectif de rassembler industriels, chercheurs, utilisateurs et formateurs dont les préoccupations relèvent à la fois de la terminologie et de l'ontologie et, de façon plus générale, de la langue et de l'ingénierie des connaissances. Elles se veulent un lieu d'échange et de partage où sont exposés problèmes, solutions et retours d'expériences tant sur le plan théorique qu'applicatif ; ainsi que les nouvelles tendances et perspectives des disciplines associées : terminologie, langues de spécialité, linguistique, intelligence artificielle, systèmes d'information, ingénierie collaborative, etc.

Christophe Roche, Président du Comité Scientifique

TOTh 2008
Terminologie & Ontologie : Théories et Applications



TOTh 08

Terminologie & Ontologie : Théories et Applications

<http://www.porphyre.org>

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



9 782951 645394

ISBN 978-2-9516-4539-4
EAN 9782951645394



Publications précédentes

TOTh 2007

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

Commandes à adresser à : toth@porphyre.org

Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2008

Annecy – 5 & 6 juin 2008



à l'initiative de :

- l'Institut Porphyre « Savoir et Connaissance »
- la Société française de terminologie
- l'Université de Savoie
- l'Université de Sorbonne nouvelle

avec le soutien de :

- l'Ecole d'Ingénieurs Polytech'Savoie
- l'association EGC (Extraction et Gestion des Connaissances)
- ISKO France
- la société Ontologos corp.

Titre : TOTh 2008. *Actes de la deuxième conférence TOTh - Annecy – 5 & 6 juin 2008*

Editeur : Institut Porphyre, *Savoir et Connaissance*

<http://www.porphyre.org>

Annecy, 2008

ISBN 978-2-9516-4539-4

EAN 9782951645394

© Institut Porphyre, *Savoir et Connaissance*



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
Anne Dourgnon-Hanoune	EDF R&D
François Gaudin	Professeur, Université de Rouen
Jean-Yves Gresser	ancien Directeur à la Banque de France
Olivier 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
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 Marseille

Comité d'organisation : Responsable Luc Damas

Samia Chouder, Joëlle Pellet

Avant propos

à la mémoire de notre collègue et ami le professeur Henri Zinglé



Nombreux sont les signes qui traduisent la nécessité de considérer la terminologie comme une discipline scientifique à part entière. Il suffit d'être à l'écoute des besoins des industriels et des préoccupations des chercheurs, de constater les mutations d'une société numérique où la connaissance tient une place prépondérante. L'intérêt porté aux conférences TOTh l'illustre bien.

Cela nous a amenés, dès la seconde édition de TOTh, à poser certains fondements pour asseoir et rythmer notre communauté qui rassemble diverses disciplines, de la linguistique à l'ingénierie des connaissances.

C'est avant tout une unité de lieu. Afin que chacun puisse s'appropriier le lieu de nos rencontres, les conférences TOTh se déroulent à Annecy. Au-delà de la beauté de la ville, on s'imagine prolonger les débats le long de ses rues médiévales et canaux devenus familiers.

C'est aussi une unité de temps. Les conférences se déroulent, chaque année, le premier jeudi et le premier vendredi du mois de juin. Vous pouvez d'ores et déjà l'inscrire sur vos tablettes.

C'est enfin, et peut-être surtout, le respect de la différence. Le comité de programme de TOTh 2008 s'est considérablement enrichi afin de traduire le nécessaire caractère pluridisciplinaire de la terminologie. Les conférences TOTh se veulent un lieu d'échange où chacun, dans sa diversité, peut s'exprimer et être respecté, un lieu de partage où les résultats de nos travaux sont accessibles¹ à tous. Terminons en citant de mémoire Saint Exupéry : « si tu diffères de moi, mon frère, loin de me léser, tu m'enrichis ».

Christophe Roche
Président du Comité Scientifique

¹ Les actes sont librement téléchargeables à l'adresse suivante :
<http://www.porphyre.org/toth>

Table des matières

THEORIES	1	APPLICATIONS	127
<i>De la typologie à l'ontologie de textes</i> Rute Costa, Raquel Silva	3	<i>TA statistique : petits corpus pour de petits sous-langages</i> Najeh Hajlaoui, Christian Boitet	129
<i>Réseau terminologique versus Ontologie</i> Sylvie Despres, Sylvie Szulman	17	<i>La place de la modélisation sémantique dans la méthodologie d'entreprise</i> Dominique Vauquier	149
<i>Pragmaterminologie : les verbes et les actions dans les métiers</i> Dardo de Vecchi, Laurent Estachy	35	<i>Modélisation des connaissances métiers du domaine de la génétique humaine en situation d'aménagement terminologique</i> Josée Di Spaldro, Pierre Auger, Maryvonne Holzem, Jacques Ladouceur	173
<i>Faut-il revisiter les principes terminologiques ?</i> Christophe Roche	53	<i>Une terminologie normée pour la maintenance des moyens de production hydrauliques</i> Anne Dourgnon-Hanoune, Philippe Rouard, Marie Calberg-Challot	197
<i>Propositions pour un réseau conceptuel des instruments de mesure œnologiques</i> Pierre Lerat	73	<i>Ontologie franco / anglaise du domaine informatique comme accès à un corpus de textes scientifiques</i> Gérald Kembellec	213
<i>Interrogations sur l'évolution du français de la gestion</i> Odile Challe	91	<i>Gestion des connaissances en médecine des assurances : modèle de représentation des connaissances et application technique</i> Juerg P. Bleuer, Kurt Bösch, Christian A. Ludwig	233
<i>Comment modéliser des concepts en rapprochant un langage orienté objet et deux normes terminologiques orientées concept ?</i> Hendrik J. Kockaert, Bassey E. Antia	105	<i>Mise en évidence de la sémantique dans un système d'aide au diagnostic des plaintes écrites associées à des situations de pollution de l'air intérieur</i> Zoulikha Heddadji, Nicole Vincent, Séverine Kirchner, Georges Stamon	241
		<i>Un témoignage issu d'une expérience professionnelle à la Banque de France et deux suggestions</i> Alain Dequier	259



THEORIES

De la typologie à l'ontologie de textes

Rute Costa, Raquel Silva

Centro de Linguística da Universidade Nova de Lisboa

Avenida de Berna, 26 – C

1069-061 Lisboa – Portugal

rute.costa@fcsh.unl.pt / raq.silva@fcsh.unl.pt

http://www.fcsh.unl.pt/clunl

Résumé :

L'importance du texte pour l'exploitation terminologique n'est aujourd'hui pas à démontrer, il nous semble néanmoins important, au stade actuel des réflexions, de revenir sur certains concepts chers à la théorie de la Terminologie comme ceux de «texte de spécialité» ou bien de «terminologie textuelle», notamment lorsque nous sommes amenés à travailler dans une perspective d'organisation des textes, comme contribution déterminante pour la représentation des connaissances. Mais que cherche-t-on et qu'espère-t-on trouver dans les textes de spécialités ? Les uns disent y trouver des termes, les autres des concepts, ou bien les connaissances elles-mêmes, alors que d'autres disent y trouver des représentations des connaissances. La nécessité de constituer une typologie de textes au sein de l'Assemblée de la République, au Portugal, a conduit les réflexions ci-dessous sur les implications théoriques d'une telle tâche, partant d'abord des caractéristiques du texte de spécialité, dans le contexte du Parlement, afin de constituer une typologie qui puisse ensuite remplir les fonctions d'une ontologie de textes parlementaires associée à une base de données de terminologie juridico-parlementaire.

Mots-clés : texte de spécialité, terminologie textuelle, typologie, ontologie de textes de spécialité, organisation des connaissances.

1. Introduction

La fin du 20^{ème} siècle est indéniablement liée à l'histoire mondiale de la communication. La démocratisation des moyens de communication et la globalisation de l'information ont définitivement contribué à l'augmentation exponentielle du marché de l'information, et non seulement sur le web. À l'image du phénomène Internet, les organisations sont aujourd'hui plus que conscientes des bénéfices et du pouvoir que procure une bonne gestion de l'information et se munissent, à large échelle, de systèmes sophistiqués de gestion et de recherche de

l'information adaptés aux besoins très précis des pratiques institutionnelles ou d'entreprises.

Les systèmes de gestion terminologiques actuels intègrent différents modules de connaissances reliés par des liens d'exploitation visant à la fois la rapidité et la précision, quel que soit le type d'information à portée terminologique recherchée. Après l'accélération technologique et informationnelle, ce sont aujourd'hui des besoins de capitalisation des ressources qui préoccupent les organisations, l'heure étant pour elles de mettre l'accent sur l'organisation des connaissances.

L'Assemblée de la République, au Portugal, est un exemple d'institution qui a mis en place un projet¹, au sein de son service interne de traduction, afin de mieux organiser et diffuser la terminologie produite et utilisée dans le cadre du Parlement. L'implantation d'une Base de Données Terminologique et Textuelle (BDTT) pour l'Assemblée de la République nous a permis, tout au long de sa conception, d'adapter sa mise en place en fonction des réalités d'usage à l'intérieur et à l'extérieur de l'institution, mais également de réfléchir à l'opérationnalisation de certains concepts, tant d'un point de vue méthodologique que théorique.

Dès lors, l'importance des textes au sein du Parlement s'est imposée dans notre démarche, la BDTT devant impérativement respecter leurs statuts et leurs contenus. C'est en partie de cette expérience que sont nées les réflexions suivantes, assumant que l'importance du texte pour l'exploitation terminologique n'est actuellement plus à démontrer, mais qu'il s'avère essentiel de revenir sur certains concepts chers à la Terminologie comme ceux de «texte de spécialité» ou bien de «terminologie textuelle», notamment lorsqu'on travaille dans une perspective d'organisation des textes comme contribution déterminante pour l'organisation des connaissances.

¹ L'Assemblée de la République a signé en 2005 un protocole de recherche avec le Centre de Linguistique de l'Universidade Nova de Lisboa – Ligne de recherche en Lexicologie, Lexicographie et Terminologie, projet qui a donné lieu à la création de la *Base de Données Terminologique et Textuelle de l'Assemblée de la République* (BDTT-AR) : responsable scientifique Rute Costa et collaboration de Raquel Silva.

2. Textes de spécialités écrits

Le texte de spécialité peut, simultanément, être compris comme la production et le produit d'une communauté de communication restreinte. Dans le texte se concentrent tous les éléments linguistiques et extralinguistiques qui résultent de l'interaction du langage avec la vie sociale, ce qui fait que le texte peut être analysé en même temps comme un processus et comme un résultat [Costa, 2006 :80]. Nous irons nous détenir plus longuement sur la description et les caractéristiques du texte de spécialité comme un résultat, étant donné qu'il est un objet d'observation et d'analyse pour ceux qui ont recours aux corpus pour identifier des termes, des concepts et en extraire des connaissances.

Mais que cherche-t-on et qu'espère-t-on trouver dans les textes de spécialités ? Les uns disent y trouver des termes, les autres des concepts, ou bien les connaissances elles-mêmes, alors que d'autres disent y trouver des représentations des connaissances.

Les différences et les relations entre termes et concepts ont été longuement débattues, que se soit du côté de la linguistique, de la cognition, de l'ingénierie ou de l'intelligence artificielle. Et ceci tant du point de vue théorique que du point de vue méthodologique. Mais ces deux réalités ne peuvent pas se substituer au concept de connaissances. Les termes désignent des concepts, qui au sein d'un métier ou d'un domaine, constituent un système ou un réseau conceptuel qui fait partie des connaissances qu'un individu doit dominer pour comprendre et produire des textes de spécialités d'un champ de savoir spécifique.

Pour notre propos, nous entendons par connaissances les *choses* qui sont sues et connues par un individu en tant que membre d'une communauté de spécialité. C'est pour avoir fait preuve du *savoir faire* et/ou du *savoir dire* qu'il est reconnu comme membre d'une communauté, de sa communauté. Ses connaissances lui permettent de produire des textes dont les contenus s'accumulent à d'autres déjà existants. En conséquence, il apparaît aujourd'hui le besoin de gérer des données que l'on retrouve dans ces textes. Néanmoins, pour ce faire, il est indispensable de savoir gérer d'abord les textes en tant qu'objets de connaissances eux-mêmes. En effet, les textes doivent d'abord être organisés comme des contenants, pour ensuite en permettre l'organisation des contenus.

Le texte est le moyen le plus efficace pour le spécialiste de communiquer avec les membres de sa communauté professionnelle. Le texte est le lieu du débat et le lieu de l'organisation des idées ; c'est le lieu de la construction et de la

déconstruction, mais aussi le lieu de l'incertain et de la polémique ; c'est le lieu des propositions et des contre-propositions, des provocations, des réponses, des défenses, des dissuasions, parce que c'est aussi le lieu de l'exposition, du risque et du jugement. En d'autres termes, le spécialiste expose de façon cohérente, en ayant recours aux mots, aux termes et à la grammaire sa vision et sa conception du monde culturellement partagé par un groupe d'individus qui forme sa communauté.

Le lecteur à qui se dirige ce type de texte se trouve à un niveau de connaissance proche de celui de l'auteur, car théoriquement, il est en mesure d'appréhender les contenus et l'intention de ce qui lui est communiqué. Ce nivellement de connaissances a de l'influence dans la façon dont l'auteur écrit son texte : il se gère une espèce de complicité qui amène à une forte présence du *non-dit* qui souvent occupe une place capitale dans le texte, et qui, d'après nous, est une des propriétés les plus caractéristiques du texte de spécialité. Établir la relation entre le *dit* et le *non-dit*, entre l'explicite et l'implicite est une des tâches du lecteur spécialiste, puisque dans toutes les situations de communication, on a tendance à signifier plus que ce que l'on dit.

Il existe forcément des intersections entre ce qu'est l'objet, sa conceptualisation et sa dénomination. Pour transmettre cette relation triangulaire qui reflète ses croyances, ses idéologies scientifiques, ses visions du monde, l'auteur s'efforce à construire son discours, qui pour lui est monoréférentiel, au sein d'un contexte donné. D'ailleurs, dans un cadre de communication de spécialité, l'auteur éprouve le besoin de restreindre le plus possible la diversité des constructions de sens pour se rapprocher idéalement d'un discours qui sera de l'ordre du monosémique. Discours monosémique qui, probablement, ne sera jamais atteint ; de toute manière, son existence serait impossible à prouver.

Le texte de spécialité est indubitablement un véhicule de connaissances et, en Terminologie, le terme y joue un rôle fondamental, étant donné qu'il est un élément nucléaire des nœuds sémantiques que l'on peut repérer dans les textes. Ces nœuds sémantiques correspondent fréquemment aux points nucléaires qui sont à la base de la construction des réseaux sémantiques et qui permettent une représentation des connaissances patentées dans le texte ou dans un ensemble de textes. Néanmoins, persistent les questions suivantes : À partir de quelles entités est-ce que l'on modélise les connaissances ? Est-ce vraiment les connaissances que l'on modélise ? Les réseaux sémantiques correspondent-ils à des systèmes

conceptuels ? Et, finalement, que représente l'entité linguistique que les uns valorisent et que les autres dévalorisent ?

Ces questions nous amènent à repenser le statut de l'unité terminologique qui actuellement ne bénéficie plus d'une attention exclusive. Le terminologue se concentre de plus en plus sur l'établissement du rapport, du lien ou encore de la relation entre deux ou plusieurs concepts au sein d'un même système conceptuel. Cette perspective peut être corroborée par une affirmation de Rey, qui préconise que : « [...] la terminologie étudie des signes. Ces signes se manifestent au moyen des formes des langues naturelles (mots, etc.), leur rapport avec ses formes doit être précisé » [Rey, 1979 : 19]. Or, c'est dans cette notion de rapport que se situe, nous semble-t-il, une partie importante du débat, car c'est dans le rapport entre ce qui est dénommé et la dénomination que se trouve l'essence du travail en terminologie et qui, en définitive, se trouve au cœur des travaux actuellement en cours. Ici, la relation entre le contexte extralinguistique et le contexte linguistique nous semble une évidence.

S'il est vrai que l'organisation des connaissances qui se trouvent à un niveau extralinguistique est un des objectifs primordiaux de la recherche en Terminologie, il n'est pas moins vrai que c'est la plupart des fois à travers l'acte de la parole, c'est-à-dire le discours, que nous pouvons accéder à la représentation de ces connaissances. La parole étant un moyen privilégié de représentation du monde en soi. La difficulté de la théorisation réside justement dans le fait que ces deux réalités – le monde et sa représentation discursive – forment une association durable et réciproquement profitable.

3. De la typologie de textes au corpus

Le besoin de constituer une typologie de textes au sein de l'Assemblée de la République a conduit nos réflexions sur les implications d'une telle tâche. À l'occasion de travaux antérieurs [Costa : 2005] nous avons déjà eu l'occasion de réfléchir sur la question de la typologie qui suppose la réunion et la classification d'un ensemble de textes sous une même étiquette. Pour cela, les textes doivent maintenir entre eux des relations de ressemblance au niveau des macro et des microstructures à travers l'identification de régularités propres à un ensemble de textes, par opposition aux régularités d'autres ensembles de textes.

Ceci dit, une typologie est le résultat de l'organisation des textes en fonction des traits qui les caractérisent et qui leurs sont communs, permettant une

classification. Cette classification permet une répartition systématique des textes en groupes ou en types auxquels sont attribués des étiquettes ou un nom générique. Ce regroupement, qui est toujours artificiel et qui dépend du point de vue du chercheur en fonction des objectifs de sa recherche, peut être de l'ordre du linguistique ou de l'ordre de l'extralinguistique.

Une typologie ne présuppose donc pas une quelconque forme de hiérarchie, de dépendance ou de relation sémantique ou conceptuelle entre les objets qui la composent. Une typologie peut se faire à partir de genres ou de types de textes. Pour Maingueneau, classer les textes en types est une activité de l'ordre du sociologique et non du linguistique, alors que le genre est constitutif de l'action verbale : « *Les genres de discours relèvent de divers types de discours, associés à de vastes secteurs d'activité sociale* » [Maingueneau, 1998 : 47]. Pour l'auteur, l'élaboration des typologies de discours et de textes est pertinente seulement si l'on tient compte du genre, concept fondateur de l'activité verbale : « *Tout texte relève d'une catégorie de discours, d'un genre de discours* » [Maingueneau, 1998:45].

Parler de types de discours signifie établir des paramètres en harmonie avec les différents secteurs de la société, dont chacun produit des discours et des textes qui peuvent être classés par types. La recherche scientifique, par exemple, est un secteur d'activité dont la production textuelle et discursive constitue un type, parce qu'elle est le produit d'une activité sociale spécifique. Ainsi, nous considérons que l'élaboration des typologies de types, ainsi que des typologies des genres résulte de l'observation des conditions sociodiscursives dans lesquelles le texte a été produit, texte qui est le témoin représentatif d'une collection de textes qui, dans son ensemble, caractérise le discours.

Un corpus de textes d'un domaine spécifique se compose idéalement de textes qui correspondent à une organisation typologique qui a pour but d'instaurer une certaine forme de représentativité ; la représentativité non au sens statistique, mais au sens de l'acceptation du texte en tant que production scientifiquement reconnue par les membres qui composent la communauté scientifique ou professionnelle, dans laquelle et par laquelle le texte a été produit. Car, ce n'est qu'en fonction de l'établissement de ces critères qu'il est possible de garantir l'adéquation des textes aux objectifs préétablis qui sont évidemment le garant de tout travail de recherche. Là encore, une compétence additionnelle s'ajoute à celles déjà requises aux terminologues ; statuer sur le texte de spécialité, en n'oubliant pas de réfléchir sur le statut des intervenants – auteur et locuteur – ainsi que sur le contexte de production et de réception.

Nous le savons, les corpus sont constitués par des collections de textes qui ont été sélectionnés et recueillis pour que l'usage de la langue, c'est-à-dire, le discours puisse être étudié en ayant recours au traitement informatique. D'après [Wynne, 2005] : « *Aujourd'hui les linguistiques de corpus offrent quelques-unes des procédures les plus puissantes pour l'analyse de la langue, et l'impacte de cette sous discipline dynamique et son expansion se fait sentir dans plusieurs domaines de l'étude du langage* ». La question que nous nous posons est de savoir quelles sont les relations entre la linguistique de corpus et la terminologie textuelle ?

Il faut, pour ce faire, nous rappeler que le recours aux textes en Terminologie n'a pas toujours été une évidence et que de longues discussions se sont tenues autour des méthodologies utilisées en Terminologie : identifier, sélectionner ou construire des terminologies en ayant recours à une méthodologie onomasiologique ou à une méthodologie sémasiologique ? Au centre de ce débat, semble se trouver la valeur que le texte peut assumer en Terminologie.

4. Linguistique de corpus et terminologie textuelle

Nous sommes d'avis, qu'aujourd'hui, la question sur la pertinence du texte pour l'exploitation terminologique n'est plus à démontrer, le texte a une valeur confirmée. Mais, l'acceptation du travail sur les textes en Terminologie nous amène à réfléchir sur la notion de Terminologie textuelle, qui a pour base le texte et ses méthodologies sous-jacentes : « *Les applications de la terminologie sont le plus souvent des applications textuelles (traduction, indexation, aide à la rédaction); la terminologie doit «venir» des textes pour mieux y «retourner». C'est parce qu'elle n'est jamais déliée du texte qu'on parle de «terminologie textuelle». C'est dans les textes produits ou utilisés par une communauté d'experts, qui se sont exprimées, et donc accessibles, une bonne partie des connaissances partagées de cette communauté, c'est donc par là qu'il faut commencer l'analyse.*» [Bourigault & Slodzian, 1999:30].

Effectivement, quand on travaille sur les textes pour obtenir de l'information à travers l'extraction automatique, le but est de pouvoir à tout moment revenir sur le texte. Nous partons du texte, pour à travers lui, traiter, organiser et structurer les termes, c'est-à-dire que nous nous séparons de lui, pour plus tard dans notre parcours revenir à lui. C'est cet aller-retour au texte qui est nécessaire à notre travail, qui incite le besoin de la normalisation des étiquetages de tout ordre, car la réutilisation des données est un fait indéniable, car nous savons tous qu'à un moment ou à un autre nous allons très probablement recourir à des corpus organisés par d'autres et à céder les nôtres.

Si le problème sur l'importance du texte en Terminologie ne se pose désormais plus, puisqu'il l'est, la difficulté essentielle est surtout de savoir ce que l'on cherche dans les textes et ce que l'on espère y trouver. D'après [Slodizan, 2006], la Terminologie textuelle s'intéresse au fonctionnement des signifiés dans les textes à caractère technique et scientifique, un courant donc plus linguistique, qui se distingue d'une perspective plus conceptuelle.

Mais nous pensons que parler de Terminologie textuelle est plus une question de méthodologie que de théorie ; une méthodologie de travail qui aurait pour base les méthodologies propres aux linguistiques de corpus, mais appliquées aux textes de spécialités. Ce qui est, selon nous, spécifique à la Terminologie textuelle, c'est le regard porté sur les textes qui, en Terminologie, présupposent une approche extralinguistique importante.

Les approches extralinguistiques nous permettent de distinguer les textes les uns des autres en fonction de leurs statuts qui leurs sont essentiellement conférés par :

1. la reconnaissance scientifique de l'auteur par la communauté à laquelle il appartient ;
2. la connaissance du public auquel se dirige le texte ;
3. la représentativité du texte pour les membres de la communauté scientifique.

Lorsqu'un texte correspond à ces paramètres, il est considéré un texte de spécialité de pleins droits. C'est idéalement à partir de ces textes que nous cherchons de l'information terminologique, basé sur des critères essentiellement sémantiques qui, néanmoins peuvent refléter le conceptuel.

Les réflexions que nous venons d'exposer sont le fruit de nos préoccupations en tant que terminologues chargés de développer un projet de Base de Données Terminologique et Textuelle au sein de l'Assemblée de la République Portugaise et où l'importance assumée par les textes a forcément dirigé nos réflexions théoriques, d'abord sur les caractéristiques du texte de spécialité, dans ce contexte dit parlementaire, afin de constituer une typologie qui puisse ensuite remplir les fonctions d'une ontologie de textes parlementaires associée à une base de données de terminologie juridico-parlementaire².

² Voir Mémoire de DEA (Mestrado): Zara Almeida (2008) Terminologia jurídico-parlamentar, Combinatórias terminológicas e Equivalência na Base de Dados Terminológica e textual da Assembleia da República – BDTT-AR, FCSH-UNL, Lisboa.

5. Vers une ontologie de textes parlementaires

Le Parlement portugais est constitué par une seule assemblée, désignée Assembleia da República, qui est considérée d'après la loi fondamentale l'assemblée représentative de tous les citoyens portugais. Outre cette fonction de représentativité, le Parlement se doit d'assurer l'approbation des lois fondamentales de la République ainsi que de veiller à l'application des principes de la Constitution, des lois et des actes du Gouvernement et de l'Administration.

Le Parlement est donc ainsi perçu par tous comme un organe de pouvoir décisionnel, véhiculant son autorité à travers la production, l'approbation et la diffusion de ses textes, lui permettant de créer un cadre conceptuel de référence capable de poser les limites de ce qui peut ou ne peut être considéré comme « constitutionnel ».

Ses textes fondateurs comme la Constitution, le Règlement (*Regimento*) et le Statut des Députés définissent les compétences et les règles de fonctionnement du Parlement ainsi que les droits et les devoirs de ses membres, garantissant les relations de séparation de pouvoirs et d'interdépendances relativement aux autres entités de souveraineté. La Constitution énumère les matières dans lesquelles la loi peut intervenir en fixant des règles ou des principes fondamentaux. Elle délimite ainsi le domaine de la loi, en précisant les modalités d'application de ces règles et principes.

Ces textes parlementaires font autorité dans la mesure où l'institution qui les produit leur confère bien un statut de référence. Par exemple, la Constitution véhicule des concepts-clés comme « démocratie » ou bien « liberté », en délimite le champ d'application et les frontières conceptuelles, rendant ainsi possible la reconnaissance et l'adoption de ces concepts dans d'autres textes parlementaires qui font autorité et dont l'on peut parfois même entendre dire que l'un ou l'autre est « anticonstitutionnel ». C'est-à-dire qu'il n'est pas en accord avec le cadre conceptuel de la Constitution et qu'il en sort du domaine d'application, pouvant aller jusqu'à remettre en cause la propre conception de démocratie, dont le cadre légal est stipulé par le texte de la Constitution.

Construire une Base de Données Terminologique et Textuelle de l'Assemblée de la République présupposait tenir compte de ce cadre référentiel et en montrer le fonctionnement à travers l'organisation conceptuelle de la

terminologie produite et utilisée au sein du Parlement mais, également, tenir compte du statut et de l'importance des textes.

Cette méthodologie n'a été possible qu'après un long travail de compréhension de ce cadre référentiel spécifique à l'Assemblée de la République. Nous nous sommes intéressés non seulement au fonctionnement interne du Parlement, mais également à l'ensemble de la procédure législative, c'est-à-dire le parcours de l'élaboration d'une loi, désigné en français par « navette parlementaire ».

La prise en compte de la diversité des textes produits par l'institution s'est dès lors imposée comme une méthodologie incontournable face à la richesse des informations à disposition pour l'organisation des connaissances produites et diffusées en contexte parlementaire. La capitalisation de ces connaissances étant, du point de vue de l'Assemblée de la République, « *une condition essentielle pour un exercice plus éclairé du droit à la citoyenneté* »³.

L'organisation des textes s'est opérée par phases, soumettant le résultat à chaque nouvelle étape à la validation d'un groupe de spécialistes constitué principalement par des juristes de l'Assemblée. La première liste de textes a servi à identifier la diversité des textes utilisés et produits, donnant ainsi naissance à une typologie de textes parlementaires ; la suivante proposait aux spécialistes une organisation par genres de textes (proposition de loi ; projet de loi ; ordonnances ; etc.) et par ordre d'importance. Cette seconde version présupposait de la part des spécialistes une validation du cadre référentiel du fonctionnement du Parlement à partir de l'organisation textuelle. Des critères de dépendance entre les textes et d'importance du statut du texte ont été introduits comme facteurs de validation.

Le résultat a donné lieu à ce que nous avons désigné par ontologie de textes de l'Assemblée de la République. Cette ontologie sera implantée très prochainement en association avec la base de données terminologique déjà existante et consultable sur le site du Parlement portugais depuis le mois de mai 2007⁴.

³ Site de l'Assemblée de la République: www.parlamento.pt

⁴ <http://terminologia.parlamento.pt/pls/TER/terwinter.home>

6. Remarques et conclusions

Il s'agit ici d'apporter quelques éléments supplémentaires de réflexion, en guise de conclusion, à propos de certaines notions qui nous semblent essentielles, commençant par la question des ontologies. Depecker et Roche [2007 :112] nous parlent de deux grands types de construction d'ontologies : la première est désignée d'ontologie «lexicale» qui dégage de l'analyse du discours et la deuxième qu'ils désignent «[...] d'ontologies conceptuelles, issues d'une conceptualisation des objets du monde que partagent une communauté de pratique. C'est ce que met en valeur l'approche conceptuelle de la terminologie.»

De notre point de vue, ces deux approches ne sont pas forcément antagoniques, elles peuvent être complémentaires. Si les textes sont des objets de connaissances et si leurs organisations, essentiellement par des relations de dépendances, présupposent des connaissances spécifiques de la part de la communauté parlementaire par exemple, alors le résultat que l'on obtient peut être considéré comme une ontologie de textes. Nous pensons bien à une ontologie de texte et non pas à une ontologie textuelle ; cette dernière pouvant être confondu avec l'élaboration d'ontologies à partir des contenus des textes et être, éventuellement, associé à la Terminologie textuelle et donc plus proches des «ontologies lexicales».

Mais nous pensons que cette méthodologie ne peut pas être appliquée à tous les domaines et que l'élaboration des ontologies de textes n'est pas toujours possible, car le texte écrit n'a assurément pas la même valeur dans toutes les communautés professionnelles. La manipulation des textes implique des connaissances extralinguistiques du domaine très pointues. C'est pourquoi nous revenons une fois de plus à la problématique de la méthodologie tellement chère aux terminologues et qui a provoqué au long des temps des discussions théoriques très effusives : l'approche onomasiologique ou l'approche sémasiologique ? D'un côté la perspective conceptuelle, de l'autre la perspective linguistique.

En tant que terminologue, nous avons la plupart des fois une approche conceptuelle, dans le sens où nous commençons toujours par observer l'usage et la façon dont la communauté professionnelle s'approprie de la langue dans un contexte donné ; comment elle rédige les textes et par quelles voies les textes sont diffusés. Cette approche nous amène à organiser des textes soit en typologies soit en ontologies de textes, selon le statut que le texte acquière au sein d'une activité ou d'un domaine.

L'approche sémasiologique, se situe généralement au niveau de l'analyse des textes et ne se réalise qu'après la réalisation de la phase antérieure. Ce n'est qu'à ce stade que nous pensons qu'il vaille la peine de recourir à l'analyse du discours pour identifier des termes ou encore des entités linguistiques à valeurs spécialisées, car nous sommes d'avis que mieux on recueille et organise les textes en fonction d'un objectif précis, plus la qualité qui résulte de l'analyse des contenus est garantie. Cependant, un corpus ne sert pas toutes les fins. Ce travail extralinguistique peut être à refaire pour chaque nouvelle réalité.

Il est vrai que la dénomination retient l'attention du terminologue, mais il se concentre surtout sur l'identification et la classification du concept et, postérieurement, sur les relations et les rapports que celui-ci maintient avec d'autres concepts au sein d'un même système conceptuel.

Quand le texte ou le discours sont le point de départ, c'est souvent par le biais des dénominations que le terminologue essaye d'accéder et d'appréhender les concepts, étant donné que ce sont elles qui nous permettent de construire des représentations possibles du monde, à travers l'acte de la parole. Mais prenons garde, car comme l'affirment Depecker et Roche [2007 :112] : «[...] les analystes ont tendance à superposer, dans un corpus de textes, structure conceptuelle et structure linguistique. Faisant cela, ils voient rarement que la structure informationnelle d'un discours, d'ordre linguistique, ne recouvre pas la structure conceptuelle du monde, d'ordre scientifique. En effet, pratiques langagières et pratiques scientifiques ne sont pas du même ordre.»

Les relations entre les concepts et les termes viennent augmenter la difficulté de l'établissement des frontières entre ce qui est du niveau du linguistique et ce qui est du niveau des connaissances. D'après Rastier [1995 :43-44], pour construire des systèmes de représentation des connaissances, il est essentiel d'agir, en simultané, sur l'état des choses, les réseaux conceptuels et les structures sémantiques. Ceux-ci sont, du point de vue de l'analyse, les trois niveaux distincts de la représentation des connaissances qui permettent, à travers des modèles formels, d'activer les systèmes informatiques en mettant en correspondance les lois de la nature, la pensée et le langage.

Les communautés qui travaillent en Terminologie ont en générale recours aux textes. En 1995, Lerat établissait une différence fine entre «langue de spécialité» et «langue spécialisée», qui laisse présumer l'existence d'une «langue en contexte de spécialités» qui est finalement le terme le plus adéquat car il repose sur le concept d'usage, de discours, dont le texte est une occurrence.

Beaucoup de travaux théoriques ont été dédiés à la terminologie, aux différentes approches théoriques, mais peu de travaux ont été voués aux caractéristiques intrinsèques aux textes de spécialités. D'un point de vue théorique, qu'est-ce qui nous permet de distinguer des textes de spécialités des autres textes ?

Comme le texte est souvent à la base de tout travail terminologique et de gestion de l'information, il serait important de savoir si on peut y trouver ce que l'on y cherche. Souvent en y trouve ce que l'on veut y voir, non ce qui y est.

Bibliographie

- Almeida, Z. 2008. *Terminologia jurídico-parlamentar, Combinatórias terminológicas e Equivalência na Base de Dados Terminológica e textual da Assembleia da República – BDTT-AR, Mémoire de DEA (Mestrado), FCSH-UNL, Lisboa.*
- Bourigault, D. & Slodžian M. 1999. «Pour une terminologie textuelle». *Terminologies Nouvelles*. N° 19. Bruxelles : Rint, pp. 29-32.
- Costa, R.. 2005. «Corpus de spécialité : une question de types ou de genres», *De la mesure dans les termes. Hommage à Philippe Thoiron*. (eds. Henri Béjoint & François Maniez]. Lyon : PUL, pp. 313 – 224.
- Costa, R.. 2006. «Texte, terme et contexte», *Actes des VIIes Journées scientifiques du Réseau Lexicologie, Terminologie et Traduction. "Mots, Termes et contextes". Daniel Blampain / Philippe Thoiron / Marc Van Campenhoudt [ed.] Paris: Editions des archives contemporaines, pp. 79 – 88.*
- Depecker L, Roche C. 2007. «Entre idée et concept : vers l'ontologie», in *Langages. Genèses de la terminologie contemporaine (sources et réceptions)*. N°168. Décembre. Paris : Armand Colin.
- Lerat, P. 1995. *Les langues spécialisées*, Paris : PUF.
- Maingeneau, D. 1998. *Analyser les textes de communication*. Paris : Dunod.
- Rey, A. 1979. *La Terminologie. Noms et Notions*, Coll. *Que sais-je ?*, Paris, PUF.
- Rastier F. 1995. «Le terme: entre ontologie et linguistique», in *Banque des Mots*. Paris : CILF, pp. 35 – 65.
- Slodžian, M. 2006. «La terminologie, historique et orientations». *17èmes journées francophones d'Ingénierie des connaissances*. Nantes. Disponible en ligne :

http://www.sdc.2006.org/cdrom/contributions/Slodzian_SDC2006.pdf [Consulté : 2008 - 05-15]

Wynne, M. 2005. *Developing Linguistic Corpora : a Guide to Good Practice*. Oxford: Oxford Bookes. Disponible en ligne: <http://ahds.ac.uk/linguistic-corpora> [consulté: 2008-05-15]

Réseau terminologique *versus* Ontologie

Sylvie Despres, Sylvie Szulman

LIPN – Université Paris 13

99 av J.B. Clément

93430 Villetaneuse

{*prenom.nom*}@lipn.univ-paris13.fr

<http://www-lipn.univ-paris13.fr/~{nom}>

Résumé :

Les méthodes de construction d'ontologies à partir de textes comportent une phase de conceptualisation au cours de laquelle s'effectue le passage du terme au concept. Dans ce papier, nous montrons comment se construit ce passage du terme au concept *via* des étapes se situant sur le plan linguistique, termino-ontologique et ontologique permettant ainsi l'articulation entre l'expression des connaissances en corpus *via* la langue et leur expression formelle *via* un langage de représentation des connaissances. La spécificité des ressources obtenues à chaque pas de la conceptualisation permet de distinguer clairement leurs usages et leur complémentarité.

Mots-clés : réseau terminologique, ontologie, conceptualisation

1. Introduction

Les méthodes de construction d'ontologies à partir de textes comportent une phase de conceptualisation [Cimiano, 2006; Aussenac et al., 2008] au cours de laquelle s'effectue le passage du terme au concept.

Cette phase de conceptualisation nécessite des traitements se situant à la fois aux plans linguistique et ontologique. Pour faciliter la présentation, nous distinguons trois grandes étapes : l'étude linguistique du corpus qui aboutit à la construction d'un réseau terminologique, l'étape termino-ontologique qui permet de construire un réseau termino-ontologique et l'étape ontologique dont la finalité est l'élaboration de l'ontologie.

L'étude linguistique permet d'organiser les unités linguistiques *via* un ensemble de réseaux terminologiques partiels qui traduisent la structure lexicale des textes. Chacun de ces réseaux terminologiques est centré sur l'étude d'un

terme représentatif du domaine. Cette étude est réalisée à partir des résultats des traitements obtenus par des outils de traitement automatique des langues (TAL).

L'exploitation des résultats issus de l'étude linguistique en vue de la construction de l'ontologie est pour une grande partie réalisée manuellement.

Un réseau termino-ontologique est construit à partir de l'interprétation, dans un cadre applicatif, des unités linguistiques (termes et relations lexicales les liant) constituant le réseau terminologique. Il est constitué d'unités dites termino-ontologiques (concepts terminologiques et relations sémantiques les liant).

Enfin les concepts de l'ontologie et les relations conceptuelles les associant sont construits à partir des unités termino-ontologiques figurant dans les réseaux termino-ontologiques. Ces concepts ontologiques sont décrits dans un langage formel, organisés dans une structure hiérarchique, liés par des relations conceptuelles et contraints par des règles et des axiomes.

L'articulation entre ces deux plans linguistique et ontologique constitue une tâche difficile puisqu'il s'agit de mettre en correspondance d'une part, des unités linguistiques caractérisées par leur polysémie, la diversité de leur emploi et d'autre part des concepts qui sont univoques et normés par le contexte de l'application. Par conséquent, il ne peut y avoir d'isomorphisme entre les deux structures. Le réseau termino-ontologique permet la transition entre les plans linguistique et ontologique.

Cette contribution insiste sur le fait, que la phase de conceptualisation ne relève pas de l'extraction de connaissances, mais consiste en une construction assurant le passage du terme au concept et permettant l'articulation entre l'expression des connaissances en corpus *via* la langue et leur expression formelle *via* un langage de représentation des connaissances.

Après avoir précisé le sens des notions utilisées, les différences entre les notions de réseau terminologique et ontologie sont clarifiées. Enfin, le passage du plan linguistique au plan ontologique réalisé dans la phase de conceptualisation est explicité. Un exemple dans le domaine juridique est utilisé pour illustrer notre propos. Enfin, l'intérêt de distinguer ces différents plans est analysé.

2. La phase de conceptualisation

Dans un schéma faisant maintenant référence (cf. figure 1), Cimiano [2006] fait apparaître les entités (termes, synonymes, concepts, hiérarchies de concepts, relations entre concepts autres que hiérarchiques, les axiomes et les règles relatifs

aux concepts représentés) nécessaires à la construction d'une ontologie à partir de textes.

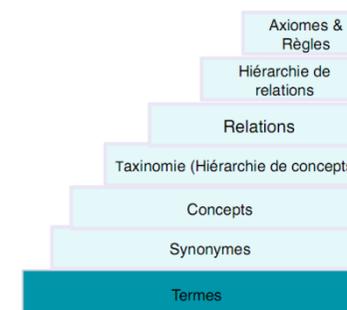


Fig. 1 : « Ontology learning layer cake » d'après Cimiano

A la lecture de ce schéma, il apparaît clairement que la nature des éléments utiles pour construire l'ontologie ne se situent pas sur le même plan. Construire l'ontologie revient à tirer parti de l'ensemble des résultats obtenus à partir des traitements réalisés sur les termes et les relations lexicales qui les lient dans une phase dite de conceptualisation.

La phase de conceptualisation est celle de la construction des concepts ontologiques à partir des résultats issus de l'analyse automatique des textes.

Elle est structurée en quatre étapes : - extraction des entités linguistiques et construction d'un réseau terminologique ; - élaboration de concepts termino-ontologiques et d'un réseau terminologique partiel ; - élaboration des concepts ontologiques et d'un réseau ontologique.

Ces différentes étapes permettent la transition du plan du discours au plan ontologique. Ce passage est réalisé en réalisant des traitements qui relèvent de la terminologie puis de la modélisation des connaissances et enfin de la représentation des connaissances. Nous avons figuré les différents plans dans lequel ce processus intervient dans le schéma extrait de [Mondary et al., 2008] figurant figure 2.

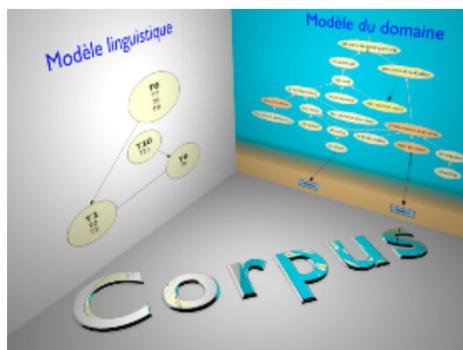


Fig. 2 : Articulation des plans textuel, linguistique et conceptuel

Passer des textes à l'ontologie, c'est donc passer du discours pris comme réalisation linguistique au modèle ontologique en passant par un plan termino-ontologique exprimé *via* un modèle conceptuel du domaine. Le modèle linguistique joue un rôle charnière entre le texte et le modèle conceptuel tandis que le modèle conceptuel assure le même rôle entre le modèle linguistique et le modèle ontologique. L'articulation de ces différents modèles requiert la construction de différents types de ressources qui sont caractérisées par la finalité du plan où elles sont situées.

3. Les notions utilisées

Afin de fixer l'usage des notions utilisées dans ce papier, nous rappelons les définitions que nous leur assignons et nous les illustrons sur un exemple extrait d'un travail dont la finalité était la construction d'une ontologie du domaine pour faire de la recherche d'information [Despres *et al.*, 2007].

Les définitions proposées dans ce paragraphe sont associées aux notions qui interviennent dans les grandes trois étapes du processus de conceptualisation. Elles sont certainement discutables, mais elles nous permettent toutefois de mieux dégager la nature des notions sur lesquelles nous travaillons.

3.1. Une présentation succincte situant l'exemple

Le domaine évoqué est celui de l'organisation de la gestion de l'hygiène, la sécurité et l'environnement (HSE) des entreprises. Le corpus de référence est

constitué d'un ensemble de textes réglementaires concernant les installations industrielles classées. Les connaissances du domaine font référence à trois contextes distincts : - le contexte métier, chaque installation relève d'un métier (chaudronnerie, peinture, ...) ; - le contexte environnement, chaque entreprise a un impact différent sur l'environnement (nuisances au niveau du bruit, de l'air, ...) ; - le contexte juridique, le cadre juridique à appliquer dépend du niveau de dangerosité de l'entreprise (risque SEVESO). Les notions servant à l'exemple sont celles de bruit et installation.

L'ontologie à construire devait aider à capitaliser les savoir-faire des experts chargés de réaliser les audits auprès de leurs entreprises clientes. Dans ce contexte, le rôle de ces experts est de déterminer, à partir des caractéristiques d'une entreprise, les éléments qui lui permettent d'être en conformité avec la législation.

3.2. Les notions intervenant dans l'étude linguistique

L'étude linguistique permet d'établir une analyse terminologique du texte guidée par l'objectif de construction d'un modèle du domaine étudié.

Définitions

Un candidat terme est un mot ou une séquence de mots susceptibles d'être retenus comme terme par un terminologue [Bourigault *et al.*, 2000] ou par un analyste et de fournir les étiquettes des concepts [Bourigault *et al.*, 2003].

Une relation lexicale est une relation entre termes (hyponymie, hyperonymie, holonymie, méronymie, synonymie, antonymie, causalité, etc.) [Cruse, 1986].

Un réseau terminologique est un ensemble de candidats termes reliés entre eux par des relations lexicales [Bourigault *et al.*, 2005].

Exemple

La figure 3 présente un réseau terminologique partiel établi à partir des termes *bruit* et *installation* qui désignent des notions centrales du domaine de la gestion HSE.

Dans l'extrait de texte sélectionné dans la figure 1, les candidats termes retenus sont ceux relatifs à la notion de *bruit* {bruit ambiant, bruit résiduel, bruit

aérien, bruit solidien, niveau de bruit, niveau de bruit ambiant, niveau de bruit global} et à la notion d'*installation* {installation, bruit de l'installation, installation en fonctionnement, hors fonctionnement de l'installation, installation à l'arrêt}.

Des relations d'hyponymie lient les candidats termes {bruit ambiant, bruit résiduel, bruit aérien, bruit solidien} au terme bruit. En outre, le texte permet d'établir une relation d'antonymie entre les candidats termes {bruit ambiant, bruit résiduel} et une relation de synonymie entre les termes {mesure de bruit et mesure de niveau de bruit}. Ces deux derniers termes suggèrent une relation fonctionnelle entre les notions de bruit et niveau de bruit.

Le terme {émergence} est également sélectionné car il est décrit comme une fonction du bruit ambiant et du bruit résiduel.

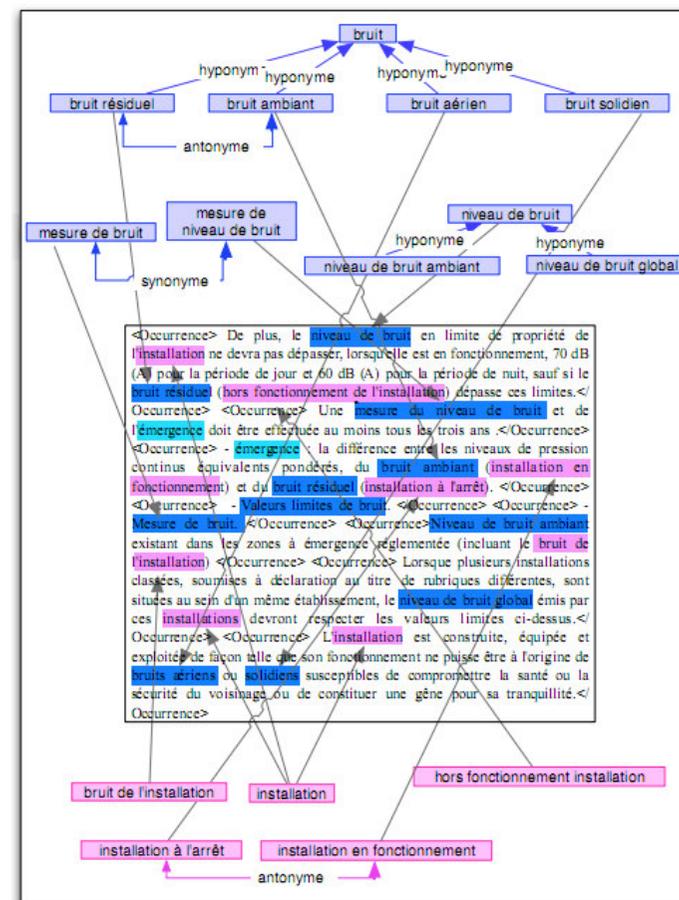


Fig. 3 : réseau terminologique partiel associé au terme bruit

3.3. Les notions associées à l'étape termino-ontologique

L'étape termino-ontologique assure une transition vers le plan ontologique et consiste à construire un modèle conceptuel à partir du réseau terminologique. Les notions de concept terminologique, relation et réseau terminologique utilisées sont introduites *infra*.

Définitions

Un concept terminologique est un terme désambiguïsé c'est-à-dire un terme ayant un sens unique dans un contexte d'interprétation [Biebow et *al.*, 1999]. Il est décrit par des propriétés structurelles et définitionnelles [Després et *al.*, 2005].

Une relation termino-ontologique est une relation entre concept terminologique (inclusion, identité, disjonction, partie-tout, etc.).

Un réseau termino-ontologique est un ensemble de concepts terminologiques reliés entre eux par des relations sémantiques.

Dans cette étape, une transition s'opère. Les termes candidats jugés pertinents servent de point de départ à la construction du modèle. Les relations lexicales qui les lient sont interprétées. Des choix sont à effectuer pour décider ce qui sera concept terminologique ou relation termino-ontologique.

Exemple

La figure 4 présente l'interprétation des unités linguistiques (candidats termes et relations lexicales) décrites dans la figure 3 dans le contexte de la gestion HSE des entreprises.

Les candidats termes « bruit », « niveau de bruit », « mesure de bruit », ... et « installation » sont organisés dans une représentation mettant à jour les relations qui les lient.

Les relations d'hyponymie reliant les candidats termes (tête, expansion) y sont traduites par des relations de subsumption (lien d'inclusion).

Par exemple, « bruit ambiant » est_une_sorte_de « bruit », « bruit résiduel » est_une_sorte_de « bruit ». La différence entre ces deux notions est due à l'état de l'installation (en fonctionnement ou hors fonctionnement). Entre outre, les notions désignées par ces deux termes servent au calcul de l'émergence. Ils sont de bons candidats au statut de concept terminologique.

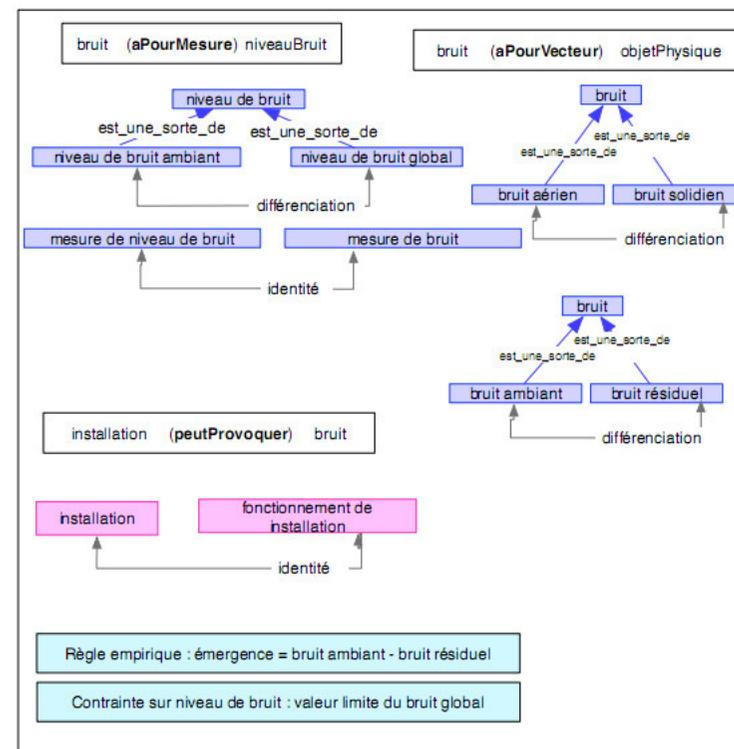


Fig. 4 : liens entre les concepts terminologiques

La différence entre les notions désignées par les deux candidats termes « bruit aérien » et « bruit solide » tient à l'origine du bruit (air ou solide). Une relation afférente au bruit peut être définie. La relation fonctionnelle « aPourVecteur » liant « bruit » à « objetPhysique » introduit un concept qui n'est pas issu du texte mais qui est utile à la définition de la relation. Cette relation peut ensuite être spécialisée afin de décrire le concept terminologique « bruit aérien » (respectivement « bruit solide »). Un « bruit aérien » « aPourVecteur » « air » (respectivement « solide »).

Les relations de synonymie sont interprétées comme des relations d'identité. Il reste alors à décider du terme qui désignera le concept terminologique associé.

Le candidat terme « mesure de niveau de bruit » synonyme de « mesure de bruit » est décrit par une relation fonctionnelle « aPourMesure » liant les concepts terminologiques « bruit » et « niveauBruit ».

Dans le contexte de notre application, la notion de bruit y est envisagée selon le point de vue de la réglementation. Le bruit est un phénomène acoustique qui peut provoquer une nuisance. Une nuisance constitue un processus dommageable qui est réglementé par la loi. La notion de bruit dans ce contexte de réglementation est défini par une mesure qui est calculée à partir du bruit en tant que grandeur physique. Le candidat terme « bruit » est polysémique. Par conséquent, deux concepts terminologiques `bruitGrandeurPhysique` et `bruitJuridique` sont définis à partir du candidat terme « bruit » afin de construire un modèle du domaine HSE. Une définition en langue naturelle est associée à chacun de ces deux concepts. Les termes les dénotant n'apparaissent pas dans le corpus étudié. En revanche, le terme « installation » est utilisé pour désigner le concept terminologique installation.

La figure 3 présente une représentation partielle du réseau termino-ontologique associé aux concepts terminologiques « bruitGrandeurPhysique », « bruitJuridique » et « installation ».

Le concept terminologique `bruitGrandeurPhysique` possède un unique sens déterminé par une définition donnée en langue naturelle, des propriétés intrinsèques et des relations le liant aux concepts terminologique `bruitJuridique` et installation.

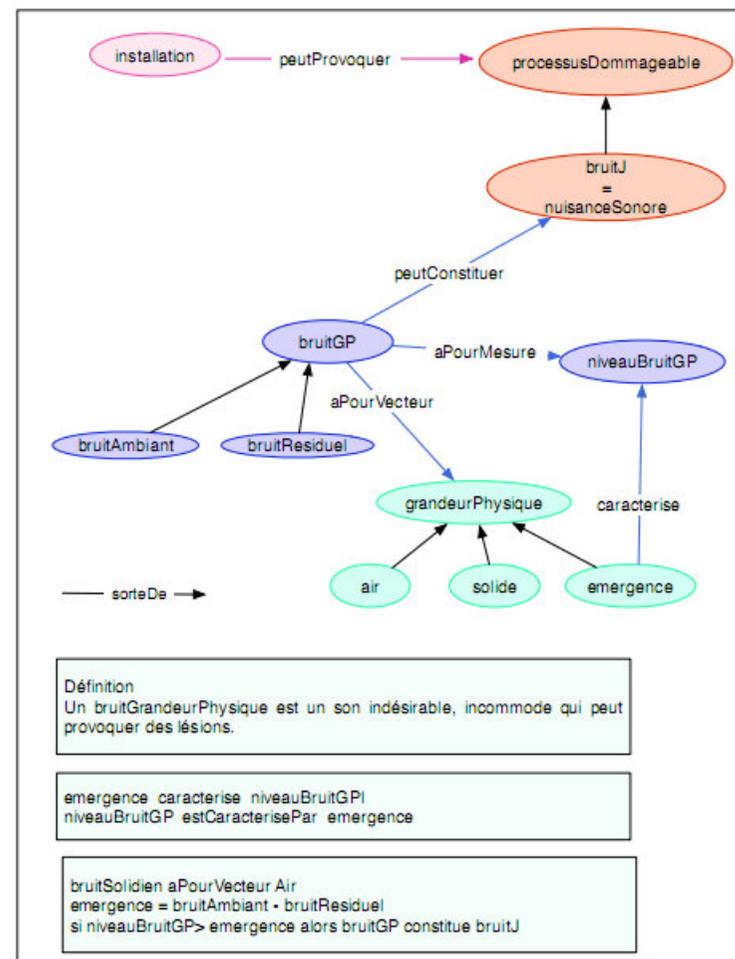


Fig. 5 : réseau termino-ontologique partiel

Définition

Un concept terminologique est défini de façon univoque par

- un terme vedette qui peut ne pas figurer dans le corpus ;
- des occurrences du terme qu'il désigne dans le corpus ;
- une définition en langue naturelle, fournie dans le corpus ou élaborée par le cogniticien ;
- un ensemble de synonymes ;
- des propriétés intrinsèques ;
- des relations hiérarchiques ou fonctionnelles les liant aux concepts terminologiques constituant le réseau.

Une relation terminologique

- possède une cardinalité ;
- peut être caractérisée par des propriétés (reflexivité, symétrie, transitivité, ...)
- peut avoir une relation inverse.

3.4. Les notions ontologiques

L'étape ontologique consiste à traduire le modèle conceptuel obtenu à la fin de l'étape terminologique dans un langage formel qui permet à la fois de s'affranchir des problèmes liés à la langue naturelle et au contexte, et d'effectuer des calculs et des raisonnements sur les entités représentées.

Définitions

Une ontologie informatique est constituée de concepts, appelés ontologique afin d'éviter toute confusion, organisés dans une hiérarchie avec héritage des propriétés, liés par des relations ontologiques existant entre eux et contraints par des règles et des axiomes. Elle permet en particulier d'implanter des mécanismes de raisonnement déductif, de classification automatique ou de recherche d'information.

Un concept ontologique est décrit dans un langage formel de représentation des connaissances. Il possède des propriétés.

Une relation ontologique (hiérarchique, descriptive) est une relation liant des concepts ontologiques, construite à partir de relations terminologiques et décrite dans un langage formel.

Une relation hiérarchique exprime un héritage des propriétés du concept. Une relation descriptive est une relation binaire qui exprime un lien entre deux concepts.

Les entités à représenter sont :

1. celles constituant la base de l'ontologie : (a) les classes représentant un ensemble d'individus partageant les mêmes propriétés. Elles sont assimilées aux concepts ontologiques. (b) les propriétés associant des paires attributs/valeurs à des individus ou restreignant des classes ; (c) les individus représentant une instance d'une classe.
2. les relations sémantiques qui comportent celles : (a) liant les entités constituant l'ontologie ; (b) décrivant les relations.

La représentation de ces différentes entités requiert de respecter les contraintes liées au formalisme adopté. Cette traduction n'est pas toujours possible et exige alors une adaptation du modèle conceptuel et des choix de représentation.

Différents formalismes de représentation des connaissances peuvent être utilisés pour décrire une ontologie. Parmi ceux existants, tous ne possèdent pas les mêmes niveaux d'expressivité et ne garantissent pas le même niveau de calculabilité.

Dans ce papier, nous adoptons la représentation dans le langage OWL DL (<http://www.w3.org/2004/OWL/>) qui repose sur les logiques de description [Baader et al., 2003].

Exemple

Un concept ontologique est décrit par une expression logique exprimant des relations du concept avec d'autres concepts de l'ontologie. La figure 5 présente une expression partielle de bruit en langage OWL utilisant la syntaxe XML.

```

bruitAmbiant est une sous classe de bruitGP
<SubClassOf>
  <OWLClass URI="&Ontology1222676008394;bruitAmbiant"/>
  <OWLClass URI="&Ontology1222676008394;bruitGP"/>
</SubClassOf>

bruitAmbiant et bruitResiduel sont des classes disjointes
<DisjointClasses>
  <OWLClass URI="&Ontology1222676008394;bruitAmbiant"/>
  <OWLClass URI="&Ontology1222676008394;bruitResiduel"/>
</DisjointClasses>

bruitGP aPourOrigine {air ou son }
<ObjectPropertyRange>
  <ObjectProperty
    URI="&Ontology1222676008394;aPourOrigine"/>
  <ObjectUnionOf>
    <OWLClass URI="&Ontology1222676008394;air"/>
    <OWLClass URI="&Ontology1222676008394;solide"/>
  </ObjectUnionOf>
</ObjectPropertyRange>

```

Fig. 6 : expression partielle du concept ontologique bruitGP

La figure 6 présente une expression partielle des différentes relations liant les concepts ontologiques bruit GP, bruitJ et la hiérarchie des relations.

```

relation entre bruitGP et bruitJ : peutProvoquerNuisanceSonore
relation entre bruitGP et air ou solide : aPourOrigine
<ObjectPropertyRange>
  <ObjectProperty
    URI="&Ontology1222676008394;peutProvoquerNuisanceSonore"/>
  <OWLClass URI="&Ontology1222676008394;bruitJ"/>
</ObjectPropertyRange>

  <ObjectPropertyRange>
  <ObjectProperty
    URI="&Ontology1222676008394;aPourOrigine"/>
  <ObjectUnionOf>
    <OWLClass URI="&Ontology1222676008394;air"/>
    <OWLClass URI="&Ontology1222676008394;solide"/>
  </ObjectUnionOf>
</ObjectPropertyRange>

hiérarchie de relations entre : peutProvoquer et
peutProvoquerNuisanceSonore
<Declaration>
  <ObjectProperty
    URI="&Ontology1222676008394;peutProvoquerNuisanceSonore"/>
</Declaration>

  <ObjectPropertyRange>
  <ObjectProperty
    URI="&Ontology1222676008394;peutProvoquer"/>
  <OWLClass URI="&Ontology1222676008394;nuisance"/>
</ObjectPropertyRange>

```

Fig. 7 : expression partielle des relations ontologiques

A partir de cette représentation, il devient possible de raisonner. Ainsi, il est possible de déduire l'émergence à partir des concepts mesureDeBruitAmbiant et mesureDeBruitResiduel et de déterminer le seuil pour lequel l'émergence est considérée comme correspondant à une nuisanceSonore qui est un concept équivalent de bruitJ.

4. Conclusion

La finalité de ce papier était de montrer comment le passage du terme au concept se construit.

Notre position est que la transition s'effectue graduellement en passant par trois plans bien identifiés : linguistique, termino-ontologique et ontologique. En cela, nous proposons une étape supplémentaire à l'approche ontoterminologique proposée par Roche [Roche, 2007]. En effet, la réalisation du modèle conceptuel est dissociée de la phase ontologique.

Le concept terminologique joue le rôle de passeur entre les plans linguistique et ontologique. Il peut-être envisagé selon deux facettes. Il a d'une part, une expression en langue, mais restreinte par une expression dans un langage semi-formel [Kassel, 2002] d'autre part, il est défini par des propriétés structurelles et fonctionnelles. Il en est de même pour les relations termino-ontologiques qui constituent une traduction des relations lexicales repérées dans le texte.

Les ressources associées à chacun de ces plans sont caractérisées par des propriétés qui les distinguent mais qui les rendent complémentaires. Ainsi, l'ontologie permet d'effectuer des tâches de classification et de raisonnement qui sont logiquement fondées tandis que le réseau terminologique permet de représenter les notions exprimées dans les textes selon le contexte de l'application visée. Le modèle conceptuel dit réseau termino-ontologique permet de changer de point de vue et d'assurer la transition entre le modèle linguistique et ontologique.

A l'issue de ce papier, il devient possible de répondre positivement à la question « Est-il possible de construire une ontologie à partir de textes ? ».

5. Remerciements

Que soient ici remerciés Jérôme Nobecourt membre du LIM&BIO et Haïfa Zargayouna membre de l'équipe RCLN pour les discussions fructueuses occasionnées par l'écriture de ce papier.

Bibliographie

Aussenac-Gilles N., Després S., Szulman S. *The TERMINAE Method and Platform for Ontology Engineering from Texts. Bridging the Gap between Text and Knowledge – Selected Contributions to Ontology Learning and Population from Text*. Eds Buitelaar P., Cimiano, P. pp. 199-223. IOS Press, 2008.

Baader F., Calvanese D., McGuinness D. L., Nardi D., Patel-Schneider P. F. *The Description Logic Handbook: Theory, Implementation, Applications*. Cambridge University Press, Cambridge, UK, 2003.

Biebow B., Szulman S. *TERMINAE : a method and a tool to build a domain ontology*, In *Proceedings of International Workshop on Ontological Engineering on the Global Information Infrastructure*, eds Benjamins V. R. and Fensel D. and Gomez Perez A., pp. 25-30, 1999.

Bourigault D., Charlet J. *Construction d'un index thématique de l'ingénierie des connaissances, Ingénierie des connaissances. De Régine Teulier, Jean Charlet* Publié par L'Harmattan, pp.29-47, 2005.

Bourigault D. , Aussenac-Gilles N. *Construction d'ontologies à partir de textes*, in TALN, *Batz sur Mer*, 2003.

Bourigault D., Jacquemin C. *Construction de ressources terminologiques*, in J.-M. Pierrel (éd.), *Industrie des langues*, Hermès, Paris, pp. 215-233, 2000.

Cimiano P., Volker J., Studer, R. *Ontologies on Demand? - A Description of the State-of-the-Art, Applications, Challenges and Trends for Ontology Learning from Text*. In *Information Wissenschaft and Praxis*, 57, pp.315-320, 2006.

Cimiano P. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer. November 2006.

Cruse, D. A. 1986. *Lexical semantics*. Cambridge, England: University Press.

Després S., Szulman S. –*Construction d'une ontologie formelle à partir d'un texte de droit*, pp. 261-281, numéro 11(2) de *International journal of theoretical and applied issues in specialized communication*. John Benjamins Publishing Company, 2005.

Després S., Furst F., Szulman S. *Construction d'une ontologie du domaine HSE*. In *actes de la conférence Ingénierie des Connaissances IC'2007*, pp. 133-144, 2007.

Kassel G. *Une méthode de spécification semi-informelle d'ontologies*. In *Actes des 13 èmes journées francophones d'Ingénierie des Connaissances IC 2002* pp. 75-87, 2002.

Mondary T., Després S., Nazarenko A., Szulman, S. Construction d'ontologies à partir de textes : la phase de conceptualisation. In actes de la conférence Ingénierie des Connaissances IC '2008, pp. 87-98, 2008.

Roche C. Le terme et le concept : fondements d'une ontoterminologie. In acte Conférence Terminologie & Ontologie : Théories et Applications Toth 2007.

Pragmaterminologie : les verbes et les actions dans les métiers

Dardo de Vecchi

Université Paris Diderot - UFR EILA
Euromed-Marseille, Ecole de management
BP 921 – 13288 Marseille
dardo.devecchi@euromed-marseille.com

Laurent Estachy

Euromed-Marseille, Ecole de management
BP 921 – 13288 Marseille
laurent.estachy@euromed-marseille.com
http://www.euromed-marseille.com

Résumé :

La description d'un domaine d'activité ou d'un métier, fait appel à la terminologie qui montre et structure une organisation notionnelle. En même temps, l'explicitation des activités d'un métier nécessite la mise en relief des formes verbales qui consolident les formes nominales qui, en terminologie, sont majoritaires. A travers un exemple d'actualité, l'article met en évidence le besoin de montrer la dépendance entre verbes et nominaux dans la description des métiers en ponctuant le possible rapprochement entre approches pragmaterminologique et ontoterminologique.

Mots-clés : métiers, compétences, verbes, actions, terminologie, pragmaterminologie, ontoterminologie.

Im Anfang war die Tat!
Au commencement était l'Action !
 Goethe, Faust, 1237

1. Introduction

Les manuels de terminologie et les travaux terminologiques montrent qu'une place prépondérante est accordée aux nominaux, les autres formes grammaticales occupant une place nettement moins importante. De leur côté, les études sur les collocations, sur les phraséologies ou sur les cooccurrences mettent en avant les rapports existant entre différentes formes grammaticales ; ces études relèvent d'une analyse de la langue spécialisée et ne retiendront pas notre attention dans le cadre de cette recherche en cours. Contrairement à ces études, nous nous intéresserons ici aux possibilités d'exploitation des verbes dans un domaine d'activité via la pertinence de leur présence dans les discours des spécialistes.

Après constater la place accordée à certains syntagmes nominaux dans les articles de presse à propos de l'affaire Kerviel, nous allons explorer la place des verbes dans quelques dictionnaires des métiers et dans quelques manuels de finance. Cela nous montrera l'utilité de la mise en avant des verbes dans la description des métiers. En nous appuyant sur cette constatation, c'est le résultat du dialogue entre l'expert financier (Estachy) et le linguiste (de Vecchi) qui montrera comment les verbes d'un domaine d'activité des opérations bancaires - où les métiers de la finance s'exercent- servent à repérer la place des acteurs dans ce domaine et à distinguer leurs métiers. Le rôle des verbes s'inscrit pleinement dans une démarche pragmaterminologique qui met en avant la place de l'action dans la structuration conceptuelle d'un domaine de connaissance. Grâce à cette méthode, nous espérons pouvoir apporter une réflexion supplémentaire au rapprochement entre terminologie et ontologies (ontoterminologie) telle que proposée par C. Roche (Roche, 2007a & 2007b). La méthode esquissée ici pourrait aussi avoir une influence dans l'enseignement des activités d'un secteur professionnel ou dans un tout autre registre dans les recherches sur la langue spécialisée.

2. Verbes : connaissances et métiers

Si, en terminologie, les nominaux sont les plus nombreux, ils ne suffisent pas à eux seuls pour décrire les actions *réalisées* dans un métier ou par les acteurs d'un

domaine d'activité : des verbes qui gravitent autour des nominaux contribuent sans doute à la description de domaine. Les connaissances et l'exercice des métiers ne peuvent être expliqués que si le lien entre nominaux se fait à l'aide des verbes et des syntagmes verbaux. Autrement dit, la description d'un *métier* ne peut ni se limiter exclusivement aux nominaux ni ignorer la dynamique des verbes du domaine d'activité où elle s'exerce. Dire un métier, c'est dire ses *actions* et non seulement ses « objets ».

2.1. Les verbes dans un fait d'actualité

L'actualité apporte souvent sur le devant de la scène le besoin d'aborder un tant soit peu la terminologie d'un domaine de connaissance afin que les informations puissent être comprises. Le 24 janvier 2008, la Société Générale a reconnu avoir perdu 4,9 milliards d'euros suite aux activités qualifiées de frauduleuses, de l'un de ses opérateurs de marchés Jérôme Kerviel. Cet événement est rapidement devenu pour les médias l'« Affaire Kerviel » : il s'agit en effet de la perte la plus importante jamais enregistrée par un établissement financier dans des activités de trading¹ sur les marchés de capitaux. Plus précisément, celle-ci résulte du débouclage de positions ouvertes non autorisées prises par Jérôme Kerviel, positions qui s'élevaient à plus de 50 milliards d'euros courant janvier 2008. Par position ouverte, on entend l'achat ou la vente simple d'actifs financiers, sans couverture annexe neutralisant les risques liés aux fluctuations de cours. Dans le cas présent, il s'agissait d'opérations portant sur des contrats de futures sur indices boursiers : Jérôme Kerviel envisageait à cette date un rebond à la hausse des principaux indices européens. A ce sujet, il convient donc de noter que par définition, des positions dites ouvertes sont de nature directionnelle et spéculative : elles ont en effet pour objectif de dégager un profit, suite à une évolution de prix favorable dans le temps, à la hausse ou à la baisse des cours.

Si le monde de l'économie et des finances était en mesure de comprendre les articles de presse, l'explication des événements se devait de présenter la terminologie opaque du métier de trader à ceux qui n'en étaient pas familiers. Ainsi, pour aider à la compréhension des articles de presse, plusieurs journaux

¹ Toutes les expressions d'origine anglaise sont considérées comme appartenant à la langue spécialisée de la finance et ne sont pas traitées comme des emprunts, en fonction de quoi elles n'apparaissent pas en italiques (que nous réservons pour faire ressortir les expressions utiles à notre raisonnement).

ont publié des mini-glossaires qui expliquaient sommairement la signification de quelques expressions propres au métier de trader. Dans ces glossaires, les nominaux ont été majoritaires, mais la fréquence inattendue et récurrente des verbes par rapport aux nominaux a attiré notre attention² ; la présence des syntagmes verbaux « prendre des positions » ou « arbitrer des marchés » revenaient fréquemment. En effet, parmi les actions qu'un trader effectue, se trouve celle de *prendre* des positions ou d'*arbitrer* différents marchés. Cette fréquence inattendue des verbes dans l'explication d'un métier mérite qu'on s'y attarde.

2.2. Les verbes dans les dictionnaires des métiers

L'observation de quelques dictionnaires des métiers et des activités professionnelles est aussi très utile pour ce qu'est de l'inclusion des verbes qui désignent des actions à effectuer.

Si l'édition 2008 du Dictionnaire comptable et financier édité par le Groupe Revue Fiduciaire ne fait mention d'aucun verbe parmi ses 810 entrées, il mentionne fréquemment des métiers –et quelques organismes (situation peu fréquente dans les textes consultés)- pour lesquels les verbes conjugués indiquent les actions effectuées. Par exemple, et dans une liste non exhaustive : le commissaire aux apports évalue, apprécie la valeur des apports et les avantages particuliers qui peuvent être stipulés lors des [...] opérations et établit un rapport. Le commissaire aux comptes certifie les comptes annuels, opère des vérifications sur des comptes. Le commissaire aux fusions vérifie les modalités de la fusion, signale des difficultés d'évaluation. L'expert comptable révisé, apprécie, tient, centralise, ouvre, arrête, surveille, redresse, organise et analyse les comptabilités des entreprises.

Parmi les 25 auteurs ayant participé à la rédaction du Petit dictionnaire de la faillite, un seul fait appel à un verbe : prononcer (pour prononcer un jugement). Ailleurs, Loïc Depecker fait mention de 132 verbes et locutions incluant un verbe parmi les 793 entrées de son Dictionnaire du français des métiers. Avec une fréquence notable de composés avec les verbes avoir, être, faire, mettre, prendre et travailler, véritables « hyperonymes » verbaux et fonctionnant comme des

² Par exemple, 1 verbe sur 7 expressions dans *20 minutes*, le 28 janvier 2008, <http://www.20minutes.fr/article/209241/France-Trader-futures-arbitrage-Qu-est-ce-que-c-est.php>

verbes support. Parmi les 1707 entrées de *Le jargon des postiers*, seuls 25 verbes apparaissent de manière explicite, 4 permettent d'être déduits de manière immédiate à partir de nominaux (affranchir, boucler, oblitérer, recouvrer) et 12 sont des syntagmes verbaux. Dans *Le petit décodeur [de l'administration]*, sur les plus de 3000 mots, expressions et sigles expliqués, les auteurs éprouvent le besoin de citer non pas moins de 412 verbes et syntagmes verbaux. Le chiffre du dernier exemple cité peut paraître important, mais il faut garder à l'esprit que ces quatre dictionnaires concernent une multitude d'activités et de métiers. Finalement, dans *Le parler des métiers*, de Pierre Perret (avec la collaboration de Gabrielle Quemada) le chiffre grandit considérablement. Le dictionnaire englobe une très grande quantité de domaines d'activité. Gabrielle Quemada écrit avec justesse : « Priorité a été donnée à la communication entre praticiens, seuls dépositaires des connaissances théoriques ou empiriques requises pour donner sens aux énoncés observés ». (Perret, 2002 : 19).

2.3. Les verbes et les métiers dans les manuels de finance

Dans les manuels de référence en finance, on constate fréquemment l'absence de formes verbales aux index. Chez Cobbaut, aucun verbe n'est cité parmi les 307 entrées de son index. Plus de 700 entrées dans l'index de l'ouvrage de Levasseur et Quintart, mais aucun verbe. La situation est analogue pour Bernard et Colli. Eiteman, Stonehill et Moffet ne présentent aucun verbe dans leur glossaire de 296 entrées leur index (trop hétérogène, noms propres, institutions, termes, etc.) ne font apparaître aucun verbe. Ailleurs, Brealey et Myers incluent 6 syntagmes verbaux à leur index : *adosser* à des actifs, *adosser* à des échéances, *désendetter* le bêta, *ralentir* le paiement des factures, *retarder* les paiements des factures, et *re-endetter* le bêta. On trouve par ailleurs pour *désendetter*, *déleverager* (suivi du commentaire : '*déleverager* « en sabir »', (Brealey et Myers, 2003 : 593)). Teulié et Topsacalian pour leur part s'ils ne mentionnent qu'un seul verbe, *réinvestir*, ils comptent parmi les rares auteurs qui indiquent des acteurs : le cambiste et le courtier.

Bodie et Merton, dans l'édition française de leur manuel *Finance* dirigée par Christophe Thibierge présentent séparément un glossaire, deux lexiques (français/anglais et anglais/français), une liste de synonymes et un index. Le glossaire contient 270 entrées (3 verbes : *se couvrir*, *être court*, *être long*), les lexiques 255 entrées pour la partie FR/EN (les mêmes verbes) et 228 entrées pour la partie EN/FR (les mêmes verbes), la liste de synonymes 110 entrées (1 verbe : *émittre* dans « émettre des actions, des obligations ») et finalement l'index de 385

entrées (5 formes verbales : *assurer*, *être illiquide*, *être liquide*, *indexer* et *répliquer* l'indice –les deux derniers étant synonymes). Ce dernier texte, remarquable par la conscience linguistique et terminologique des auteurs, insiste sur la place des termes dans ce domaine de connaissance ; ce serait un exemple – rare - de la nécessité de montrer les actions dans un domaine de connaissance comme celui des finances.

De manière générale, ce n'est pas seulement l'absence de verbes qui étonne mais également l'absence de la mention des métiers et ce d'autant plus que la liste de métiers extraite des sites internet fait apparaître plus de trente métiers, dont celui de trader³ par lequel nous avons commencé nos observations, et dont les activités s'entrecroisent.

L'exploration d'un métier, d'une profession ou plus largement d'une activité spécifique nécessite l'explicitation de formes de la langue spécialisée dans ce secteur. Si la langue se spécialise pour dire une connaissance (Lerat, 1995), elle se spécialise aussi pour dire la réalité immédiate de l'activité qui peut être plus spécifique que celle d'un domaine de connaissance. On spécialise la langue pour dire non seulement ce qu'on *conçoit*, mais aussi ce qu'on *fait*. Il en résulte que les actions ont un rapport fondamental aux termes utilisés dans une activité et, dans le cas des métiers et de manière incontournable, les verbes articulent pratiques et notions.

L'étude du verbe dans la réflexion terminologique reste éparse (L'Homme, 1995). Dans la description des métiers là où l'action se réalise, la situation en est une autre. Dans l'énonciation, la description et l'explicitation de l'action, il apparaît clairement que la présence des verbes spécifiques est nécessaire afin de répondre aux questions suivantes : en quoi *consiste* le métier de « x » ? ou que *fait* un « x » ? L'absence des formes verbales pertinentes dans la réponse à ces questions laisserait la description incomplète.

³ Pour les opérateurs de marché, comme le trader, l'observation de la dynamique des fluctuations de prix ainsi que l'analyse des sources de ces fluctuations constituent une part essentielle de l'activité et une référence centrale au cœur de processus de prises de décisions extrêmement complexes. Le métier des opérateurs relève d'une compréhension d'un ensemble hétérogène de conventions ainsi que d'une adaptation permanente à leurs éventuelles mutations. De fait, l'absence de la mention des opérateurs de marché dans les manuels surprend.

3. Méthodologie d'obtention des verbes

Il est facile d'imaginer qu'aujourd'hui pour comprendre l'affaire Kerviel au-delà des descriptions succinctes de la presse généraliste, un néophyte chercherait sur Internet un glossaire d'accès facile plutôt qu'un dictionnaire spécialisé d'accès plus difficile. En conséquence, nous avons cherché un glossaire de travail en fonction des critères suivants :

- être de taille limitée, mais suffisante pour pouvoir être cohérent et apporter les informations nécessaires pour comprendre le domaine d'activité ou les informations fournies par la presse ;
- rester proche du métier de trader et ne pas être trop généraliste en s'écartant du sujet ;
- se trouver parmi les premiers résultats affichés par Google ;
- être accessible gratuitement (certains sites sont payants ou figurent dans des anciens numéros payants d'un journal).

En fonction de ces critères, notre choix s'est porté sur le glossaire suivant (présenté d'ailleurs comme un lexique) : <http://www.fairinvest.com/21.htm>. Il contient en l'état⁴ 36 termes et vise essentiellement le vocabulaire des traders. Il est présenté par son auteur, Michel Cataneo, comme suit : « *Ce petit lexique de base vous aidera à vous familiariser (si besoin est) avec le vocabulaire utilisé par les traders* » (*ibid.*).

Chacun de ces 36 termes a été redéfini par l'expert dans un double but. Premièrement, afin de vérifier l'usage des formes en présence, notamment à l'oral où beaucoup d'expressions échappent aux textes écrits pour des raisons diverses (style, argots, « ça ne s'écrit pas », considération de manque d'importance, etc.), mais qui font partie de la réalité des pratiques. Deuxièmement, dans le but de capturer les verbes utilisés dans ces « nouvelles » définitions et surtout dans les explications fournies par l'expert à propos de cette terminologie. Il ne faut pas oublier que les spécialistes ne s'interrogent guère sur leur propre mode d'expression, ils ne sont pas linguistes ; pour eux, linguistiquement, tout va de soi et les problèmes n'apparaissent qu'en cas d'incompréhension manifeste ou de manque de consensus. En dépit de l'aspect fastidieux de l'exercice par rapport à une extraction automatique des verbes, force est de constater que l'observation de l'oral s'avère efficace puisque de cette manière sont apparus des acteurs, tel

⁴ Site consulté en mars 2008.

que le *sniper* dont la mention n'est pas systématique et qui pourtant existe et qui agit dans ce secteur d'activité, même si son rôle n'est pas primordial sans quoi il serait mentionné comme c'est le cas, entre autres, pour le *daytrader* dans le glossaire de *fairyinvest.com*. C'est donc 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. L'exemple suivant montre de manière synthétique la méthode de travail conjointe du financier et du linguiste :

Terme de départ extrait du site Internet de référence : « bull ».

(*fairyinvestment.com* fait apparaître : « Bull : expression imagée désignant un taureau pour qualifier une tendance haussière des marchés »).

Redéfinition du terme : qu'est-ce que bull ? C'est un type de marché : « un *marché bull* » (par opposition à un *marché bear*).

Recherche des verbes qui accompagnent ces termes : que fait-on dans un marché bull ? On *achète* dans un marché bull.

Recherche des agents grammaticaux susceptibles d'effectuer ces actions : qui achète dans un marché bull ? Un *investisseur* achète dans un marché bull.

Recherche des compléments : qu'est-ce qu'un investisseur achète dans un marché bull ? Un investisseur y achète des *actions*.

4. Résultats

Grâce à cette méthode de travail, nous avons obtenu une première liste de 15 acteurs de la finance⁵ que par recoupement par expert a été ajustée à 4 (acteurs retenus) : le *broker*, le *sales*, le *gestionnaire des fonds* et le *trader*. Nous avons recueilli aussi d'une seconde liste (non exhaustive) de 44 verbes seuls que nous avons ensuite accompagnés de leurs compléments. Par exemple : *négoier*, puis *négoier des actifs*, *négoier des fonds*, *négoier des options*.

Les listes obtenues ont été reliées graphiquement afin de visualiser les liens qui unissent acteurs, activités (par le biais des verbes) et objets sur lesquels

portent ces activités (par le biais des mêmes verbes + leur complément) comme schématisé dans le tableau suivant où ne figure qu'une sélection.

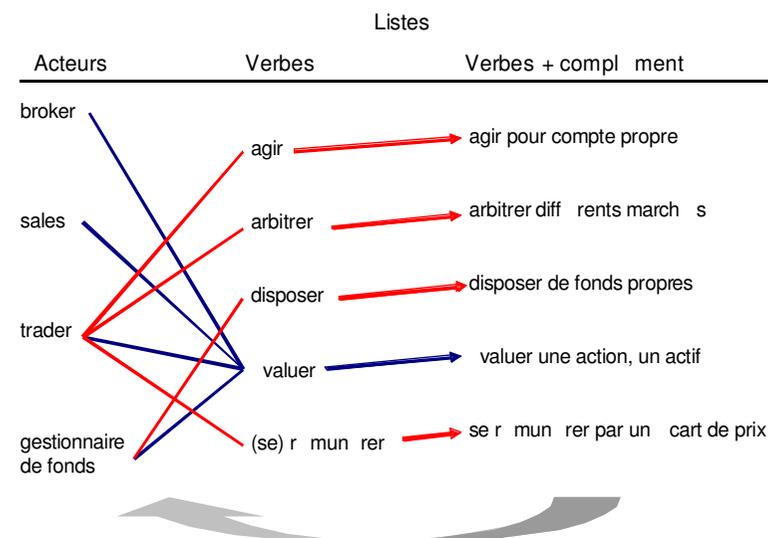


Fig. 1 Liens entre acteurs, verbes et verbes + complément

Dans un premier temps, cette représentation des liens entre listes nous ont permis de constater les faits suivants :

- le nominal seul (*bull*) ne suffit pas pour comprendre l'activité, une mise en discours est nécessaire (on *achète* dans un marché bull) ;
- le verbe seul (*acheter*) ne suffit pas pour décrire l'activité, le syntagme verbal est nécessaire (on achète dans un *marché bull*) ;
- certaines activités peuvent être effectuées par un même agent, en bleu dans le schéma (*un actif est évalué par un broker, par un sales, par un trader ou par un gestionnaire de fonds*) ;
- ces spécificités de l'action renvoient aux acteurs (flèche grise) ;

⁵ Broker, contrarians, contrôleur de risque, courtier, daytrader, investisseur, opérateur, sales, sales en Asie, scalpeur, sniper, trader, trader junior, trader senior, trend followers. Nous

- certaines activités, en nombre restreint, ne sont effectuées que par un seul acteur (*le trader agit pour compte propre, arbitre différents marchés et se rémunère par un écart de prix*).

Dans un second temps, l'analyse des acteurs des marchés financiers et de leurs interactions nous permis d'établir le schéma suivant qui est commenté ensuite :

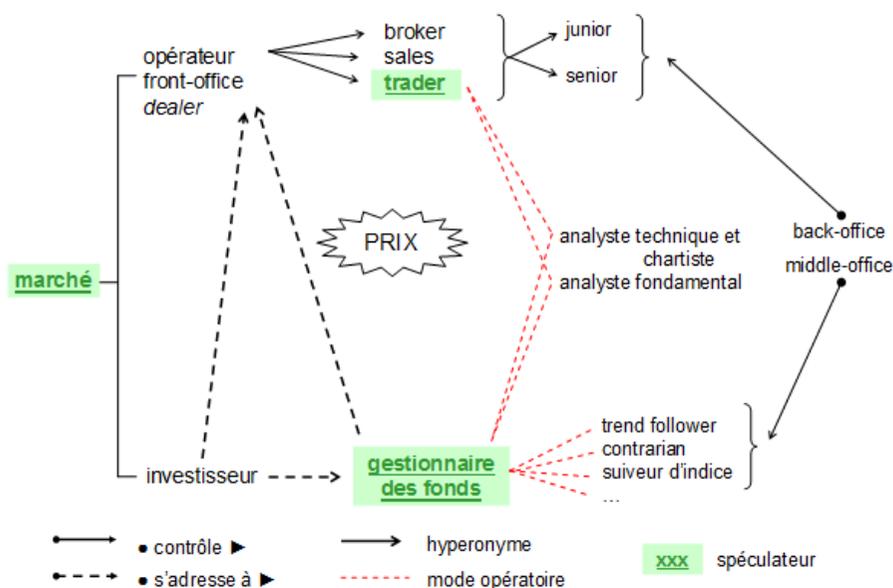


Fig. 2 Interactions entre intervenants sur les marchés financiers

- 1- L'ensemble des *investisseurs*⁶ que nous qualifierons de finaux sur les marchés de capitaux (particuliers, entreprises, Etat) confie les sommes à leur

⁶ Les mêmes intervenants (particuliers, entreprises, Etat) peuvent également intervenir en tant que demandeur de fonds et *émettent* actions, obligations, etc. Dans ce cadre, on peut alors considérer les banques d'investissement comme des acteurs *agissant* en tant qu'intermédiaires entre investisseurs et émetteurs. On parle alors de finance directe. Le

disposition à des *gestionnaires de fonds*: fonds mutuels-Sicav, fonds de pension, hedge-funds, fonds souverains, etc. Ainsi quelques milliers de gestionnaires de fonds dans le monde opérant sur les principales financières (New York, Londres, Tokyo, etc.) gèrent l'épargne mondiale ainsi mise à leur disposition.

- 2- Ces gestionnaires de fonds s'adressent ensuite à des banques d'investissement et plus précisément à leurs opérateurs de salle de marché (en anglais : *front office*). L'essentiel des opérations traitées sur les marchés de capitaux transite donc par ces salles : qu'il s'agisse d'achat ou de vente d'actions ou d'obligations, d'opérations de prêts/emprunts sur les marchés monétaires, d'achat ou de vente de devises sur le marché des changes, ainsi enfin que toute opération sur les produits dérivés associés à ces actifs financiers (options, swaps, futures).
- 3- La confrontation des offres et des demandes sur l'ensemble de ces marchés fait l'objet d'une gestion effectuée par les différents traders opérant pour le compte de leurs banques. Cette confrontation aboutit à l'établissement de prix fluctuant au gré des flux traités entre traders pour le compte d'investisseurs finaux mais également au gré des positions pour compte propre prises par ceux-ci, reflets de leurs propres anticipations (position ouverte, spéculation). Les analyses dites techniques ou chartistes ainsi que les analyses dites « fondamentales », celles se référant à des données macroéconomiques, constituent la base essentielle des anticipations des opérateurs de marchés⁷.
- 4- Enfin, agents de back office, de middle office, contrôleur de risques/auditeurs et comptables assurent la bonne fin des opérations et les différents suivis traités ainsi que leurs compatibilités avec les normes de risques retenus par chacun des établissements intervenants sur les marchés de capitaux.

schéma présenté ci-dessus a volontairement été simplifié pour ne pas nuire à la compréhension de la chaîne de relations interactives développées ci-dessus.

⁷ Investisseurs (gestionnaires de fonds) mais également émetteurs *anticipent* en permanence l'évolution des prix sur les marchés soit pour *prendre* des positions nouvelles, soit pour *couvrir* des risques, leurs outils d'analyse sont identiques à ceux utilisés par les traders des banques d'investissements : analyses chartistes et fondamentales servent donc de base aux décisions d'investissements de l'ensemble des intervenants.

A ce stade, il nous est apparu intéressant de constater que la notion de « prix », en tant que quasi acteur, est très fréquemment reprise dans le vocabulaire des opérateurs de marché. Au point que cette notion semble pouvoir être assimilée à un opérateur central dominant susceptible d'influencer l'ensemble des intervenants. Cette conjecture nous est apparue comme pouvant constituer un point d'entrée original dans les débats opposant individualisme et holisme méthodologique au sein de la communauté des économistes. Une réflexion est donc en cours sur ce point.

Une terminologie purement centrée sur les nominaux de la langue spécialisée dans la finance ne mettrait pas en évidence cette dynamique ; il est nécessaire d'inclure dans la terminologie de la finance d'autres paramètres qui tiennent compte des acteurs et des activités, ce qui peut être fait par le biais des verbes que, considérés comme pivots du discours, permettent une autre lecture des opérations de marché et de ceux qui les effectuent. Ainsi, il n'est pas possible de comprendre *ce qu'un trader fait* qu'en signalant les activités propres au métier et qui le distinguent des autres métiers de la finance ; le lexique de *fairyinvest.com* oriente dans l'explication mais ne lui donne pas corps, sa didactique est restreinte tout comme les glossaires dans Bodie et Merton (§ 2.3). Ces glossaires existent pour un répertoire de manière synthétique les éléments utiles à une connaissance en donnant leur définition, mais non pour expliquer l'activité où ils sont en scène, ce que l'expert peut faire. En résumé, pour comprendre ce qui s'est passé dans le cas de cette affaire, la liste des nominaux propres à la finance ne suffit pas : il faut aussi expliquer ce qu'on *fait* dans ces métiers. Ce besoin expliquerait la quantité hors du commun des verbes dans les glossaires publiés par la presse (toute proportion gardée) ou dans les dictionnaires de métier et en moindre mesure des manuels de finance comme on a pu le voir. On peut alors conclure que la connaissance du *métier* nécessite impérativement la mention des « verbes du métier ».

Bien que des recherches plus poussées soient nécessaires, il nous semble qu'outre l'explication des métiers des opérations financières, le domaine d'activité pourrait être expliqué en prenant comme point de départ les verbes qui signalent les activités qui y sont effectuées pour être ensuite accompagnées des compléments. Par exemple, *acheter, agir, arbitrer, couper, se couvrir, placer, porter, se rémunérer, traiter, transmettre, etc.* ouvrent la voie à la prise en compte aux nominaux du domaine : *acheter* → *des actions, des fonds, des options...*, *agir* → *pour compte propre, arbitrer* → *différents marchés, couper* → *une perte, une position, se couvrir* →

contre un risque, placer → *un ordre, porter* → *une position, se rémunérer* → *par un écart de prix, traiter* → *un ordre, transmettre* → *une confirmation, etc.*

On voit pourquoi le fait qu'il s'agisse ou non de collocations, est relativement peu important et ce sont les cooccurrences de ces verbes avec de nominaux qu'il faut rechercher. Quant à la phraséologie, l'expert tient le dernier mot car c'est lui qui détient la vérité de l'usage de la parole spécialisée.

5. Conclusion

Les terminologies et les ontologies d'un domaine de connaissance – et plus particulièrement d'un domaine d'activité ou d'exploitation⁸- gardent un lien très étroit avec les métiers et avec leurs pratiques. Ces derniers fournissent le corpus que les premières traitent et ce indépendamment des besoins de description, de transmission ou de normalisation (Rey, 1993 : 55) qui motivent les terminologies et les ontologies. Dans ce va-et-vient entre connaissance et pratique, il est nécessaire de prendre en considération un « savoir-effectuer » dans lequel les verbes occupent un rôle essentiel et solidaire des nominaux comme nous l'avons montré au long de cet article. Il faut alors inclure l'activité et sa position dans le temps, comme critères terminologiques – là où il est pertinent de l'exploiter car cela n'est pas toujours nécessaire- dans le domaine à étudier. La dynamique entre connaissance et pratique, dans l'exercice d'un métier, peut ainsi être prise en compte par l'approche pragmaterminologique des notions dans un domaine (de Vecchi, 2007). Cette approche considère dans les termes⁹ quatre aspects indissociables : linguistico-cognitif (terminologie « classique »), social (communautés de pratiques et d'exercice des professions, donc socioterminologique), temporel (validité dans le temps et évolution des termes pour une communauté) et d'action (activités concrètes et intentions de communication des termes - notamment lors de leur formation). Dans une terminologie ou dans une ontologie d'un domaine d'activité ou d'un métier, un terme – signe aura sa place si ces dernières sont en mesure de considérer ces

⁸ Pour le développement de la fragmentation de la notion de domaine de connaissance voir de Vecchi 2005, in bibliographie.

⁹ Nous considérons le terme comme l'aboutissant sémiotique d'un processus de conceptualisation. *Aboutissant*, car c'est le résultat saisissable de l'activité mentale qui le forme. *Sémiotique*, car ce résultat est un signe (qu'il soit en langue ou non est une autre affaire). *Processus de conceptualisation*, car une conception se fait dans le temps et non de manière instantanée.

quatre aspects pour être consensuelles, cohérentes, précises et évolutives pour être acceptées et exploitées (Roche, 2005 : 57).

Aujourd'hui, un métier évolue en s'adaptant aux contraintes de la société qui l'entoure et il est indéniable que les termes reflètent cette adaptation. Les termes – signes (nous insistons) n'ont pas la même valeur, en même temps *et* avec les mêmes implications pour tous les membres d'une société, surtout s'ils n'effectuent pas les mêmes activités. Pour cette raison, nous avons non seulement des dictionnaires, surtout spécialisés, mais aussi des ontologies notamment dans la perspective de Roche : « Définie pour un objectif et un domaine particulier, une ontologie est pour l'ingénierie des connaissances une représentation d'une modélisation d'un domaine partagé pour une communauté d'acteurs » (*ibid.*).

Si la terminologie garde un rapport essentiel à la langue spécialisée et les ontologies à l'ingénierie des connaissances, elles partagent néanmoins avec les communautés d'acteurs les concepts que les acteurs emploient. Autrement dit, ces deux disciplines ne s'excluent pas l'une à l'autre, elles se complètent dans deux volets distincts du traitement de ce que les acteurs activent dans l'univers de la pensée : le concept. Ces volets apparaissent dans la figure 3, où le traitement en trois dimensions permet une visualisation. Nous l'avons vu dans cet article, les acteurs ne sauraient être évacués du circuit : ils détiennent le *savoir* et *valident* les termes en usage tant pour la terminologie que pour des ontologies.

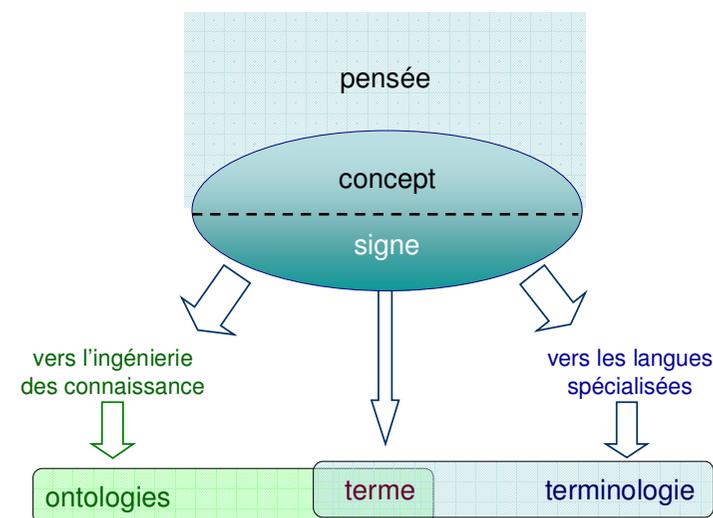


Fig. 3 Le concept à la charnière de la pensée, de la terminologie et des ontologies

Bien que cette recherche sur le rôle des formes verbales reste modeste et soit en cours, il nous semble évident que la prise en compte des verbes et des actions dans la description d'un domaine d'activité, et donc des métiers, est impérative. Les verbes n'en sont pas la panacée, mais ils ne sauraient être omis, et moins encore les syntagmes verbaux¹⁰. Une description du métier basée essentiellement sur des nominaux aboutit à une description statique où les interactions sont absentes. Or, une prise en considération des verbes (qu'ils soient d'action, d'état ou de relation) rend cette description dynamique. L'approche

¹⁰ Il faut également garder à l'esprit que les textes orientés vers l'action (modes d'emploi, recettes de cuisine, notices explicatives, etc.) s'articulent souvent autour des verbes et non seulement des nominaux. Ce n'est pas gratuit si les recettes de cuisine énoncent l'action puis les objets utilisés et non le contraire : *préchauffez, beurrez, coupez, battez, étalez, enfournez, servez, accompagnez.*

pragmaterminologique pourrait apporter à l'ontoterminologie proposée par Roche (Roche, 2007), la dynamique notionnelle des verbes que nous avons montré ici de manière à compléter la liste de caractères formels des termes.

Outre les aspects théoriques que nous venons d'aborder, les applications de nos résultats peuvent être diverses. La considération des verbes permettrait pour un domaine d'activité - et en parallèle à celle des collocations, cooccurrences et phraséologies- un autre accès à la langue spécialisée ; les opérations de finance en sont un exemple (§ 4). Bomati a appliqué notre méthode à propos des verbes dans l'administration en France (Bomati, 2008). En traduction, la méthode pourrait contribuer à l'étude des verbes (cf. L'Homme, 1995). En pédagogie et pour un domaine d'activité, voire d'exploitation s'il s'agit d'intégration dans une entreprise (de Vecchi, 2008), la méthode servirait à signaler des activités et leurs acteurs. Mais pour l'instant, il s'agit des pistes à explorer.

Bibliographie

- Bodie, Z. et Merton, R. (2001), *Finance, Paris, Pearson Education Village Mondial*
- Bomati, Y. et al. (2008), *L'administration en bons termes – 1000 mots clés pour utiliser et comprendre le langage administratif, Paris, Vuibert*
- Brealey, R. et Myers, S. (2003), *Principes de gestion financière, Paris, Pearson Education, 7^e éd.*
- Cobbaut, R. (1997), *Théorie financière, Paris, Economica, 4^e éd.*
- Eiteman, D., Stonehill, A. et Moffet, M. (2004), *Gestion et finance internationales, Paris, Pearson Education France*
- Girardin, C. (1995), « Trader, aux frontières du néologisme », in *Terminologies nouvelles*, n° 14, déc. 1995
- Giraud, Pierre-Noël (2001), *Le commerce des promesses, Paris, Seuil*
- Keynes, J. (1936), *Théorie générale de l'emploi, de l'intérêt et de la monnaie, Paris, Petite Bibliothèque Payot, éd 1969*
- Kocourek, R. (1991), *La langue française de la technique et de la science, Wiesbaden, Brandstetter*
- Lerat, P. (1995), *Les langues spécialisées, Paris, PUF*
- Levasseur, M. et Quintart, A. (1998), *Finance, Paris, Economica, 3^e éd.*

L'Homme, M.-C. (1995), « Définition d'une méthode de recensement et de codage des verbes en langue technique : applications en traduction » in *Technolectes et dictionnaires, Volume 8, numéro 2.*

L'Homme, M.-C. (2004), *La terminologie : principes et techniques, Montréal, Les Presses de l'Université de Montréal*

Orléan, A. (2006), *Connaissance et Finance : de l'hypothèse d'objectivité du Futur à l'hypothèse conventionnelle, version du 11 06 2006, site personnel d'André Orléan : <http://www.pse.ens.fr/orlean>*

Orléan, A. (2005), "The Self-referential Hypothesis in Finance", éd. Touffut J.P., *The Stability of Finance in Europa, Paris, Albin Michel, à paraître (site personnel d'André Orléan : <http://www.pse.ens.fr/orlean>)*

Orléan, A. (2005), *Réflexions sur l'hypothèse d'objectivité de la valeur fondamentale dans la théorie financière moderne. Représentations et croyances sur les marchés financiers, Paris, Economica, Coll. : Recherche en Gestion*

Orléan, A. (2004), « Efficience, finance comportementale et convention : une synthèse théorique. Conseil d'Analyse Economique », in Boyer R., Debove M. et Plihon D., *Les crises financières, Rapport du Conseil d'Analyse Economique, octobre 2004, Compléments A, 241-270*

Orléan, A. (2003), « Les Marchés sont-ils rationnels ? », in *La Recherche, Mai, n° 354*

Orléan, André (2001), « Comprendre les Foules Spéculatives : Mimétisme informationnel, autoréférentiel et normatif » in *Crises Financières, Paris, Economica*

Orléan, A. (1999), *Le pouvoir de la finance, Paris, Odile Jacob*

Rey, A. (1992), *La terminologie : noms et notions, Paris, PUF*

Roche, C. (2005), « Terminologie et ontologie », in *Langages, n° 157, pp. 48-62*

Roche, C. (2007a), « Le terme et le concept : fondements d'une ontoterminologie » in *Actes de la première conférence TOTh 2007,*

Roche, C. (2007b), *Plaidoirie pour une ontoterminologie, in Actes de la 12^e journée scientifique de la Cellule de recherche en linguistique, 17 novembre 2007 <http://crl.exen.fr/index.php?file=Journees&name=Novembre%202007>*

Teulié, J. et Topsacalian, P. (2000), *Finance, Paris, Vuibert, 3^e éd.*

(de) Vecchi, D. (2005) « La terminologie dans la communication de l'entreprise, approche pragmaterminologique », in *Cahiers du CIEL, Université Paris 7 EIL-A, mars 2005*, pp. 71-83

(de) Vecchi, D. (2007) « Terminologie et sciences de gestion. Le cas des entreprises : vers une pragmaterminologie » in *Actes du colloque international : Terminologie : approches transdisciplinaires, Gatineau, Québec 24 mai 2007, actes en ligne* : <http://www.uqo.ca/terminologie2007/documents/deVecchi.pdf>

(de) Vecchi, D. (2008) « La langue comme facteur d'intégration et communication en entreprise » in *Actes du colloque Langue, économie, entreprise* Université Sorbonne Nouvelle, 27, 28 & 29 mars 2008, à paraître

Dictionnaires

1. Bernard, Y. et Colli, J.-C. (1989), *Dictionnaire économique et financier*, Paris, Seuil
2. Bomati, Y. et al. (2008), *L'administration en bons termes – 1000 mots clés pour utiliser et comprendre le langage administratif*, Paris, Vuibert
3. Depecker, L. (1995), *Dictionnaire du français des métiers*, Paris, Seuil
4. Genet, J. et al. (1999), *Le jargon des postiers*, Paris, La maison du dictionnaire
5. Groupe Revue Fiduciaire (2008), *Dictionnaire comptable et financier*, Paris, Groupe Revue Fiduciaire
6. Institut français des praticiens des procédures collectives, IFPPC (2006), *Petit dictionnaire de la faillite*, Paris, La documentation française
7. Perret, P. (2002), *Le parler des métiers*, Paris, Robert Laffont
8. Le Robert (2004), *Le petit décodeur. Les mots de l'administration en clair*, Paris, Le Robert

Faut-il revisiter les *Principes terminologiques* ?

Christophe Roche

Equipe Condillac « Ingénierie des Connaissances »
Laboratoire Listic – Université de Savoie
Campus Scientifique
73 376 Le Bourget du Lac cedex
roche@univ-savoie.fr
<http://ontology.univ-savoie.fr>

Résumé :

La société numérique, en réclamant l'opérationnalisation des terminologies à des fins de traitement de l'information, a réactualisé la terminologie wüstérienne. La dimension conceptuelle retrouve une place prépondérante qu'elle avait perdue au profit d'une lexicographie de spécialité. L'émergence de la notion d'ontologie (issue de l'ingénierie des connaissances) en terminologie en est une illustration.

Si les *Principes terminologiques*, tels qu'ils sont définis dans le manuel de Felber et les normes ISO qui s'en inspirent, sont toujours au cœur de la terminologie classique, leur mise en œuvre computationnelle n'est pas sans poser quelques problèmes. Dans le cadre de cet article, nous nous attacherons, tout en rappelant leur importance, à revisiter ces *Principes* selon un double point de vue : celui de la logique, au sens mathématique du terme, et celui de l'épistémologie, au sens de la théorie de la connaissance. Nous serons ainsi amenés à clarifier et préciser certains d'entre eux afin de tenir compte des acquis de l'ingénierie des connaissances et des systèmes formels. La notion d'*ontoterminologie*, terminologie dont le système notionnel est une ontologie formelle, est le résultat d'une telle démarche.

1. Contexte

La terminologie wüstérienne a fait l'objet de nombreuses critiques [Humbley 2004], [Campenhout 2006], [Slodzian 1995]. Elle ne répondrait pas, autant qu'il le faudrait, aux attentes des utilisateurs tant d'un point de vue linguistique – la langue naturelle n'a pas vocation à être normée – que scientifique – la définition d'un terme se doit de reposer sur un système logiquement fondé. L'approche onomasiologique serait inappropriée à une démarche terminologique.

Sans rentrer dans l'étude des raisons de cet « échec » relatif – ce n'est pas l'objet de cette contribution –, il s'explique principalement par la confusion qui existe entre les discours scientifiques et la conceptualisation du domaine. Même lorsque la priorité est donnée au concept, celui-ci est trop souvent confondu avec les mots qui en parlent. Alors que distinguer concept et mot, permettrait de distinguer la *définition du concept*, en tant que spécification logique, de la *définition du terme* comme explication linguistique¹. Cela éviterait également certaines contraintes telles que la bi-univocité² qui n'a pas lieu d'être. En effet, un terme en discours, même normé, donne bien lieu à la construction d'un signifié³.

Face à ces critiques, la terminologie s'est tournée vers l'étude de sa partie la plus immédiate, la plus visible, celle des discours de spécialité⁴. Les textes et les termes (unités de discours) qu'ils contiennent sont des données objectives sur lesquelles des méthodes scientifiques peuvent être appliquées – la sémantique distributionnelle en est un exemple [Harris 1968]. Le concept a disparu au profit du mot qui en parle – et non du signe qui le définirait⁵. Aujourd'hui « être, c'est être dit et non plus être pensé » [Roche 2007a]. La terminologie est devenue une lexicographie de spécialité et relèverait de la linguistique.

Si le succès que connaît l'approche linguistique de la terminologie depuis les années 90 ne se dément pas – en particulier grâce à l'informatique qui aujourd'hui encore lui ouvre de nouvelles perspectives –, elle ne permet cependant pas de

¹ « 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].

² Si la bi-univocité facilite la communication univoque, elle n'est néanmoins pas obligatoire.

³ Le concept structure la réalité de manière stable, indépendamment de la langue – il y a ici bi-univocité entre le concept et sa dénomination. Le concept ne doit pas être confondu avec le mot qui le désigne – un terme, en texte, s'emploie comme un mot – et dont le signifié découpe de manière contingente une réalité construite en discours.

⁴ Une terminologie est l'« ensemble des désignations appartenant à une langue de spécialité » [ISO 1087], [ISO 704].

⁵ Nous distinguons la langue de spécialité qui parle du concept des langages formels qui le définissent. Rappelons que, s'il n'y pas d'expression de pensées sans signes, les lois qui régissent les langages formels ne sont pas les mêmes que celles qui gouvernent la langue naturelle.

répondre aux besoins d'une société où l'information cède de plus en plus le pas à la connaissance. La capitalisation des savoirs et des compétences, la formation, la conception de systèmes d'information, etc. sont autant de domaines qui reposent sur une conceptualisation du domaine d'activités – même dans le cas de la gestion de documents spécialisés, la solution réside davantage dans la prise en compte d'une organisation extralinguistique que dans les termes employés [Berners-Lee *et al.* 2001], [Tricot *et al.* 2006]. Cependant, les *choses* ne sont pas telles qu'elles sont *dites*. La conceptualisation d'un domaine ne peut-être extraite d'une seule analyse linguistique des textes⁶ : « dire n'est pas concevoir » [Roche 2007b]. Une lexicographie de spécialité n'a pas pour objectif de définir une modélisation d'un domaine ; elle est ici inopérante.

Les applications liées aux traitements de l'information nécessitent de plus une représentation computationnelle des concepts, c'est-à-dire une représentation manipulable par un ordinateur. La terminologie n'offrant pas les moyens de son opérationnalisation, comme nous le verrons dans cet article, cette problématique est aujourd'hui principalement abordée par l'ingénierie des connaissances. Cette dernière, à travers la notion d'ontologie, comprise comme une spécification formelle des concepts d'un domaine et des termes pour en parler [Gruber 1992], [Guarino *et al.* 1994], [Staab *et al.* 2004], réactualise la prédominance du concept. L'ontologie constitue aujourd'hui une des voies les plus prometteuses pour la représentation des systèmes notionnels [Roche 2005].

2. Terminologie

Le risque est alors grand que la terminologie, en tant que discipline autonome, disparaisse, ou soit absorbée, au profit d'une lexicographie de spécialité ou d'une ingénierie des connaissances ; la réduisant, pour la première, à une étude de phénomènes linguistiques⁷ ou, pour la seconde, à une problématique de représentation computationnelle de connaissances.

⁶ 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. Les connaissances scientifiques sont nécessaires à la compréhension des textes de spécialité, elles ne peuvent donc en être extraites : « Car on ne voit jamais personne devenir médecin par la simple étude des recueils d'ordonnances » Aristote, *Ethique à Nicomaque*, X, 10, 1181b.

⁷ Où, dans le cadre d'une sémantique distributionnelle, les mathématiques jouent un rôle de plus en plus important.

Or, la terminologie, en tant que discipline scientifique, est indispensable si on considère que son objet premier est de comprendre le monde, décrire les objets qui le peuplent et trouver les mots justes pour en parler. Bien que la terminologie classique s'inscrive dans cette finalité, force est de reconnaître qu'elle ne donne pas entièrement satisfaction tant d'un point de vue logique (en proposant des définitions consistantes) que computationnel (à travers une représentation du système notionnel qui puisse donner lieu à un calcul informatique). La terminologie classique, si elle veut continuer à exister en tant que discipline scientifique autonome, se doit de revisiter⁸ ses *Principes terminologiques*.

Notre objectif n'est pas ici d'émettre une critique de plus envers la terminologie classique. Bien au contraire, notre contribution se place délibérément sous une filiation wüstérienne par son approche scientifique. Nous souhaitons, à travers une lecture des fondements de la terminologie, montrer l'importance de la logique – la terminologie est une science – de l'épistémologie – la compréhension du monde en relève – et, cela est plus nouveau, des modèles computationnels – la terminologie doit s'opérationnaliser. A cette fin, nous étudierons les *Principes terminologiques* portant sur la construction du système notionnel sous les fourches caudines des systèmes formels⁹ – si la définition du terme relève de la langue naturelle, la définition du concept relève quant à elle d'un langage formel¹⁰. Nous essayerons, dans la mesure du possible, de comprendre pourquoi certains problèmes peuvent se poser et, le cas échéant, de proposer des formulations plus précises des *Principes*. Nous nous attacherons également à harmoniser le vocabulaire avec celui de la logique et de l'ingénierie des connaissances, disciplines devenues aujourd'hui incontournables. Notre étude portera uniquement sur les concepts, les caractères et leurs relations. La partie qui traite des termes (désignations) ne sera pas abordée ici.

⁸ Bien qu'il soit dit, à juste titre, que la *Théorie générale de la terminologie* entretient des liens étroits avec de nombreuses disciplines telles que l'ontologie, la logique et l'épistémologie, il est également précisé que « des recherches plus poussées s'imposent à cet égard » [Felber 1987].

⁹ C'est-à-dire de systèmes, tels que la logique, qui reposent sur un langage artificiel, à la syntaxe et sémantique clairement définies, pour l'écriture des formules et sur des règles de réécriture pour leur manipulation.

¹⁰ Ce qui limite le champ du connaissable aux expressions bien formées du langage.

Note : Par Principes terminologiques, en abrégé Principes, nous entendons l'étude des concepts (ou notions) et de leurs rapports. Nous nous référerons, dans le cadre de cette contribution, au Manuel de terminologie [Felber 1987] et aux normes affiliées à la terminologie classique dans leur version française, à savoir la norme [ISO 704] qui présente « les principes et méthodes de terminologie » (dans sa version d'avril 2001) et la norme [ISO 1087] (dans sa version de février 2001) qui fournit une « description systématique des concepts appartenant au domaine de la terminologie ».

3. Concept (notion)

Sachant que « tout travail terminologique a pour point de départ les notions » [Felber 1987], la première tâche consiste à définir ce que l'on entend par « concept » (ou notion).

Le *concept* est un « élément de pensée » [Felber 1987]. Comme le précise la norme française ISO 704, les concepts « sont considérés comme des représentations mentales d'*objets* dans un contexte ou un domaine spécialisé ».

Plus qu'une « unité de pensée » [ISO 704], nous proposons de définir le concept comme une *unité de compréhension*, non pas au sens de [Temmerman 2000] où se confondent classification et conceptualisation, mais au sens d'une unité permettant d'appréhender la réalité dans sa diversité. En effet, le concept est une connaissance portant sur une pluralité de choses, une « représentation intellectuelle permettant de viser le réel suivant des déterminations abstraites et générales et non dans sa singularité concrète » [Dictionnaire de philosophie 1995] – littéralement le mot de *concept* signifie « ce qui est pris ensemble » [Depecker *et al.* 2007].

Les normes ISO introduisent les notions de « concept unique » et « concept général ». La définition de « concept unique » n'est pas très claire et les exemples cités ne font qu'entretenir l'ambiguïté. Le « concept unique » désigne-t-il un concept ne décrivant qu'un seul objet, c'est-à-dire un concept dont l'extension se réduit à un seul élément ? Ou n'est-il introduit que pour uniformiser le vocabulaire et désigne-t-il l'objet lui-même ? Un objet¹¹ est une connaissance

¹¹ Nous distinguons la *chose* de l'*objet* (même si parfois nous employons indifféremment l'un ou l'autre lorsque cela ne porte pas à conséquence). Si la chose impacte nos sens, elle ne nous est connue que par l'intermédiaire de l'objet que notre intellection projette sur la chose.

singulière (même s'il n'existe qu'au regard du concept qui le fonde¹² [Heidegger 1971]) alors qu'un concept est une connaissance portant sur une pluralité de choses. Que cette pluralité se réduise à une seule entité ne signifie pas qu'un concept soit un objet, ils demeurent de nature différente et le concept reste « général ». Les notions de « concept unique » et de « concept général » ne se justifient ni d'un point de vue épistémologique, logique ou computationnel.

« Unité de compréhension » est une définition qui reste trop générale pour être opérationnelle. C'est pourquoi il est précisé qu'un concept est créé par une « combinaison unique de caractères » [ISO 1087] appelée *compréhension* ou *intension* du concept¹³. De manière duale, on appelle *extension* du concept l'ensemble des objets qui relèvent du concept – un objet relève d'un concept si et seulement si il en possède les caractères (condition nécessaire et suffisante¹⁴). L'extension d'un concept ne contient donc que des objets et aucun concept (qu'ils lui soient spécifiques ou partitifs).

4. Caractère

Le *caractère* joue un rôle essentiel en terminologie. « Propriété abstraite d'un objet » [ISO 1087], il permet de définir mais aussi de structurer et de distinguer les concepts. A cette fin, on distingue deux types de caractères, ceux qui sont essentiels, c'est-à-dire « indispensables pour comprendre le concept » [ISO 704], de ceux qui ne le sont pas.

Cette distinction, telle que définie ici et pour indispensable qu'elle soit, soulève le problème de la *nature* du caractère essentiel. Ce caractère particulier porte-t-il sur l'objet (« propriété abstraite d'un objet ») ou sur le

¹² Si dans une théorie donnée (point de vue), un objet relève bien d'un concept unique (conceptualisation) qui définit sa nature et décrit sa structure, il peut néanmoins appartenir à plusieurs ensembles selon les propriétés qu'il vérifie (classification). *Concept* et *ensemble* sont deux types différents de connaissances portant sur une pluralité de choses qu'il sera nécessaire de prendre en compte. Nous y reviendrons ultérieurement.

¹³ Notons C_p l'ensemble des concepts et C_r l'ensemble des caractères. Un concept c est défini (via l'opérateur de définition Ξ) par un ensemble de caractères (son intension) : $c \Xi \{ p_1, p_2 \dots p_n \}$, $p_i \in C_r$.

¹⁴ Notons Ob l'ensemble des objets. Un objet o est défini par un ensemble de caractères : $o \Xi \{ p_1, p_2 \dots p_k \}$, $p_i \in C_r$. Soient $o \in Ob$ et $c \in C_p$, $o \in c \Leftrightarrow \forall p_i \in c, p_i \in o$

concept (« indispensable pour comprendre le concept ») ? N'y-a-t-il pas confusion entre définition, celle du concept, et description, celle de l'objet ?

L'épistémologie nous apprend que le concept est plus qu'un ensemble de caractères communs aux objets qu'il subsume¹⁵. Il traduit également une connaissance portant sur leur nature¹⁶. En effet, *deux objets relèvent d'un même concept, non pas parce qu'ils partagent les mêmes caractéristiques, mais inversement, c'est parce que les caractéristiques essentielles du concept leur sont applicables*. Ce sont ces caractères essentiels qui font le sujet « autre » et qui, s'ils lui étaient retranchés, il ne serait plus ce qu'il est. En cela, ils ne peuvent être soumis « au plus ou au moins ». Les caractères essentiels définissent les concepts – au sens où ils permettent de comprendre la chose – et les organisent en système.

L'ingénierie des connaissances, quant à elle, insiste davantage sur la description des objets qu'à leur compréhension. Le propos ici n'est pas tant de définir l'objet que de le représenter à des fins de manipulation par un ordinateur, et en particulier de pouvoir représenter les différents états dans lesquels il peut se trouver. Un objet, connaissance singulière, se décrit comme un ensemble de caractères, appelés ici *attributs*, auxquels sont associées des valeurs. Les valeurs attachées aux attributs constituent autant de déterminations particulières de l'objet (par exemple, pour un agitateur en chimie, la *capacité d'agitation H₂O*, la *plage de vitesse* ou la *viscosité maximale du milieu*). Les attributs ne font pas le sujet « autre », mais seulement de qualité « autre ». Ils traduisent des connaissances *valuées* (des connaissances accidentelles soumises « au plus ou au moins »), dont la présence ou l'absence, si elle rend la description du sujet plus ou moins complète, ne change pas son essence. L'objet est avant tout support des attributs, l'expression de sa nature se limite en général au choix du nom du concept dont il dépend.

¹⁵ « Les faits particuliers ne sont pas simplement rassemblés : il y a un élément nouveau qui s'ajoute à la combinaison par l'acte même de pensée qui effectue la combinaison ; il y a une conception de l'esprit qui est introduite dans la proposition générale, et qui ne se trouvait dans aucun des faits observés » [Whewell 1938].

¹⁶ « La langue nomme effectivement monnaie aussi bien les pièces que les billets ; néanmoins, si la définition de l'argent parle de moyen d'échange officiellement reconnu, ce n'est pas parce que c'est le point commun aux choses désignées par la langue sous le terme de 'monnaie', mais inversement, c'est parce qu'ils sont tous deux des moyens d'échange officiellement reconnus que pièces et billets sont désignés par le terme de monnaie » [Rickert 1997].

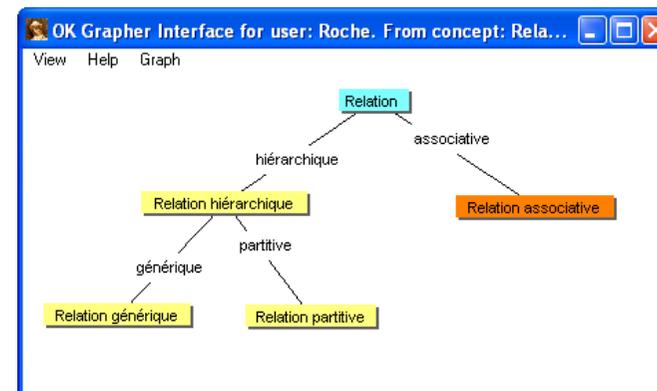
Cette différence entre les caractères est fondamentale. Elle traduit, ainsi que nous venons de le voir, des connaissances de nature différente. Les caractères essentiels définissent et structurent les concepts. Ceux qui ne le sont pas décrivent les objets et, à travers les valeurs qui leur sont attachées, les différents états dans lesquels ces objets peuvent se trouver. Appelons, afin de s'accorder avec le vocabulaire de la représentation des connaissances, *attributs* de tels caractères descriptifs, et intéressons nous de plus près aux seuls caractères essentiels.

Les caractères essentiels définissent mais aussi distinguent les concepts entre eux – définir c'est aussi délimiter, inclure autant qu'exclure. Ainsi, un caractère distinctif est un caractère essentiel utilisé pour « distinguer un concept d'autres concepts associés » [ISO 1087]. Mais un caractère peut être distinctif pour certains concepts et ne pas l'être pour d'autres. Les rapports qu'entretiennent de tels caractères avec les relations génériques doivent être précisés (comment se transmet, en particulier par la relation générique, la qualité de distinctif ou de commun aux concepts coordonnées et à ceux qui leur sont subordonnés ?). Les normes ISO introduisent de plus la notion de « type de caractère » pour désigner des caractères servant de critères de subdivision. Au-delà d'exemples guère convaincants (la couleur apparaîtra, pour de nombreux domaines, davantage comme une qualité, c'est-à-dire un attribut valué, que comme un critère de subdivision), se pose, comme précédemment, le problème de la gestion de ces caractères : sont-ils exclusifs, peut-on les combiner, comment se propagent-ils par la relation générique, etc. Toutes ces notions traduisent une même préoccupation, celle de vouloir capturer le réel dans un maillage le plus précis possible en explicitant les liens entre les éléments modélisés. Les caractères dépendants du manuel de terminologie, que les normes ISO n'ont hélas pas repris, participent de la même idée. Il reste cependant à spécifier de manière formelle ces notions et leurs liens afin de pouvoir répondre aux questions soulevées.

5. Relations

Un concept ne prend pleinement sens que dans la mesure où il s'insère dans un système de concepts [Thoiron 1996]. La norme ISO 704 précise que « les concepts n'existent pas en tant qu'unités de pensées isolées, mais sont toujours en relation les uns par rapport aux autres ».

Les *Principes* distinguent deux types de relations entre concepts¹⁷ : les relations hiérarchiques, regroupant les relations génériques et partitives, et les relations associatives. Les premières jouent un rôle central dans la mesure où elles mettent en ordre le système notionnel et nous permettent ainsi d'en appréhender et maîtriser la complexité – *la science est une mise en ordre du réel*. Les secondes traduisent un lien, considéré comme non hiérarchique, entre concepts « fondé sur l'expérience » [ISO 1087] tel que les relations de cause à effet, de producteur à produit, etc.



5.1. Relation générique

La relation générique¹⁸ est ici formellement définie. Deux concepts sont liés par une relation générique si l'intension du premier, appelé concept générique ou superordonné, est incluse dans l'intension du second, dénommé concept spécifique ou subordonné, et que cette dernière contient au moins un caractère

¹⁷ Les relations hiérarchiques et associatives sont des relations binaires liant deux concepts entre eux. Deux concepts x et y sont liés par une relation R , parfois noté xRy ou Rxy , si et seulement si le couple qu'ils forment appartient à l'ensemble des couples qui définit la relation R ($R \subseteq C_p \times C_p$). Ce que nous noterons par $(x, y) \in R$.

¹⁸ Notons R_g la relation générique ($R_g \subseteq C_p \times C_p$) et C_{re} l'ensemble des caractères essentiels ($C_{re} \subseteq C_r$). Dire que x est superordonné à y , ou inversement que y est subordonné à x , c'est dire que le couple (x,y) appartient à R_g et sera noté par l'expression $(x, y) \in R_g$. $(x, y) \in R_g \Leftrightarrow x \subset y$ et $\exists p \in y - x$ tel que $p \in C_{re}$ (un caractère distinctif est un caractère essentiel qui distingue deux concepts).

distinctif supplémentaire. Cette relation est irréflexive¹⁹, asymétrique²⁰ et transitive²¹. Elle définit donc un ordre strict – une hiérarchie – sur des concepts, ordre non total dans la mesure où il peut exister des concepts non comparables.

Intension et extension entretiennent un rapport inverse. Par la définition de l'appartenance d'un objet à un concept (voir la note 14), l'extension d'un concept spécifique est incluse dans les extensions des concepts qui lui sont génériques. Ainsi, un objet relevant d'un concept donné relève également des concepts qui lui sont superordonnés. La relation générique ordonne les concepts de nature comparable.

Cependant, si l'inclusion de l'extension d'un concept dans l'extension d'un autre implique bien que l'intension du premier contient celle du second, cela ne suffit pas, au regard de la définition que donnent les *Principes* de la relation générique, à ordonner les concepts. En effet, rien ne garantit l'existence d'un caractère distinctif. Dans une telle situation, comment interpréter le fait qu'un même objet puisse relever de deux concepts différents qui n'entretiennent aucun lien hiérarchique ?

La définition de la subsomption (relation générique), pour laquelle l'inclusion des intensions est une condition nécessaire et suffisante, requiert des ensembles homogènes de caractères (ou du moins une définition qui garantirait cette condition). Il est donc impératif, si l'on veut conserver la distinction entre caractères essentiels et ceux qui ne le sont pas, de séparer la définition du concept (en termes de caractères essentiels) sur laquelle porterait la subsomption, de sa description (exprimée sous la forme d'attributs valués).

5.2. Relation partitive

Il est souvent plus facile de décrire la chose telle qu'on la perçoit, telle qu'elle nous apparaît à travers les éléments qui la composent, que de définir ce qu'elle est (sa nature). La relation partitive, ou mérologique, joue donc un rôle important.

¹⁹ $\forall x \in Cp, (x, x) \notin Rg$ (un concept ne peut être générique, respectivement spécifique, à lui-même).

²⁰ $\forall x, y \in Cp, (x, y) \in Rg \Rightarrow (y, x) \notin Rg$ (un concept superordonné à un autre ne peut lui être subordonné).

²¹ $\forall x, y, z \in Cp, (x, y) \in Rg \text{ et } (y, z) \in Rg \Rightarrow (x, z) \in Rg$ (si x est superordonné à y et que y est lui-même superordonné à z alors x est superordonné à z).

Elle exprime une relation interne entre un tout, le concept intégrant, et ses parties, les concepts partitifs, sans imposer de contrainte particulière quant à la nature de ses constituants. Les *Principes* qualifient la relation partitive de hiérarchique afin de traduire l'idée qu'une chose peut être comprise par niveaux de détails de plus en plus fins ou de plus en plus englobants. Cependant, la relation partitive n'est pas une relation d'ordre²². Il n'y a pas subordination de la partie au tout, comme il y a subordination de l'espèce au genre. Si ce qui est affirmé du genre l'est pour ses espèces, ce n'est le cas ni du tout pour ses parties ou inversement des parties pour le tout qui les contient.

Faire de la relation partitive une relation définitoire et non uniquement descriptive soulève de nombreux problèmes. C'est prendre le risque de confondre les caractères du concept intégrant avec ses concepts partitifs. Même si on peut identifier des parties « essentielles » (au sens d'obligatoire) et des parties « distinctives » (qui différencieraient deux concepts intégrants), un concept partitif n'est pas un caractère – de manière générale, un concept n'est pas un caractère, et réciproquement, un caractère n'est pas un concept²³. Une telle approche nécessite de préciser les rapports, s'il en existe, entre la compréhension d'un concept intégrant et, d'une part, sa structure et, d'autre part, les compréhensions de ses parties. Elle nécessite également de clarifier ce que peut être un concept partitif obligatoire composé uniquement de parties facultatives (comment la relation partitive propage-t-elle, si elle le fait, les propriétés d'obligatoire et de facultatif pour les constituants ?), si l'existence d'une partie obligatoire se traduit par un caractère essentiel du tout, etc. Si les *Principes* proposent, pour la description des concepts, des idées intéressantes sur la base de la relation partitive, il reste à les spécifier de manière plus formelle.

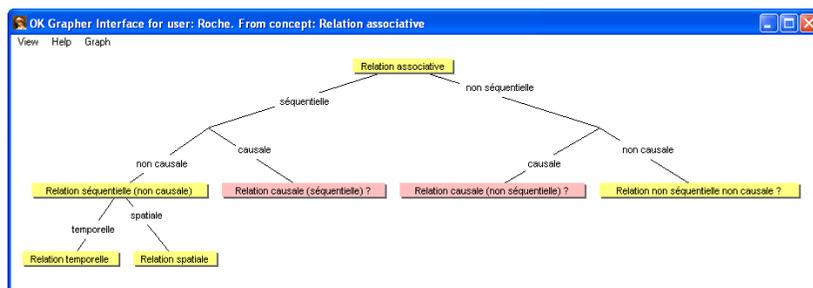
²² Notons $Part(x)$ l'ensemble des parties de x ($Part(x) \subseteq Cp$) et Rp la relation partitive ($Rp \subseteq Cp \times Cp$). On dit que x contient y, ou que y est une partie de x, noté $(x, y) \in Rp$, si et seulement si $y \in Part(x)$. De cette définition ne découle aucune propriété des relations d'ordre. Cependant, si nous pouvons poser l'irréflexivité de Rp – le tout ne peut se contenir lui-même, et l'asymétrie – un élément ne peut à la fois contenir un élément et être contenu dans ce même élément – la transitivité ne peut être affirmée – les niveaux hiérarchiques s'effaceraient.

²³ Même si on aimerait pouvoir désigner l'ensemble des objets possédant un caractère particulier. Ceci est un autre problème que la logique permet de résoudre de manière élégante.

5.3. Relations associatives

Les relations associatives sont des relations externes entre concepts (externes au sens où elles ne sont pas nécessaires à la compréhension des concepts liés), non hiérarchiques (ni génériques, ni partitives), fondées sur l'expérience. Elles n'imposent aucune contrainte a priori quant à la nature des concepts liés. Les relations associatives sont des relations binaires, ce qui implique que les relations de plus grande arité doivent être traduites en un ensemble de relations dyadiques.

La norme ISO 1087 donne quelques exemples de relations associatives : les relations séquentielles, temporelles et causales. On regrette l'absence de caractères distinctifs dans la définition de ces relations. Ils auraient pu lever certaines ambiguïtés. Ainsi, une relation causale est-elle également séquentielle ? Les exemples fournis le laissent penser (« action » et « réaction », « explosion nucléaire » et « retombées »). Sachant qu'il faudrait alors se poser la question de la proximité spatiale ou temporelle (une relation séquentielle est une « relation associative fondée sur la proximité spatiale ou temporelle » [ISO 1087]).



5.4. Systèmes de concepts

Les différents types de relations (hiérarchiques et associatives) structurent le champ conceptuel (« groupe non structuré de concepts » [ISO 1087]) en autant de systèmes de concepts : générique, partitif, associatif et mixte (en combinant les relations). L'élaboration de ces systèmes repose principalement sur l'analyse de la compréhension et de l'extension des concepts, ce qui suppose qu'elles leur sont antérieures.

5.5. Relations ontologiques

Le terme « ontologie » n'apparaît pas dans les normes ISO 704 et 1087. Le manuel de terminologie parle de rapports ontologiques pour désigner des

rapports indirects entre les notions et en particulier les relations partitives (mérologiques).

L'ontologie n'a pas la même signification pour l'ingénierie des connaissances qui la définit comme une spécification formelle des concepts et de leurs relations²⁴. La relation la plus importante est alors celle de subsomption (relation générique), là où la terminologie [Felber 1984] parle d'un rapport logique (subordination logique). Sachant que l'ontologie constitue aujourd'hui une thématique à part entière et une des voies les plus fructueuses pour la modélisation et la représentation computationnelle du système notionnel, nous préconisons l'emploi du mot ontologie au sens de l'ingénierie des connaissances. Et ce d'autant plus que des objets ne peuvent être mis en relation qu'après avoir été définis, c'est-à-dire après que l'ontologie, au sens étymologique du mot, ait été construite.

6. Définitions

Pour les *Principes*, les définitions sont des « représentations d'un concept » [ISO 1087] et « doivent refléter le système de concepts » [ISO 704]. Les définitions ne créeraient²⁵ ni les concepts – ils le sont par une « combinaison unique de caractères²⁶ » –, ni le système notionnel. Les définitions sont données a posteriori, lorsque les combinaisons uniques de caractères ont été construites et les relations identifiées.

On distingue ici deux types de définitions, par compréhension et par extension.

²⁴ L'ontologie, dans son sens premier, est la science de l'être en tant qu'être, indépendamment de ses déterminations particulières. Cette définition est plus proche de l'ingénierie des connaissances – même s'il existe ici une certaine confusion entre « être » et « existence » – que de la terminologie.

²⁵ A comparer avec la création de l'espèce par sa définition en genre prochain et différence spécifique.

²⁶ Précisons cependant que, si toute combinaison unique de caractères crée un concept, celui-ci n'est pas nécessairement porteur de sens pour le domaine.

6.1. Définition par compréhension

La définition par compréhension, semblable à une définition en genre et différence, se compose du concept superordonné immédiatement supérieur suivi du ou des caractères distinctifs. Mais, un même concept pouvant être subordonné directement à plusieurs concepts, lequel choisir²⁷ ? Que des définitions différentes puissent aboutir à une même combinaison unique de caractères n'est, en soi, pas gênant. Les *Principes* demandent également de préciser les caractères qui distinguent un concept de ceux qui lui sont coordonnés. L'intérêt de cette demande est d'autant plus limité que les caractères distinctifs des concepts coordonnés vis-à-vis de leur concept supérieur commun distinguent ipso facto les concepts coordonnés entre eux²⁸ (ou du moins devraient, dans une conceptualisation bien pensée, les distinguer).

Les définitions reposant sur une relation partitive ou une relation associative ne sont pas satisfaisantes d'un point de vue formel dans la mesure où l'élaboration de la combinaison unique de caractères définissant le concept ainsi construit n'est pas clairement spécifiée. Quelle est la signification d'un caractère distinctif dans le cas d'une relation partitive ? Peut-on comparer les combinaisons de caractères d'un tout et de ses parties ? Des parties entre elles ? De concepts intégrants entre eux dans leur structure ? Et que de dire des « caractères qui qualifient la relation » associative [ISO 704] ?

6.2. Définition par extension

La définition par extension d'un concept générique ou intégrant consiste à énumérer les concepts qui lui seront subordonnés (spécifiques dans le cas d'un concept générique, partitifs pour le concept intégrant). Ce type de définition ne doit pas être confondue avec la définition en extension d'un ensemble (respectivement d'un concept) qui consiste à énumérer les objets qui composent l'ensemble (respectivement qui relèvent du concept). C'est pourquoi nous parlerons également, afin d'éviter toute confusion, de définition par énumération.

Si l'on comprend bien l'intérêt de ce type de définition, elle soulève immédiatement le problème de la combinaison unique de caractères qui

²⁷ Soient c_1 , c_2 et $c_3 \in C_p$ et $c_1 = \{a\}$, $c_2 = \{b\}$, $c_3 = \{a, b\}$. c_1 et c_2 sont deux concepts immédiatement supérieurs à c_3 et donnent lieu à deux définitions différentes de c_3 .

²⁸ Ce que fait la définition aristotélicienne en genre prochain et différence spécifique.

identifierait le concept ainsi défini. Comment peut-elle se construire à partir des combinaisons de caractères des concepts subordonnés ?

7. Rapports logiques

La définition par extension (par énumération) d'un concept générique n'est pas sans rappeler la disjonction de notions du manuel de terminologie – définie dans ce cas à partir des extensions des notions. Cette dernière soulève des problèmes identiques quant à la détermination des caractères du concept générique ainsi construit. Mais l'idée est la même : pouvoir construire de nouveaux concepts, qu'ils soient génériques, spécifiques ou intégrants, à partir de concepts existants. La conjonction de notions, absente des normes ISO, est intéressante à ce titre. Cependant, définir l'intension du concept spécifique résultant comme l'union des intensions n'échappe pas à certaines contradictions. Comment interpréter un concept issu de la conjonction de deux concepts coordonnés (c'est-à-dire issus d'un même concept générique) qui se différencieraient par un caractère distinctif (peut-on, à la fois, être une chose et son contraire) ?

8. Conclusion et perspectives

Les *Principes terminologiques* insistent, avec raison, sur l'importance de la conceptualisation du domaine comme fondement de la terminologie – il ne peut y avoir de terminologie sans connaissances. La construction du système notionnel est cependant ardue. Elle l'est d'autant plus qu'elle doit expliciter des connaissances souvent tacites qui, rarement décrites dans les documents scientifiques et techniques, rendent la présence des experts et leur collaboration au travail terminologique indispensables. Afin de nous aider dans cette tâche, les *Principes* proposent un certain nombre de paradigmes qui traduisent une volonté scientifique de mise en ordre du réel – un structuralisme mathématique – à travers des systèmes de concepts mis en relations.

Cependant, la définition de ces paradigmes souffre d'imprécisions qui rendent l'opérationnalisation des terminologies difficile et expliquent l'ascendance de l'ingénierie des connaissances dans ce domaine – du seul point de vue computationnel, les *Principes* doivent être revisités. Ces imprécisions trouvent leur

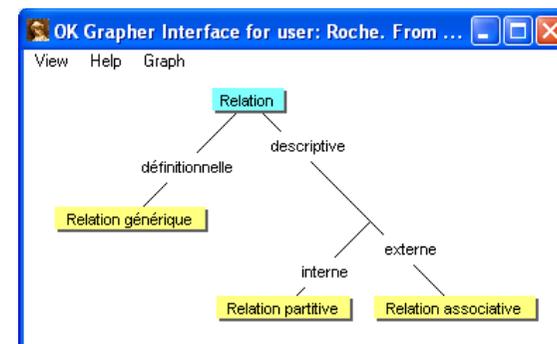
origine dans l'utilisation de la langue naturelle²⁹ inadaptée à décrire une dimension qui lui est étrangère. La définition des *Principes* doit suivre la même rigueur avec laquelle ces *Principes* aspirent à ordonner le réel. L'utilisation d'un langage formel à la syntaxe et à la sémantique clairement définies est inévitable, non seulement pour lever les ambiguïtés, mais aussi parce que la conceptualisation est une activité scientifique. Les langages de représentation issus de l'intelligence artificielle permettront in fine, avec le même esprit de formalisation et de précision mais dans un registre différent et complémentaire, l'opérationnalisation des terminologies sur la base d'une spécification logique du système notionnel.

Les logiques [Baader *et al.* 2003] et les langages de l'intelligence artificielle [Sowa 2000], [Brachman *et al.* 1985] sont avant tout des systèmes de représentation. Leur finalité n'est pas de comprendre³⁰ le monde, mais de le décrire, de manière formelle pour les premiers, à des fins de calcul par un ordinateur pour les seconds. Leur utilisation, dans le cadre de la construction du système notionnel, doit être guidée par des principes épistémologiques qu'il reste à préciser. Ces principes doivent être posés en gardant à l'esprit cet objectif de formalisation et d'opérationnalisation. Indiquons, pour conclure, quelques pistes en ce sens.

²⁹ « Si c'est une tâche de la philosophie de rompre la domination du mot sur l'esprit humain en dévoilant les illusions qui souvent naissent presque inévitablement de l'utilisation de la langue pour l'expression de relations entre des concepts, et en libérant la pensée de ce dont elle est atteinte uniquement par la nature du moyen d'expression linguistique, alors mon idéographie, développée plus avant pour ces buts, pourra devenir un outil utile aux philosophes. Elle ne restitue assurément pas la pensée purement, étant donné que cela n'est guère possible par un moyen extrinsèque de représentation ; mais d'une part, on peut restreindre ces écarts à l'inévitable et l'inoffensif, d'autre part, le fait qu'ils soient d'une toute autre sorte que ceux qui sont propres à la langue, procure déjà une protection contre une influence unilatérale de l'un de ces moyens d'expression » G. Frege, *L'idéographie*.

³⁰ Comme tout langage, les langages formels découpent la réalité selon des structures qui leurs sont propres. Ainsi, la logique du 1^{er} ordre le fait en termes de prédicats sur lesquels s'appliquent des calculs formellement définis (calcul des prédicats). Cependant, un tel langage ne sait traduire directement certaines différences fondamentales. Il utilise le même formalisme pour représenter aussi bien des caractères essentiels (par exemple Homme(x)) qu'accidentels (Malade(x)).

La notion de concept, en tant qu'unité de compréhension, permet d'appréhender la diversité des objets qui peuplent la réalité. Il aurait, à cette fin, une double fonction : celle de comprendre ce qu'est une chose et celle de la décrire. La première définit la chose en termes de caractères essentiels³¹. La seconde la décrit comme une connaissance singulière, sous la forme d'attributs valués^{32,33}. Les caractères essentiels sont issus de la raison. Ils participent à la définition des concepts et les structurent en un système – un squelette – sur lequel sont projetés les attributs qui décrivent les objets tels qu'ils sont perçus et dont les valeurs traduisent des connaissances contingentes. *Définition et description sont deux notions qu'il est important de distinguer*. Elles se retrouvent dans la typologie des relations elles-mêmes. Les définitions ne se limitent pas à traduire, a posteriori, une structure préalablement construite. Elles créent³⁴, dans une même opération, les concepts et la structure.



³¹ Un caractère est dit essentiel pour un objet lorsque, retranché de l'objet, celui-ci n'est plus ce qu'il est.

³² Contrairement à un caractère essentiel, retrancher un attribut de l'objet ne change pas sa nature, mais le rend seulement de description incomplète.

³³ Nous préférons l'expression d'« attribut valué » à celui de « caractère inhérent » (issu du vocabulaire de la terminologie classique) afin d'insister sur ce qui, dans l'objet, relève du descriptif et du contingent.

³⁴ Si on considère que la définition d'un concept se fait par rapport à des notions antérieurement définies, plus connaissables que le definiendum.

Si tout objet relève d'un concept, il peut également appartenir à des ensembles différents³⁵. L'ensemble, tout comme le concept, est une connaissance portant sur une pluralité de choses. Il s'en diffère néanmoins dans la mesure où les objets qu'il regroupe peuvent être de nature différente, à condition qu'ils satisfassent la même loi. La notion de concept se trouve enrichie de celle de définition intensionnelle d'un ensemble. Il ne se limite plus à une combinaison unique de caractères mais devient une fonction à valeur de vérité. Le concept est prédicatif [Frege 1971], la logique lui ouvre ses portes. Il se combine alors à l'infini, par conjonction, disjonction, négation etc. pour devenir une formule bien formée.

L'apparition de nouveaux paradigmes traduit cette volonté de vouloir intégrer les différentes sources auxquelles puise la terminologie. Ainsi, l'*ontoterminologie* [Roche 2007a], terminologie dont le système notionnel est une ontologie formelle, insiste sur l'importance des principes épistémologiques qui président à la conceptualisation d'un domaine – 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 et normés) qui en parlent, tout en distinguant les définitions formelles des concepts (spécifications logiques) des définitions en langue naturelle des termes (explications linguistiques).

³⁵ Les ensembles peuvent être distincts, se contenir, se recouvrir, se chevaucher, laissant libre cours à un découpage moins stricte de la réalité que ne le feraient les concepts. Conceptualisation et classification sont deux opérations différentes de l'esprit trop souvent confondues.

Bibliographie

- Aristote (1997), *Ethique à Nicomaque (Vrin)*.
- Baader F., Calvanese D., McGuinness D., Nardi D. and Patel-Schneider P. (2003), *The Description Logic Handbook (Cambridge University Press)*.
- Berners-Lee T., Hendler J. and Lassila O. (2001), « *The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities* », *Scientific American Magazine*, May 17.
- Brachman R.J., Levesque H.J. (1985), « *Readings in Knowledge Representation* », Morgan Kaufmann Publishers, Inc. 1985.
- Campehondt M. Van (2006), « *Que nous reste-t-il d'Eugen Wüster ?* », *Colloque international Eugen Wüster et la terminologie de l'Ecole de Vienne, Paris, 3-4 février 2006*.
- Depecker L., Roche C. (2007), « *Entre idée et concept : vers l'ontologie* », *Revue Langages*, 168, décembre 2007, pp. 106-114 (Éditions Larousse).
- Dictionnaire de Philosophie (1995)*, Baraquin N., Baudart A., Dugué J., Laffitte J., Ribes F. et Wilfert J. (Armand Colin).
- Felber H. (1984), *Manuel de terminologie (Paris, Unesco)*.
- Frege G. (1971), *Écrits logiques et philosophiques (Paris : Éditions du Seuil)*.
- Frege G. (2000), *Idéographie (Vrin)*.
- Gruber (1992), « *A Translation Approach to Portable Ontology Specifications* », *Knowledge Acquisition*, 5(2), pp. 199-220.
- Guarino N., Carrara M. and Giaretta P. (1994), « *An Ontology of Meta-Level Categories of Knowledge Representation and Reasoning* », *Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning (KR94) (Morgan Kaufmann)*.
- Harris Z.S. (1968), *Mathematical Structures of Language (R.E. Krieger Publishing Company, Inc)*.
- Heidegger M. (1971), *Qu'est-ce qu'une chose ? (Gallimard, « Tel »)*.
- Humbley J. (2004), « *La réception de l'œuvre d'Eugen Wüster dans les pays de langue française* », *Cahier du C.I.E.L. 2004*, pp 33-51.

ISO 704 (2001), *Travail terminologique*. ISSN 0335-3931.

ISO 1087-1 (2001), *Vocabulaire*. ISSN 0335-3931.

Lerat P. (1995), *Les langues spécialisées*, Paris, PUF.

Rickert H. (1997), *Théorie de la définition* (Gallimard, « nrf »).

Roche C. (2005), « Terminologie et ontologie », *Revue Langages*, 157, mars 2005, pp. 48-62 (Éditions Larousse).

Roche C. (2007 a), « Le terme et le concept : fondements d'une ontoterminologie », *Conférence TOTh 2007, Terminologie & Ontologie : Théories et Applications, Anney 1^{er} juin 2007*, pp. 1-22.

Roche C. (2007 b), « Dire n'est pas concevoir », *18èmes journées francophones d'Ingénierie des Connaissances, Grenoble 4-6 juillet 2007*, pp. 157-168.

Slodzian M. (1995), « Comment revisiter la doctrine terminologique aujourd'hui », *La banque des mots*, n°7-1995

Sowa J. (2000), *Knowledge Representation*, (Brooks/Cole).

Staab S. and Studer R. (2004), *Handbook on Ontologies* (Springer).

Temmerman R. (2000), *Towards new ways of terminology description*, (Benjamins Publishing).

Thoiron P., Arnaud P., Béjoint H. et Boisson C.P. (1996), « Notion 'd'archi-concept' et dénomination », *Meta*, vol. 41, n°4, pp. 512-524.

Tricot C., Roche C., Foveau C. et Reguigui S. (2006), « Cartographie sémantique de fonds numériques scientifiques et techniques », *Document Numérique : Visualisation pour les bibliothèques numériques*, 9 (2), 2006, pp. 13-36.

Whewell W. (1938), *De la construction de la science*, (Vrin).

Propositions pour un réseau conceptuel des instruments de mesure œnologiques

Pierre Lerat

UMR 7187 LDI, Université Paris XIII
34, rue N.D. de Recouvrance, F – 45000 Orléans
pierre.lerat@wanadoo.fr

Résumé :

Les qualités attendues d'une terminologie sont avant tout celles qui permettent une gestion cohérente des connaissances lexicalisées. C'est le mode de gestion des relations entre concepts qui est essentiel pour élaborer un réseau conceptuel. Il importe donc de commencer par distinguer clairement (graphiquement) les concepts et les termes. Il faut ensuite concevoir une méthode d'élaboration des réseaux conceptuels en langues naturelles (modèle « opérateur-argument » de Harris). Il faut enfin mettre à l'épreuve sa méthode sur un échantillon d'exemples. C'est ce qui est proposé ici, à propos de trois concepts reliés entre eux: « densimètre électronique », « lecture » et « saccharose ». Le corpus de base est le *Recueil des méthodes internationales d'analyse des moûts et des vins* (Paris, OIV, 2008), qui fait autorité en la matière.

Mots-clés : réseau conceptuel en langues naturelles, opérateur, argument

1. Introduction

Il y a réseau conceptuel et réseau conceptuel, selon que l'orientation est lexico-sémantique ou ontologique formelle¹. Les propositions faites ci-dessous sont à la charnière de la terminologie et de l'ontologie : elles utilisent comme unités des concepts terminologiques, définis par analyse d'un corpus mais non formalisés, formulés dans des textes par des termes préférentiels et leurs synonymes. Ces concepts terminologiques, c'est-à-dire spécialisés, sont reliés entre eux en un réseau conceptuel en langues naturelles (au pluriel, puisque peu importe la langue de travail si les définitions sont assez partagées pour autoriser leur traduction) ; il faut aller plus loin dans le travail logique si l'on veut

¹ Voir notamment, dans ce recueil, le texte de Sylvie Després.

opérationnaliser de tels résultats en vue de traitements non linguistiques d'informations, et le terminologue linguiste n'a pas compétence pour le faire, mais il est permis de penser que l'identification rigoureuse des concepts peut à tout le moins constituer un préalable utile dans ce cas ².

Si l'on admet avec l'équipe de SUVA³ que les données sont des représentations symboliques de faits (représentations dont les langues naturelles sont un exemple privilégié), que les informations sont des données mises en contexte et les connaissances des informations reliées entre elles, on peut dire que les textes spécialisés véhiculent des connaissances spécialisées, où l'analyse des contextes conduit à des informations, dont la décontextualisation aboutit à des données, on voit clairement pourquoi l'expertise du domaine ou du métier est nécessaire : sous les mots il y a des informations, sous les informations des connaissances.

La terminologie à base de concepts, dans l'esprit du *Dictionnaire de la machine-outil* de Wüster, permet en effet d'identifier des concepts spécialisés clairement distincts dès lors que la matière s'y prête. Elle rend également possibles des terminologies de domaines, de métiers et de tâches, pour peu que les définitions, les sources, les dénominations et les concepts connexes soient exploités selon des « dimensions » appropriées (au sens de « type de mise en relation des connaissances », dans la conception de l'équipe de SUVA).

La collecte des termes passe nécessairement par un tri dans les séquences de mots. Le filtre proposé ici est ce que j'appelle « collocation conceptuelle », c'est-à-dire « les relations typiques entre un prédicat et un objet auquel il est applicable » (Lerat, 1995 : 104) et effectivement appliqué dans le corpus. Et comme *prédicat* est dangereusement polysémique dans nos disciplines, il sera plus clair d'utiliser la terminologie du fondateur de cette approche : *opérateur / argument* (Harris, 1971, 1976 et 2007).

Comment constituer un réseau conceptuel qui ne soit pas un artefact de corpus dans une langue donnée ? Il faut d'abord distinguer clairement concepts et termes, ce à quoi invite la terminologie classique. Il faut ensuite ne pas se

limiter aux relations hiérarchiques, ce qu'autorise une approche fondée sur une distinction entre expressions saturées et expressions insaturées. Il importe également de ne pas être esclave du corpus ; en particulier, seule la veille terminologique permet de relier les concepts du réseau à autant de définitions quand le corpus ne contient pas toutes les définitions souhaitables, c'est-à-dire le plus souvent ⁴. Enfin, il ne faut pas non plus être esclave de la surface syntaxique des textes, du fait de la très grande variabilité des formulations conceptuellement équivalentes, mais prendre en compte les « collocations conceptuelles » au sens précisé ci-dessus.

Après une présentation de ces principes, il est proposé ici, à titre d'illustration de la méthode, un échantillon concernant trois concepts connexes considérés dans les limites d'un domaine (l'œnologie) et d'un sous-domaine (les méthodes d'analyse) : « densimètre électronique », « lecture » et « saccharose ». Le corpus où sont puisées les concordances est une « publication normative » ⁵ ; c'est l'autorité mondiale en la matière, dont les recommandations sont intégrées aux règlements de l'UE portant sur l'œnologie. Le consensus sur les concepts, nécessaire pour une application impérative dans tous les États membres, garantit en l'occurrence la validité des équivalences terme à terme d'une langue à l'autre. Ces équivalences se trouvent dans les versions parallèles des règlements européens. Ainsi, la polysémie du mot *lecture* et de ses équivalents ne nuit nullement à la participation de « lecture » à un réseau conceptuel des méthodes d'analyse en œnologie, comme on le verra ; de même, l'interdisciplinarité de ce carrefour de la chimie, de la technologie et de l'agro-alimentaire n'empêche nullement la cohérence d'une terminologie à base d'ontologie professionnelle.

Enfin, pour illustrer la nécessité de distinguer clairement terme et concept, l'exemple de *concentration* met en évidence l'intérêt de prendre en compte d'un côté 3 jeux de relations correspondant à 3 concepts et de l'autre les chaînes de caractères cooccurrentes du mot dans ces 3 types d'emplois. Seul l'un de ces derniers met en jeu des instruments de mesure, mais les 3 concernent l'œnologie, où la synonymie et les différences dans les traductions montrent que la

² L'usage d'arborescences, comme dans la communication de de Vecchi et Estachy, est à tout le moins une bonne visualisation des relations, donc une aide à la formalisation.

³ Ces distinctions importantes sont empruntées à Bleuer, Bösch et Ludwig (voir leur texte dans ce recueil).

⁴ Pour établir la terminologie d'un texte spécialisé, il faut nécessairement prendre en compte « la place du non dit » (Costa et Silva, dans ce recueil).

⁵ *Recueil des méthodes internationales d'analyse des vins et des moûts*, Paris, Organisation Internationale de la Vigne et du Vin (OIV), 2008, www.oiv.org/publications (téléchargeable au format .pdf).

distinction des domaines et sous-domaines n'est pas une panacée en terminologie.

2. Principes de terminologie générale

Le concept terminologique est un objet construit et non pas une conceptualisation individuelle par un « je-ici-maintenant », ni même le sens d'une forme lexicale. C'est un contenu de connaissances normé et nommé de façon consensuelle. Ses caractéristiques sont les suivantes :

- il est interlinguistique

Ex. 1 : fr. « titre alcoométrique volumique acquis » = es. « grado alcohólico volumétrico adquirido » = en. « actual alcoholic strength by volume » = de. « vorhandener Alkoholgehalt »

Ici le réseau conceptuel est construit à partir de dénominations conventionnelles en français parce que notre colloque est francophone ; s'il était germanophone le concept serait appelé « vorhandener Alkoholgehalt ». Ce qui compte est le rapport nécessaire avec la définition traduite dans chacune des langues de l'UE. : il s'agit d'une grandeur égale au « nombre de volumes d'alcool pur à une température de 20 °C contenus dans 100 volumes du produit considéré à cette température » (R07). Certes, toute dénomination spécialisée ne se traduit pas aussi commodément. Dans les sciences humaines et sociales, ainsi que l'a observé depuis longtemps Alain Rey (1979 : 45) et avant lui Eugen Wüster (1985 : 90), la terminologie est difficilement séparable des connaissances générales et des cultures nationales. Il n'est pas pour autant interdit de parler de terminologie juridique, si l'on est conscient de sa spécificité : *Assembleia* en portugais (exemple de Costa et Silva), comme *Assemblée Nationale* en français, est définissable dans toute langue mais non traduisible, *Parlement européen* est définissable et traduisible dans toutes les langues de l'UE, *Assemblée Générale des Nations Unies* dans toutes les langues de l'ONU. Aussi bien, le service de traduction de la Cour de Justice Européenne de Luxembourg a un seul terme en toute langue pour ce qui dénomme l'institution italienne *Corte d'appello* (en italien quelle que soit la version), dont la définition n'est pas la même que pour *Cour d'appel* (dénomination française dans toute langue de l'UE).

- dans une même langue les concepts peuvent être exprimés au moyen de plusieurs dénominations synonymes

Ex. 2 : en œnologie, « enrichissement » a pour dénominations es. *aumento artificial del grado alcohólico natural* (R07), mais aussi *enriquecimiento* (R90)

ils peuvent être exprimés à l'aide d'un élément de nomenclature alphanumérique (comme les formules chimiques) ou numériques (comme ci-dessous)

Ex. 3 : « jus de raisin » correspond à *Code NC 2209 61* et à *jus de raisin* (R07)

il renvoie à une entité qui, si elle est concrète, peut faire l'objet de représentations iconiques

Ex. 4 : « aréomètre », « balance hydrostatique », « densimètre », « pycnomètre », « réfractomètre » dans les catalogues de fabricants ou de distributeurs

- il est gagé sur des sources textuelles (ici, *Recueil*, R90 etc.)

3. Insaturation et saturation

3.1. Logique et langue

Il importe de ne pas confondre les concepts et les dénominations. Or les meilleurs s'y laissent prendre. Ainsi, Harris, qui utilise le couple « opérateur / argument », déclare que « l'opérateur est un constructeur d'assertions, approximativement un prédicat par rapport à ses arguments » (1976 : 26), mais également que « opérateurs et arguments sont des mots, un verbe comme *manger* est un opérateur à deux arguments » (p. 7). Non : dans le premier cas, « manger » introduit une affirmation, il a une « fonction assertive consistant à doter l'énoncé d'un prédicat de réalité » (Benvéniste, 1966 : 154), dans le second c'est une expression prédicative (un verbe), non un prédicat logique, et il sert à construire une phrase en français. Pour matérialiser la distinction, j'écrirai entre guillemets les concepts et en italiques minuscules les expressions.

En outre, pour assurer un statut univoque à des dénominations utilisées aussi ailleurs, par exemple en économie (comme *concentration* ou *valeur*), il convient de les spécifier. Une autre caractéristique d'une bonne dénomination de concept est son usage effectif dans les textes normatifs. Il importe également que leur choix ne soit pas exclusivement dépendant du corpus, car les classes gagnent à être assez générales pour permettre leur réemploi. Ainsi, le densimètre n'est pas

seulement un « instrument de mesure œnologique », mais plus largement un « instrument de mesure ».

Ex. 5 : *liquide* -> « corps liquide », *solution* -> « solution chimique », *valeur* -> « valeur de grandeur »

Enfin, il est également nécessaire, pour construire des réseaux conceptuels cohérents, de relier tout concept à son générique le plus bas dans la hiérarchie.

Ex. 6 : « densimètre électronique » /GÉNÉRIQUE : « densimètre »/SPÉCIFIQUE : « densimètre électronique à résonateur de flexion »

Ce qui caractérise en effet la définition spécialisée, c'est le recours à l'« hyperonyme de niveau immédiatement supérieur » (Lerat, 1995 : 182).

3.2. Les collocations conceptuelles

Si l'on admet, avec Harris, que des prépositions comme *sur* et *pour* sont des « indicateurs d'arguments » (1976 : 26), il faut leur reconnaître un flou conceptuel que n'ont pas en général les concepts spécialisés. Il en va de même pour les autres prépositions.

Ex. 7: « sont déterminées (...) *par* aréométrie ou densimétrie *par* la balance hydrostatique » (R90)

Le premier *par* introduit une méthode, le second un moyen.

Ex. 8 : *densimétrie* sur la balance hydrostatique, *densimétrie* par la balance hydrostatique, *densimétrie* utilisant la balance hydrostatique (R90)

Une grammaire des cas y perdrait son latin : le lieu et le moyen ne sont pas distingués, et quand on veut clarifier on a recours à un verbe.

Ce qui est à retenir du point de vue de la collocation conceptuelle est le fait que pour spécifier le concept de « densimétrie » on ajoute une expression insaturée (utilisant appelle un complément) ou un « indicateur d'argument » suivi d'un nom d'entité saturé (c'est-à-dire n'ayant pas besoin d'être complété). Ce qui compte ici, c'est donc uniquement le lien (conceptuel) entre la méthode et l'instrument, non la manière (linguistique) de l'exprimer. De la même façon, le concept de « densimétrie » se lexicalise en français au moyen de deux expressions saturées : *mesure de densité* et *densimétrie*.

3.3. Les concepts non saturés

Là où il existe une nominalisation déverbale de même sens que le verbe correspondant et utilisée dans le corpus, c'est elle qui sera retenue pour nommer un concept. Les raisons ne sont pas philosophiques mais pratiques. La principale est statistique : dans les textes spécialisés, les formes nominales dominent, même pour exprimer des actions (voir notamment Lerat, 2007). C'est nettement le cas pour les opérations vitivinicoles dans un corpus en espagnol⁶; ici les formes sont en concurrence plus égale. Le type de discours joue en effet un rôle : *mesurer*, *calculer*, *déterminer* et *lire* sont des consignes empruntées telles quelles au *Recueil*.

Ex. 9: *mesure(s)* (nom, y compris dans des locutions figées) 143 fois dans R90, *calcul(s)* 79 fois, *détermination(s)* 80 fois, *lecture(s)* 11 fois / *mesurer* (et ses formes conjuguées) 79 fois, *calculer* 46 fois, *déterminer* 70 fois, *lire* 17 fois

Il en irait de même pour les nominalisations d'adjectifs : on ne rencontre guère dans le corpus les adjectifs correspondant à *répétabilité*, *reproductibilité* et *sensibilité* : *répétabilité* 54 fois, *reproductibilité* 51 fois, *sensibilité* 10 fois / *répétable* et *reproductible* 0 fois, *sensible* 3 fois. Or ce sont bien des concepts insaturés qui sont exprimés dans les deux cas : de même que la lecture est la lecture de quelque chose, la sensibilité est, pour quelque chose (un instrument de mesure), la propriété d'être sensible (à des variables).

3.4. Les classes de concepts spécialisés

La notion de classe d'objets, utilisée au laboratoire LDI (« Lexiques, Dictionnaires, Informatique ») de Villetaneuse, se heurte à diverses objections⁷:

⁶ Dans la version espagnole de R7, voici les proportions entre noms d'opérations vitivinicoles et verbes correspondants: acidificación 14 / acidificar 0 ; almacenamiento 3 / almacenar 0 ; arranque 64 / arrancar 15 ; centrifugación 3 / centrifugar 0 ; comercialización 22 / comercializar 5 ; concentración 10 / concentrar 24 ; cosecha 22 / cosechar 5 ; desacidificación 12 / desacidificar 0 ; destilación 21 / destilar 0 ; elaboración 15 / elaborar 14 ; etiquetado (N) 22 / etiquetar 1 ; fermentación 21 / fermentar 18 ; filtración 3 / filtrar 0 ; injerto 2 / injertar 3 ; mezcla 9 / mezclar 3 ; plantación 43 / plantar 49 ; prensado (N) 4 / prensar 1 ; replantación 17 / replantar 1 ; sobreprensado (N) 3 / sobreprensar 0 ; sobreinjerto 1 / sobreinjertar 1 ; vinificación 34 / vinificar 1 (Lerat, 2008).

⁷ Voir les doutes a priori de M. Gross (1981 : 49) et d'A. Guillet (1986 : 100), ainsi que l'argument cognitiviste de R. Temmerman selon lequel toute catégorisation est « le résultat d'une intersection entre le langage et l'esprit » (1997 : 55, n. 3).

elle correspond à l'usage dans la langue générale, où « presque tous les termes sont polysémiques » (Gross et Mathieu-Colas, 2001 : 69), elle est conçue *in vitro*, elle est intuitive. En terminologie, au contraire, il importe d'avoir en vue une application déterminée, car les points de vue varient avec les professions et les tâches, donc il faut construire les classes spécialisées dont on a effectivement besoin. En outre, une terminologie à fondement textuel est nécessairement *in situ*, donc lexicale lors de la collecte des « candidats termes », ce qui la fragilise conceptuellement, notamment du fait des anaphores. Enfin, c'est le corpus lui-même qui guide l'intuition, ainsi que la connaissance du domaine, donc le risque de confusion entre compétence culturelle et compétence linguistique reste une menace. D'où l'importance de la normalisation en terminologie. Au reste, les classes de concepts spécialisées sont des constructions empiriques, donc révisables et construites de façon incrémentale.

Ex. 10: en exploitant les exemples ci-dessus, on a déjà < « instrument de mesure » : « aréomètre », « balance hydrostatique », « densimètre », « pycnomètre », réfractomètre »>

4. Méthode d'élaboration d'un réseau conceptuel

Peu importe le point de départ : partir des opérations, concepts insaturés qui conduisent aux objets concernés (Lerat, 2007 et Lerat, 2008), ou partir des instruments, qui conduisent à ce qu'on fait avec, c'est-à-dire des opérations, qui à leur tour mettent en jeu des objets appropriés. Ce qui est capital, en revanche, c'est de travailler sur un corpus spécialisé dont on respecte l'esprit parce qu'on est familiarisé avec sa culture. Et comme aucun corpus textuel n'explicite toutes les connaissances que sa lecture présuppose, ce qui le rendrait illisible, il faut aussi utiliser sélectivement des référentiels externes porteurs de définitions fiables et de représentations iconiques parlantes. D'où la nécessité d'une veille terminologique bien documentée. On peut distinguer trois types de veille complémentaires.

4.1. La veille lexicale

Il ne s'agit pas ici de néologie événementielle, mais de vitalité des dénominations spécialisées émergentes en tant qu'indice de pertinence conceptuelle. Les outils utilisés dans le cas présent sont les suivants :

- les fréquences brutes de chaînes de caractères sous Google

Ex. 11 : *réfractomètre numérique* (389) / *réfractomètre digital* (258)

On vérifie ainsi que la forme normalisée, bon candidat pour dénommer le concept de « réfractomètre numérique », a plus d'usage que la « forme à éviter » (GDT)

pour le français et l'anglais, le *Grand dictionnaire terminologique* (GDT) de l'Office de la langue française, précisément

Ex. 12 : *sucrose*, anglicisme, est un synonyme canadien de *saccharose*

pour les langues de l'UE, *LATE* et *Eurotermbank*

Ex. 13 : pour *passer automatique d'échantillons* (R90), il existe un synonyme, *changeur automatique d'échantillons* (LATE), qui est un calque de l'anglais *automatic sample changer*

Effectivement, ce terme a un peu d'usage (30 attestations sous Google avec un *s* à échantillon, 3 avec le singulier), mais moins que l'autre (557 au pluriel, 107 au singulier). On trouve aussi *changeur d'échantillons* (96), plus fréquent que *passer d'échantillons* (29).

- pour le français, TermSciences

Ex. 14 : *passer échantillon* ressemble à un descripteur documentaire sans « mots vides » et n'est pas défini. À la vérité, les seules définitions rencontrées parfois dans cette ressource sont en anglais ; c'est le cas pour *saccharose*, grâce à un jeu d'interfaces entre TermSciences, la base INSERM, NML (National Library of Medicine) et MeSH (Medical Subject Headings).

L'absence de définitions (sauf domaines privilégiés) limite donc l'utilité de cette ressource pour la terminologie proprement dite. En revanche, il est compréhensible que le documentaliste de l'INIST ait traduit par un calque *sample changer*, qui est dans le résumé en anglais du texte référencé; aussi bien, si l'on veut tester la vitalité de cette expression en français, on trouve 66 attestations sous Google, ce qui linguistiquement est un témoignage de plus du « style télégraphique » chez les industriels et les scientifiques, et non pas seulement chez les documentalistes.

Ex. 15 : *passer automatique* est un descripteur raccourci car le résumé de l'article belge référencé par l'INIST comporte *passer automatique d'échantillons*, mais c'est aussi un terme usuel (1400 attestations).

Ainsi, *TermSciences* est une ressource plus intéressante pour le lexicologue que pour le terminologue.

- pour le français encore, la *Base de terminologie* du CILF

Certes, contrairement à *GDT*, mais avec beaucoup moins de moyens, le CILF ne fournit pas une information terminologique au-delà de ses dictionnaires propres, mais il a des entrées nombreuses et une définition à chaque fois.

4.2. La veille normative

Le *Recueil* lui-même est à la fois scientifique, technique et normatif, ce qui favorise le travail terminologique. Il gagne toutefois à être complété, notamment pour les définitions, par des sources complémentaires : les règlements vitivinicoles de l'UE, les normes internationales (notamment en matière de métrologie et d'électrotechnique) et nationales (AFNOR), citées de façon aléatoire par *LATE*.

Ex. 16 : *résonateur* : « Appareil ou système susceptible d'entrer en oscillation par résonance avec un autre oscillateur » (CEI 05-45-050)

Ex. 17 : *oscillateur* : « Dispositif produisant un courant alternatif dont la fréquence est déterminée par les caractéristiques propres du dispositif » (NF C 01-151)

Ainsi, il est confirmé que les concepts de « dispositif » et de « système » comprennent, en matière d'instruments de mesure, le concept d'« instrument », qui se dit *instrument* ou *appareil*, indifféremment.

4.3. La veille industrielle

La veille industrielle est une notion complexe, qui inclut souvent la veille concurrentielle. Elle n'est pas sans rapport avec les normes, car la conformité à celles-ci est un argument de vente très important, notamment chez les industriels des pays anglophones et germanophones, ainsi qu'en Chine et en Italie, pour ce qui concerne les instruments de mesure utilisés en œnologie. Les définitions courantes disent plus ou moins ceci :

Ex. 18 : *veille industrielle* : « surveiller l'environnement technologique, commercial, réglementaire etc. pour en anticiper les évolutions »⁸

Les brevets ont donc une importance particulière, ainsi que les catalogues, qui contiennent souvent de bonnes définitions.

⁸ www.pacac.cci.fr/arist/veille.html

Ex. 19 : « Les densimètres numériques mesurent la densité, la masse volumique ainsi que d'autres valeurs associées (% d'alcool, degrés Brix, degrés API etc.) avec une très haute précision, en un temps très court » (site de Mettler Toledo, avec son autorisation)

5. Esquisse d'un réseau conceptuel

Un réseau conceptuel n'est pas une base de connaissances, mais une base de données terminologiques susceptible de liens hypertextuels avec une base de données textuelles et une base de données iconiques.

La nécessité de renvoyer de concepts globaux lexicalisés à des définitions consensuelles, et donc de dénominations normatives à des jeux complexes de propriétés, résulte de la nature des choses : un *oscillateur à quartz à enceinte à température régulée* (en anglais *oven controlled crystal oscillator*) n'a pas la simplicité sémantique d'une chaise ou d'un fauteuil : c'est un « oscillateur piloté par un résonateur à quartz, dans lequel le résonateur au moins est à température régulée » (CEI 561-04-05). En outre, elle est conforme à la façon classique de faire de la terminologie : il s'agit de considérer les concepts comme des ensembles de « caractères » des choses, et non pas comme des traits sémantiques distinctifs au sein d'une langue et d'une culture.

Un réseau conceptuel tel qu'il est conçu dans ce travail a pour fonction de traiter systématiquement les liens (interlinguistiques) entre les concepts (insaturés et saturés) et les liens (intra-linguistiques) entre les concepts et les termes d'une langue. Il est de bonne méthode de partir des concepts insaturés (opérateurs) car ils conduisent aux concepts saturés (leurs arguments appropriés).

Voici quelques concepts majeurs tirés du *Recueil*. Les concordances sont exploitées en tant que « collocations conceptuelles », donc interprétées par le terminologue, et non pas comme simples cooccurrences.

Les premiers champs constituent les conditions de validité du réseau : dénominations correspondantes (DN), générique (G), spécifique (SP), « comprenant » (CT), « compris dans » (CS), exemple(s) (E) (voir Calberg-Challot *et al.*, 2007 : 133-134), source(s) (SO), définition(s) (DF).

Ce qui suit la parenthèse ouvrante réunit par classes conceptuelles les concepts associés présents dans le corpus, et non pas les formulations en tant que telles.

Ex. 20 : g/L -> « gramme par litre »

Les instruments de mesure se prêtent bien à des arborescences, G et SP autorisant l'héritage des propriétés.

Ex. 21 : « densimètre électronique »/DN : fr. *densimètre électronique*/G : « densimètre » /SP : « densimètre électronique à résonateur de flexion »/CT : « afficheur numérique », « calculateur électronique », « cellule de mesure », « enceinte thermostatée », « passeur automatique d'échantillons », « thermostat »/E : DMA 35N (Autriche), MD-300S (Canada), DENSIMAX (France), DE51 (USA)/SO : *Recueil*, R90, R00, R04, R05, LATE/DF : « Appareil de mesure de la masse volumique ou de la densité d'un liquide » (LATE) relié à un ordinateur/

(< « opération de mesure » : « étalonnage », « lecture », « mesure »> ; < « denrée alimentaire » : « vin »> ; < « méthode d'analyse » : « densimétrie »> ; < « substance chimique » : « distillat », « fluide », « solution chimique »> ; < « grandeur physico-chimique » : « masse volumique », « répétabilité », « reproductibilité », « taux alcoométrique volumique », « température »> ; < « valeur de grandeur » : « degré Celsius », « gramme par millilitre »>)

Nota bene : la relation COMPRENANT prend en compte des classes cumulatives en extension ; ainsi, il n'est pas nécessaire qu'un densimètre électronique comporte un passeur automatique d'échantillon(s).

La pertinence des collocations conceptuelles exploitant le modèle opérateur / argument(s) de Harris est manifeste dans l'exemple ci-dessous : « lire » un < « instrument de mesure »> ou lire une < « valeur de grandeur »> est une opération technique, réalisable aussi bien par un dispositif que par un humain, et dans des métiers variés (dont ceux de l'œnologie ne sont que des cas particuliers).

Ex. 22 : « lecture »/DN : fr. *lecture, lire*/G : « opération de mesure »/SP : « lecture spectrophotométrique »/SO : *Recueil*, R90, R00, R04/DF : « Décodage des valeurs fournies par un appareil de mesure » (R90)/

(< « instrument de mesure » : « balance hydrostatique », « densimètre », « ionomètre », « manomètre », « spectrophotomètre », « thermomètre », « tube de mesure »> ; < « denrée alimentaire » : « moût de raisin », « vin »> ; « lieu » : « échelle graduée », « éprouvette », « récipient », « thermomètre », « tige d'aréomètre »> ; < « substance chimique » : « corps liquide », « saccharose », « solution chimique »> ; < « grandeur physico-chimique » : « absorbance », « courbe d'étalonnage », « masse volumique apparente », « pH », « potentiel »,

« pourcentage », « pression », « surpression », « signal », « température », « teneur », « tension », « titre alcoométrique apparent », « volume »> ; < « valeur de grandeur » : « degré Celsius », « gramme par litre », « millivolt », « pascal »>)/

Le fait qu'il puisse exister plusieurs définitions en fonction des communautés de travail et de leurs intérêts ne saurait créer de confusion si l'on explicite les liens entre définitions et sources, en particulier. Par exemple, le concept de « saccharose » a une définition chimique et une définition œnologique à la fois, bien qu'il s'agisse du même objet, comme on peut en juger ci-dessous :

Ex. 23 : « saccharose »/DN : fr. (*le*) *saccharose* (France), (*le*) *sucrose* (Québec)/G : « sucre »/ SO : *Recueil*, R90, R04, R07, LATE/DF : « Sucre, tiré de la betterave ou de la canne à sucre, utilisé dans certaines régions pour augmenter la teneur naturelle en sucre des moûts et, par conséquent, le degré alcoolique des vins. » (LATE) ; Glucide (C₁₂H₂₂O₁₁) qui par hydrolyse se décompose en molécules de glucose et de fructose (en chimie)/

(< « opération de mesure » : « détermination », « dosage », « expression », « lecture », « recherche »> ; < « instrument de mesure » : « réfractomètre »> ; < « denrée alimentaire » : « moût de raisin », « vin »> ; < « méthode de mesure » : « chromatographie », « pH-métrie », « RMN »> ; < « substance chimique » : « corps liquide », « eau déminéralisée », « solution chimique »> ; < « grandeur physico-chimique » : « concentration », « densité », « masse volumique », « température », « titre massique »> ; < « valeur de grandeur » : « % », « degré Celsius », « gramme par kilogramme », « gramme par litre »>)/

Nota bene : le réseau conceptuel qui peut être esquissé à partir de ces trois concepts passe par « lecture » : le densimètre électronique n'est pas fait pour lire une teneur en saccharose, mais l'opération de lecture est commune à tous les appareils, y compris aux réfractomètres, qui, eux, sont utilisés pour mesurer la concentration de saccharose.

6. Les connaissances terminologiques

Ce que la simple analyse automatique ou semi-automatique de la surface des textes ne permet pas de repérer, l'exploitation des collocations conceptuelles le repère et l'éclaire. Il est possible d'exploiter le réseau conceptuel en langues naturelles de deux façons : soit comme donnée brute pour des systèmes formels, soit comme donnée terminologique utilisable en rédaction technique, en traduction ou en terminographie. Dans les deux cas, le réseau a un amont (l'étude

d'un corpus) et un aval : l'un des deux types d'exploitation, précédé ou non d'une validation par ce non corpus chaotique mais immense et d'accès immédiat qu'est Internet.

Prenons l'exemple de *concentration*, mot polysémique y compris en langue spécialisée, et jusque dans le sous-domaine de l'œnologie. Les collocations conceptuelles du corpus conduisent à distinguer trois concepts distincts, donc trois termes au moins, correspondant à trois dimensions : la composition chimique et la fabrication, qui sont en relation avec des instruments différents (pour mesurer ou pour vinifier), ainsi que la dégustation, qui a ses professionnels (les sommeliers), mais qui n'est pas quantifiable (dans l'état actuel de la technologie) parce qu'il s'agit d'une analyse sensorielle (gustative).

Ex. 25 : « concentration »/DN : fr. *concentration*, (*substance*) *concentré(e)*, *teneur*/G : « grandeur physico-chimique »/SO : R4, R90, LATE/DF : « Quantité d'une substance (...) contenue dans l'unité de volume (...) expression numérique de la teneur » (LATE)/(« denrée alimentaire » : « moût de vin », vin »> ; < « substance chimique » : « acide sulfurique », « analyte », « clarifiant », « éthanol », « méthanol », « sucre », « sulfite »> ; < « valeur de grandeur physico-chimique » : « degré Brix », « gramme par litre », « milligramme par kilogramme »>)/

Conceptuellement, il s'agit d'un état (opposable à la dilution) ou d'une propriété (par exemple la « teneur Brix ») d'une « substance chimique » mesurable en pourcentage dans un produit.

Ex. 26 : « concentration »/DN : fr. (*lors de la*) *concentration*, *concentrer*/G : « pratique de fabrication »/CT : « déshydratation »/SO : R7/DF : Pratique de fabrication consistant à réduire le pourcentage d'eau dans une denrée alimentaire (d'après R7)/(« denrée alimentaire » : « jus de raisin », « moût de raisin », « vin »> ; < « traitement physico-chimique » : « congélation », « évaporation », « osmose inverse »>)/

Conceptuellement, il s'agit cette fois d'une opération applicable à d'autres denrées alimentaires liquides (soupes, jus de fruits etc.), dont le résultat est un « concentré ».

Ex. 27 : « concentration »/DN : fr. *concentration*, (*vin*) *concentré*/G : « propriété organoleptique »/SO : R4, pro/DF : Teneur élevée en extrait sec (selon l'âge de la vigne, le rendement, le millésime) (d'après R4 et sites de

professionnels)/(« denrée alimentaire » : « vin »> ; < « substance chimique » : « arôme », « matière sèche », « tanin »>)/

Conceptuellement, il s'agit d'une propriété (ou qualité) : un faisceau de sensations gustatives propres à l'œnologie.

Il est possible de vérifier, infirmer ou compléter les collocations conceptuelles tirées du corpus normatif en examinant sous Google les chaînes de caractères *concentration du vin*, *vin concentré* et *concentrer le vin*. Voici les résultats :

- pour la concentration chimique, on peut compléter la liste des substances chimiques susceptibles de se trouver dans le vin, avec *acide lactique*, *acide tartrique*, *ammonium C*, *anhydride sulfureux*, *arsenic*, *ion Fer III* et *polyphénol*, et aussi ajouter, dans le cas de la concentration en sucre, la collocation conceptuelle < « instrument de mesure » : « réfractomètre »/

- pour la concentration par déshydratation, on rencontre un exemple (« cryoconcentration ») et un instrument (« machine à osmoser »)

- pour la concentration sensible aux papilles, rien de plus.

Ainsi, la dialectique du terme et du concept est productive en amont (fouille dans le corpus) et en aval (navigation) de l'analyse (formelle ou non) des définitions partagées dans une communauté (nationale ou internationale) de travail.

Le bénéfice de cette confrontation avec un « non corpus tout venant » tel que Google est également lexical. Contrairement à une approche en termes de « syntagme terminologique », on voit bien que l'unicité de la chaîne de caractères *concentration du vin* masque les différences formelles (transformations verbale et adjectivale possibles ou non) et sémantiques (trois concepts distincts). Sur le plan théorique, il apparaît que l'apport de Harris est plus important, pour la terminologie, par l'utilisation du couple opérateur / argument que par celle de l'analyse distributionnelle.

7. Conclusion

Un mauvais reproche que l'on pourrait faire à cette conception du réseau conceptuel serait celui de circularité. Certes, on ne part pas de primitives, qui seraient illusoires dans le cas d'une technologie, mais de « caractères » observables, propriétés des choses perçues ou pensées. La démarche consiste à aller de concept insaturé à concept saturé, et réciproquement. Ce qui serait un

défaut dans une lexicographie à l'ancienne est une fonctionnalité productive : non pas une circularité, mais un mode de circulation, et de circulation en tous sens, une navigation interne, gagée sur des définitions..

A-t-on vraiment esquissé ici un échantillon de terminologie validable, qui soit « consensuelle, cohérente, précise » (Roche, 2007 : 7) ? Si tel est le cas, le relais peut être pris par l'ingénierie des connaissances, pour aller jusqu'à du « partageable, réutilisable et calculable » (même page). Dans le cas contraire, on aura au moins réalisé un type de travail terminologique utile : l'identification de concepts spécialisés et des termes correspondants.

Bibliographie

- Benvéniste, E., *Problèmes de linguistique générale*, Paris, Gallimard, 1966
- Calberg-Challot, M., Candel, D., Roche, C., « De la variation des usages au consensus terminologique: vers un dictionnaire de l'ingénierie nucléaire » in *Terminologie et Ontologie*, 2007, p. 119-141
- Conseil international de la langue française, *Base de terminologie*, www.cilf.org
- Eurotermbank : *Banque de données terminologiques multilingue*, www.eurotermbank.com
- Gross, G. et Mathieu-Colas, M., « Description de la langue de la médecine », *Meta*, vol. 46, n° 1, 2001, p. 68-81, www.erudit.org/revue/meta/2001/v.46/n1
- Gross, M., « Formes syntaxiques et prédicats sémantiques », *Langages*, n° 63, 1981, p. 7-52
- Guillet, A., « Représentation des distributions dans un lexique-grammaire », *Langue française*, n° 69, 1986, p. 85-107
- Harris, Z.S. : *Structures mathématiques du langage (1968)*, trad. C. Fuchs, Paris, Dunod, 1971
- Harris, Z.S. : *Notes du cours de syntaxe*, Paris, Senil, 1976
- Harris, Z.S. : *Langue et information (1988)*, trad. A. H. Ibrahim, Paris, CRL, 2007
- LATE, *InterActive Terminology for Europe*, <http://iate.europa.eu>
- Lerat, P. : *Les langues spécialisées*, Paris, PUF, 1995
- Lerat, P. : « Les nominalisations en -tion dans un texte techno-administratif » in *Terminologie et Ontologie*, 2007, p. 79-92

Lerat, P. : « La terminologie communautaire des opérations vitivinicoles », à paraître dans les actes du II Congreso internacional sobre la lengua de la vid y el vino y su traducción, Soria, 2008

Office de la langue française, *Grand dictionnaire terminologique*, www.granddictionnaire.com

Rey, A. : *La terminologie : termes et notions*, Paris, PUF, 1979

Roche, C., « Le terme et le concept : fondements d'une ontoterminologie » in *Terminologie et Ontologie*, 2007, p. 1-22

Temmerman, R., « Questioning the univocity ideal. The difference between sociocognitive Terminology and traditional Terminology », *Hermes. Journal of Linguistics*, n° 18, 1997, p. 51-91, <http://hermes2.asb.dk/archive>

Terminologie et Ontologie : théories et applications, Annecy, Institut Porphyre, 2007, www.porphyre.org/toth

TermSciences, *portail terminologique multidisciplinaire*, CNRS, www.termosciences.fr

Wüster, E. : *Dictionnaire multilingue de la machine-outil. Volume de base Anglais - Français*, Londres, Technical Press, 1968

Wüster, E. : *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie (1979)*, Copenhague, Handelshøjskolen, 1985

Corpus

Recueil des méthodes internationales d'analyse des vins et des moûts, Paris, Office International de la Vigne et du Vin (OIV), vol. I, 2008

Règlement (CEE) n° 2676/90 de la Commission du 17 septembre 1990 déterminant les méthodes d'analyse communautaires applicables dans le domaine du vin ⁹ (version consolidée, 1999) (R90)

⁹ Le site de toute la législation communautaire est <http://eur-lex.europa.eu>. Avertissement : « Seule fait foi la version de la législation européenne telle que publiée dans les éditions papier du Journal officiel de l'Union européenne ».

Règlement (CE) n° 2870/2000 de la Commission du 19 décembre 2000 établissant des méthodes d'analyse communautaires de référence applicables dans le secteur des boissons spiritueuses (R00)

Règlement (CE) n° 440/2003 de la Commission du 10 mars 2003 modifiant le règlement (CEE) n°2676/90 (R03)

Règlement (CE) n° 128/2004 de la Commission du 23 janvier 2004 modifiant le règlement (CEE) n° 2676/90 (R04)

Règlement (CE) n° 355/2005 de la Commission du 28 février 2005 modifiant le règlement (CEE) n° 2676/90 (R05)

Proposition de Règlement portant organisation commune du marché vitivinicole et modifiant certains règlements, Bruxelles, 4-7-2007 (R07)

Interrogations sur l'évolution du français de la gestion

Odile Challe

Maître de conférences HDR

Directrice adjointe du CICLaS

Université Paris-Dauphine

Membre associé du DILTEC Paris Sorbonne Nouvelle

Odile.Challe@dauphine.fr

Résumé :

Le langage joue un rôle crucial dans le contexte de l'entreprise actuelle. La communauté des gestionnaires produit des discours en langue française dans lesquels fleurissent selon les spécialités des termes très techniques qui en chassent d'autres. On peut s'interroger sur ces apparitions : s'agit-il de nouveaux concepts ou d'influence par traduction ? À partir d'exemples professionnels pris dans la gestion, la comptabilité voire le droit en entreprise, on se demande si les nouveaux termes ne reprennent pas un concept ancré dans la culture historique de la spécialité. La notion du concept flou sera abordée par l'étude de quelques euphémismes.

Plan

L'euphémisme dans la langue dans l'entreprise.

Terminologie spécialisée en gestion

Exemples d'euphémismes en droit de l'entreprise

Conclusion : les besoins

1. L'euphémisme dans la langue de l'entreprise

Avec les mutations du monde économique, le langage se démultiplie en discours suite à la prolifération des supports. L'économie détourne des mots connus et en apporte d'autres qui se diluent. Le terme *entreprise* par sa polysémie illustre la difficulté à décrire les réalités que vivent ceux qui y travaillent. Substantivation du participe passé féminin du verbe qui signifiait au départ attaquer¹, il désigne une action parfois hostile, qui prend un sens marchand par confusion avec le verbe "emprendre" pour mettre en œuvre. Avant d'être une action de commerce, elle était une opération militaire. Aujourd'hui, il renvoie à une organisation de production de biens ou de services à caractère commercial, d'où les syntagmes dérivés comme chef d'entreprise, comité d'entreprise mais c'est aussi un lieu de vie.

Mon intervention s'appuie sur les travaux que je mène à l'université de Paris-Dauphine depuis 1985 sur l'étude des discours en sciences des organisations : gestion, économie, droit des affaires internationales, management, etc. En particulier, je ferai référence aux résultats d'un colloque que j'avais organisé avec le doyen Joël Monéger, réunissant autour de la problématique de la précision dans la langue de l'entreprise, des collègues spécialistes de droit économique et de comptabilité (Université Paris-Dauphine, 22 octobre 2004).

Parallèlement, mon activité d'enseignement à des chercheurs ou salariés étrangers m'a fait prendre conscience que le monde du travail s'ancre dans une culture et donc dans une langue, expérience confirmée dont je retire que les non-francophones spécialistes se heurtent particulièrement à deux phénomènes : la métaphore et l'euphémisme. Mon but est ici de poursuivre mon interrogation sur la précision qui me semble faire défaut dans les discours en entreprise pourtant en langue de spécialité, au regard de l'objet de ce colloque, le classement des connaissances métiers, mais en l'élargissant d'un point de vue universitaire aux classements théoriques des sciences des organisations.

Devenue essentielle dans un contexte d'internationalisation du monde du travail, une langue de spécialité ne se traduit pas aisément car elle est souvent loin d'être précise. C'est par rapport à cette non-précision que la langue de l'entreprise

¹ Source : *Le dictionnaire historique de la langue française*, Le Robert, éd. 1995, p. 700.

est abordée, en me limitant cette fois-ci² à la forme euphémique de quelques termes.

Les spécialistes en gestion interrogés ont fourni de précieuses informations sur les concepts de leurs disciplines. On pourrait croire qu'en comptabilité, contrôle de gestion, finances ou droit économique, les termes sont extrêmement précis, il n'en est rien. Dès que sont prises en compte les conditions concrètes d'utilisation de la langue dans une communauté donnée, la vision d'un instrument linguistique abstrait coupé de sa logique sociale de fonctionnement s'étirole.

Nombreux sont les travaux sur les discours de vulgarisation (Moirand 2007) qui véhiculent des termes sinon précis du moins définis dans la sphère des spécialistes qui les a forgés. Les journalistes économiques ont le devoir de traduire en langage accessible par leurs lecteurs-auditeurs-spectateurs les termes qui font écran et non d'inventer un autre jargon³. Ils ne sont que des relais et des diffuseurs parfois excellents pédagogues sans doute au prix d'euphémismes. Mais en amont, les discours premiers émanent soit des acteurs de l'entreprise, soit des concepteurs de théories économiques, juridiques et managériales. Or ces deux mondes échangent entre eux des concepts, les uns du terrain comme celui de *flexibilité* au travail qui a promu la récente *flexsécurité* issue du modèle danois, les autres dans leurs publications, conférences, expertises et formations, voire interventions médiatiques. Les termes évoluent puisqu'ils désignent des réalités mouvantes.

Dans ce cadre où plusieurs discours s'entrecroisent, métiers liés à l'entreprise et théories en sciences des organisations, les conditions de performativité⁴ deviennent primordiales et l'acheminement au grand public renforcé par l'usage d'internet résulte souvent d'un mouvement euphémique, volontaire ou non. Rappelons brièvement que l'euphémisme est un trope, autrement dit une figure dans laquelle on emploie les mots⁵ avec un sens différent du sens habituel, qui

² Par ailleurs, j'ai commencé à traiter de la métaphore spécialisée en gestion (sous presse, Ecole Polytechnique).

³ Philippe Simmonnot, « *Le Monde* » et le pouvoir, Paris, les presses d'aujourd'hui, 1977, p. 118.

⁴ Le fait de prononcer certains verbes exprime leur valeur d'action, ce qui est le cas de "la séance est levée".

⁵ Discuté en linguistique, le **mot** reste encore plus utilisé en entreprise que **terme**.

peut aller jusqu'à « exprimer le contraire de ce que l'on veut dire », c'est alors une antiphrase, comme le signale Georges Mounin⁶. Dumarsais rappelait qu'en grec l'euphémisme est le discours de bon augure, et Sourieux et Lerat qu'il est à la fois ornement de discours et technique de communication⁷. Chaque fois qu'un mot est jugé inconvenant il est remplacé par un mot supportable, grâce aux vertus de l'adoucissement, selon Jean Carbonnier⁸. Le cas du droit appliqué à l'entreprise est à cet égard révélateur.

Pour François Terré, actuellement président l'Académie des sciences morales et politiques et fortement impliqué dans la modernisation de la terminologie en droit, c'est par une pensée consciente que le spécialiste recourt tant aux euphémismes qu'aux litotes ou aux hyperboles comme outils qui servent à acclimater le droit. Pour lui, « l'absence fréquente d'univocité des discours inhérente à la communication est particulièrement problématique en matière juridique⁹ », surtout quand elle concerne l'articulation avec le monde économique car, estime-t-il, le point de vue sur les faits juridiques donne naissance à des *valeurs*, c'est-à-dire des objets chargés de signification. Le performatif se combine alors avec le descriptif et le narratif. Il considère que le droit économique diffère des courants dominants en droit classique par ses concepts, ses catégories, ses qualifications. Ce droit anciennement droit des affaires, correspond au droit commercial venant de la langue du négoce fortement liée à la comptabilité. De nos jours, les présidents pas seulement de banque alourdissent leurs prises de parole du poids des performances financières, ce qui faisait dire à Jacques Attali et Marc Guillaume que les ouvrages économiques sont tristes et figés, loin d'être simples et gais : « leur présentation ennuyeuse veut signifier qu'on y parle sérieusement, qu'on y énonce une vérité scientifique solennelle et ennemi de la facilité »¹⁰.

⁶ Georges Mounin (dir.), *Dictionnaire de la linguistique*, PUF, 1974.

⁷ Jean-Louis Sourieux et Pierre Lerat, "L'euphémisme dans la législation récente", Dalloz 1983 chron. 221 s.

⁸ Jean Carbonnier, *Essais sur les lois*, Rép. Defrénois, 1995, p 284.

⁹ François Terré, "Droit du langage". In Odile Challe (dir.), *Langue français spécialisée en droit*, Economica, 2007, p. 1-6.

¹⁰ Jacques Attali et Marc Guillaume, *L'anti-économique*, PUF, 1974, p. 5.

Le langage spécialisé tend à devenir une institution ; à chaque besoin de réforme, il entre dans le mouvement vers l'euphémisation. S'agit-il de néologismes, de rajeunissement de la langue ?

2. Terminologie spécialisée en gestion

Mon objectif ici ne reprend pas les travaux de la commission du ministère des Finances. Ayant participé à la création en 1984 de l'équipe Lecticiel, à l'école normale supérieure de Saint-Cloud au sein du Credif (Centre de Recherche et d'études pour la diffusion du français), je me réfère à notre recherche d'alors visant à appliquer à l'informatique la démarche développée dans une méthode de lecture de textes spécialisés en sciences économiques et sociales¹¹. Sur le plan de la francophonie, cette méthode a marqué de nombreuses générations, de l'Amérique latine à la Syrie, et a largement contribué à la diffusion à l'étranger de la langue française spécialisée. Elle consistait à regrouper à l'intérieur d'un même texte, les mots de la spécialité dans des ensembles conceptuels fortement liés à la structuration du domaine de références. A partir de ces classements, des tableaux pédagogiques pouvaient s'élaborer sur la base de textes en sciences économiques et sociales : le champ des agents économiques avec ses sous-ensembles (entreprises, particuliers...), le champ des objets économiques qui dans un texte se retrouvent sous l'hyperonyme monnaie (moyens de paiement, crédits, liquidités, réserves, dépôts...), le champ des opérations économiques (*régler* la circulation monétaire, *régulation* du marché monétaire, *contrôler* les moyens de paiement, moyens de paiement que *créent* les banques, qui *accordent* des crédits...). Un autre groupe d'éléments sert à caractériser les objets ou les opérations, souvent des quantificateurs (*volume* ou *quantité* des moyens de paiement, *montant* de la monnaie, de 11% à 12%, du *total* des dépôts). Le type d'opération en jeu peut être la relation entre les *prix* et les *profits*. Cette approche à l'origine didactique correspond à la conception du Professeur Alfandari qui définit le droit économique comme l'étude « des cadres juridiques de l'économie, des agents économiques, des objets économiques et des activités économiques »¹².

a. L'euphémisme en comptabilité générale

Prenons à présent les résultats obtenus auprès de collègues gestionnaires à l'université de Paris-Dauphine. Tout d'abord, en comptabilité générale, Jacques

¹¹ Denis Lehmann et al. *Lire en français les sciences économiques et sociales*, 1979, Didier.

¹² Elie Alfandari, *Droit des affaires*, Litec, 1993, p. 3.

Richard constate que depuis une trentaine d'années, la comptabilité capitaliste a entrepris de renouveler sa terminologie. Il affirme que c'est sous l'influence principale des auteurs américains que ce renouvellement s'effectue et au moyen d'euphémismes qui visent d'une part à éviter des mots trop dévalorisés voire brutaux, d'autre part à trouver des mots qui soient de bon augure et qui se propagent efficacement. Il en veut pour preuve la disparition de plusieurs termes majeurs dans sa discipline remplacés par des euphèmes (ou *bonnes expressions*) : parmi les acteurs, le terme de capitalistes-proprétaires, parmi les objectifs, l'expression « recherche de profit », et parmi les modes d'action, on en parle plus de domination.

Dans le premier cas, le terme de capitaliste est plus dévoyé que le terme de propriétaire. Le capitaliste-propriétaire désignait l'apporteur de capital. Il désigne actuellement l'actionnaire. Il aurait pu être remplacé par l'expression « apporteur de capital » mais c'est le terme d'investisseur qui l'a emporté. A regarder de près, le concept d'investissement est beaucoup plus large car il peut provenir d'actionnaires, mais également d'obligataires ou de banquiers. L'harmonisation mondiale des comptabilités financières¹³ en cours en 2005 restreint les investisseurs aux seuls fournisseurs de capitaux à risque (*providers of risk capital*) distincts des prêteurs (*lenders*), rang auquel sont relégués les créanciers. L'IASB va donc à l'encontre d'une tradition comptable européenne continentale qui distinguait le capital des propriétaires et le capital des prêteurs.

Ainsi, parmi les personnes qui aident l'entreprise en lui fournissant des ressources, les Allemands distinguent ceux qui apportent un capital propre (*Eigen Kapital*) et ceux qui apportent un capital étranger (*Fremd Kapital*) là où le plan comptable français de 1982 distinguait les capitaux propres et le passif externe. On peut donc parler d'euphémisme dans la mesure où le terme d'investisseur réduit l'apport de capital à l'actionnaire, comme si celui-ci jouait le rôle essentiel dans le financement de l'entreprise et oblitère les autres sources d'apport. Le concept d'investisseur serait apparu avec l'élaboration de la théorie financière moderne des années 1960-1990. Déjà en 1927, son père spirituel, Irving Fisher¹⁴, opposait les *stockholders* et les *crédores*.

¹³ Menée par l'*International Accounting Standard Committee* ou *Standard Board*, sis à Londres : IASC/IASB.

Un autre terme entre comptabilité et finances a disparu, celui de « recherche de profit » au profit de « création de valeur ». il aurait pu être remplacé par celui des années 50-60, de « résultat résiduel ». On peut voir dans le choix de mot quelque peu ambigu de « valeur » un rapprochement voulu avec la création de valeur sonnante et trébuchante pour l'actionnaire dans les dividendes qu'il espère toucher. Il semble que ce soit à la fin du XX^e siècle que les manuels de finance aient banni le terme de profit : « la maximisation du profit » est devenue un objectif insensé pour un dirigeant d'entreprise, il est plus élégant de parler de maximisation certes mais de la VAN, mis pour Valeur actuelle nette. Aujourd'hui, l'entreprise doit créer de la valeur, hier de la richesse. On peut se demander si les concepts de profit pur du XIX^e siècle à l'avènement de l'industrie et de la Bourse et les concepts de valeur actuelle nette ou plus récemment de valeur créée sont des concepts différents. Il s'agit du solde bénéficiaire qui reste d'un apport de capital une fois que l'on a déduit le coût du capital, autrement dit, du supplément de rentabilité.

La valeur créée ou valeur ajoutée économique (*economic added value = VAE* !), apparue dans les années 90, ne bénéficie qu'à l'actionnaire puisqu'elle correspond à la part de la valeur ajoutée qui revient au seul actionnaire après satisfaction de son besoin en profit normal selon le coût du capital. La création ne se confond-elle pas avec la captation ?

b. L'euphémisme en contrôle de gestion

La comptabilité de la gestion, à usage interne, est devenue le contrôle de gestion, par opposition avec la comptabilité générale ou financière (crédits clients et fournisseurs...). Ce changement d'appellation présente lui aussi un cas intéressant dans ce domaine où la terminologie usitée en France a dans les dernières années évolué. Henri Bouquin qui demandait de remettre en cause le rôle du langage financier¹⁵ pour plus de simplification, voit dans la proposition de contrôleurs de gestion de se présenter comme « consultants internes » ou encore « économistes », « conseillers internes » et même « experts en ingénierie de la performance », un souci d'atténuer la menace perçue dans le mot contrôle. Philippe d'Iribarne a montré qu'à la différence des cultures néerlandaises et nord-américaine, les Français acceptent mal l'idée de contrôle qu'ils perçoivent comme déshonorante renfermant un soupçon, un manque de confiance.

¹⁵ Henri Bouquin, *Les fondements du contrôle de gestion*, PUF, 1994, p. 92.

Le dictionnaire historique de la langue française rappelle que le contrôle était lié au poinçon des orfèvres mais omet l'histoire du comptable. Aujourd'hui encore, les fiscalistes parlent du registre des rôles, autrefois le double du rouleau de parchemin moyenâgeux qui portait les enregistrements comptables, c'est l'origine du contre-rouleur, terme repris par les Anglais, jusqu'à ce que Colbert succède à Fouquet comme contrôleur général des finances aujourd'hui ministre en 1665 sous Louis XIV. Pour le spécialiste Henri Bouquin, le contrôle est un euphémisme car ce que les Français appellent « contrôle de gestion » est désigné par les Anglo-Saxons « *management control* ». Le premier contrôle est une vérification, le deuxième *control* est avoir le contrôle sur, autrement dit maîtriser.

c. L'euphémisme en comptabilité financière

Enfin, la comptabilité générale dite financière à cause de l'expression américaine *financial accounting*, n'est pas comme la comptabilité analytique interne à l'entreprise mais participe à la régulation de ses relations avec l'environnement. D'après Bernard Colasse, membre du comité de terminologie de l'Institut des Comptables agréés du Québec, un principe de prudence apparu au XIX^e siècle, impose de l'enregistrement de moins-values potentielle mais interdit celui de plus-values. Il voit dans cette sous-estimation de la valeur des actifs et du résultat de l'entreprise une sorte d'euphémisation puisqu'il y a recherche d'atténuation de la réalité. Parmi les théories de recherche sur l'entreprise, il renvoie à la théorie la plus actuelle en finance dite de l'agence qui accorde la prédominance aux relations entre dirigeants et actionnaires. L'actionnaire détient le pouvoir de gestion qu'il délègue au l'agent, celui qui exerce le pouvoir, sur les salariés. La relation peut aller de la confiance à la méfiance, d'où la recherche dans la comptabilité de pacification entre le mandant et le mandataire, de médiation.

La comptabilité devant légitimer l'entreprise aux yeux de ses partenaires, elle produit des informations sur les flux financiers et ne fait pas l'économie d'euphémismes qui jouent un rôle quasiment social.

3. Exemples d'euphémismes en droit de l'entreprise

En étudiant l'euphémisme dans la législation récente¹⁶, Sourieux et Lerat ont dégagé de l'euphémisme une caractéristique de la « *despécification* ». C'est ainsi que ces auteurs expliquaient le phénomène d'inclusion propre à ce trope. Si la despécification est l'action de minimiser ou neutraliser les différences, il s'agit

¹⁶ J.-L. SOURIOUX et P. LERAT, *op. cit.*, p. 221.

d'utiliser, par euphémisme, une notion volontairement vague, générale, englobant, parmi autres choses, la réalité de ce que l'on veut désigner. Ce phénomène d'inclusion ne serait-il pas la résultante d'un langage juridique de droit économique imprécis ? On peut avancer avec Sourieux & Lerat que l'euphémisme engendre des concepts mous, et pas seulement en droit de la concurrence.

En droit boursier, la technique de l'offre publique d'achat que l'on désigne par l'abréviation OPA a été imaginée pour éviter les aléas d'un ramassage en Bourse. On distingue alors couramment les OPA amicales et inamicales. Le risque d'une OPA n'est que la conséquence normale d'une cotation des actions à la Bourse : « *Une OPA est presque toujours ressentie par la société cible au mieux comme un séisme au pire comme une agression* ¹⁷ ». Cette qualification est euphémique car *inamicale* voire même *agressive* puisqu'elle demeure une atténuation de la réalité, bien qu'elle soit très répandue parmi la doctrine ou la pratique.

L'euphémisme exerce une fonction adoucissante du moins atténuante curieusement en lien avec la métaphore. Ainsi, que les mots révèlent comment ; le premier exemple de manipulation révèle les différentes techniques par lesquels les cours sont manipulés comme celui dite de la bouilloire¹⁸. Cela consiste à manipuler rapidement une valeur dont le marché est étroit et sensible, à la hausse. De nombreux ordres d'achat sont passés pour persuader les spéculateurs de l'imminence d'une opération sur ce titre, ce qui permet de tromper les investisseurs vu la brusque croissance des volumes des transactions qui engendre l'augmentation artificielle du cours des titres. La seconde technique est celle du carambouillage¹⁹, exemple classique d'escroquerie combinant l'usage d'une fausse qualité et l'emploi de manœuvres frauduleuses. Elle consiste à revendre une marchandise sans l'avoir payée.

L'euphémisme par atténuation se retrouve aussi en matière de droit des entreprises en difficulté. Le "*droit des entreprises en difficulté* ²⁰" se disait autrefois le *droit des faillites*. Or, la faillite plonge ses racines dans les logiques issues du droit romain des obligations. Les termes de banqueroute et de faillite terrifiaient ceux

¹⁷ Y. GUYON, *op. cit.*, p. 645.

¹⁸ W. JEANDIDIER, *Droit pénal des affaires*, Précis Dalloz, 2000, p. 148.

¹⁹ W. JEANDIDIER, *op. cit.*, p. 12.

²⁰ J. PAILLUSSEAU, *Du droit des faillites au droit des entreprises en difficulté...*, Mélanges Houin, 1985.

qui en étaient victime. D'essence pénale, ce mot recouvre une dimension négative qui est assimilée à l'échec. Or les lois de 1807, de 1889 et de 1955 utilisent encore cette expression alors qu'elle est trop brutale et demande d'être adoucie. Les réformes de 1967 ont fait émerger un droit concernant ces entreprises en difficulté. Pour Mme Pérochon, « *il ne s'agit pas ici d'un euphémisme analogue à ceux qui ont conduit des pays sous-développés aux pays en voie de développement ou des aveugles aux non-voyants sans que rien n'ait changé, hélas, dans leur situation. Le changement de terminologie reflète un changement de perspective réel*²¹ ». Dès que la faillite est devenue un problème de société, on observa le développement de modifications sémantiques. On parle même parfois de défaillance d'entreprise avec les lois de 1985 et 1994 : « *le droit des faillites devient le droit des entreprises en difficulté, droit des procédures collectives ou d'insolvabilité, ou droit du redressement et de la liquidation judiciaires*²² ». Cet adoucissement est-il le fait exclusif du législateur français à moins que le droit européen vise les procédures d'insolvabilité. Il est à signaler que la nouvelle procédure concernant le droit du surendettement est baptisée « *rétablissement personnel* », là aussi par atténuation.

Soinne souligne « *l'impérieuse nécessité d'un changement de dénomination ; le terme même de liquidation de biens est révélateur. On liquide, c'est-à-dire qu'il faut vendre, disperser, réaliser*²³ ». Vu la connotation du mot « *liquidation* », on peut souhaiter une formule plus adoucissante : « *Le débiteur doit savoir que la procédure ouverte à son encontre n'a pas pour vocation ou pour nature de l'enterrer définitivement sur le plan patrimonial ou financier. Il s'agit bien au contraire de le rétablir dans l'intégralité de ses espoirs* ».

Les termes euphémiques dans le langage de l'entreprise et du droit lié ne sont-ils pas monnaie courante ? Quand le Code de commerce institue une « *période d'observation* », destinée à évaluer la situation économique du débiteur, on constate un effet d'atténuation de la réalité. Non seulement on observe le débiteur voire on l'épie, mais surtout, on l'assiste. S'agit-il d'une période de « *gestion intermédiaire* » selon les termes de M. Chaput ? La loi du 25 janvier 1985 qui répartit la charge et les pouvoirs de gestion entre les différents acteurs de la procédure, la « *période suspecte* » ou période courant de la cessation des paiements à l'ouverture de la procédure, ne manque-t-elle pas de précision ? Ce laps de temps

²¹ F. PEROCHON, *Entreprises en difficulté*, LGDJ, 2001, p. 4.

²² P. ROUSSEL-GALLE, *De l'évolution sémantique à l'hypocrisie des mots*, Les Petites Affiches, n°70, p. 3.

²³ B. SOINNE, Réalisme et confusion (à propos du projet de loi réformant le droit des procédures collectives), *D.* 2004, p. 1506.

est-il favorable à la triche du débiteur, à un souci de fraude de sa part ou de son cocontractant ? « *Suspect* » n'est-il pas d'une certaine façon, une manière d'adoucir la réalité et « *période douteuse, louche, équivoque ou interlope* » n'est-elle pas l'une des formules euphémiques traduisant un malaise dans cette discipline, amenant sans doute les observateurs experts à constater une certaine faillite " *du droit des faillites*²⁴ ».

4. Conclusion : concepts flous mais besoins précis

La floraison des discours récemment produits par l'actualité économique en constante évolution parfois non anticipée par les spécialistes, engendre l'apparition de mots recyclés parés de valeurs déclarées modernes. Par différents exemples puisés dans un éventail discursif allant de la comptabilité des entreprises au droit économique, nous avons voulu montrer que les nouveaux mots ne sont pas toujours plus clairs que les précédents, n'introduisent pas un concept très différent, et peuvent être la marque d'une recherche d'adoucissement à moins que l'euphémisme ne soit manipulation.

Des notions telles que profit ou actionnaire sont au cœur des débats actuels que provoque la crise des « actifs toxiques ». Avec l'appui éclairé de nos collègues universitaires, nous nous proposons de poursuivre l'étude de ces termes qui germent dans les prises de parole des dirigeants de banque ou d'entreprise et des journalistes économiques, sans oublier le discours juridique attendant et à venir. La toxicité appartient aussi à cette masse de concepts flous sur l'utilité desquels nous ne cessons de nous interroger.

Le champ ouvert depuis longtemps reste immense mais d'ores et déjà, comme nous l'avons écrit avec le Doyen Andrée Brunet en conclusion de notre article sur la précision du langage des lois (Brunet, Challe 2007) :

« Ainsi, un bon usage des normes floues, loin d'entraîner la ruine du droit, contribue à l'enrichir par cet apport d'équité qui lui est indispensable ». On peut estimer que le flou est légitime en ce qu'il rend le droit plus humain, « évitant au plus faible d'être opprimé par le plus fort ».

Opposer la flexibilité du droit à la précision du droit, c'est « reprocher au premier de rendre le droit trop incertain et à la seconde de le rendre trop rigide ». Toutefois, le flou ne saurait supplanter la précision car s'il envahit le droit, « ce

²⁴ J.-J. NEUER, *Vers un droit des faillites plus souple*, La Tribune, 2 juin 2004.

n'est pas seulement la précision qui recule jusqu'à disparaître, mais c'est l'idée même de droit qui s'effondre » dans la mesure où, comme on la vu dernièrement, trop de place risquerait d'être laissée à l'arbitraire du juge.

« À la vérité, la précision doit demeurer première sinon trop de flexibilité dissout le droit ».

Ce travail d'analyse de termes avec des spécialistes co-producteurs de ces types de discours, nous est apparu indispensable non pour porter un jugement sur l'utilisation de ces termes mais pour répondre à des besoins de formations langagière à de futurs professionnels comme à de spécialistes qui oeuvrent, c'est-à-dire agissent, décident dans un contexte international donc forcément multilingue (et multiculturel). A nos yeux, ces besoins sont très précis.

Cette présentation fait ressortir des besoins dans différentes directions que nous résumerons ainsi : Les langues de spécialités sont à étudier en lien avec la vision conceptuelle des spécialistes. La formation à ces langues s'entend avec une dimension historique et dans un souci critique. La traduction suppose une approche des cultures. Quand des secteurs se normalisent comme la comptabilité ou le droit, il importe de développer des termes dans plusieurs langues, à commencer par les langues communautaires.

Bibliographie

Alfandari, E. (1993), « Droit des affaires », Litec, p. 3.

Attali J., Guillaume M. (1974), « L'anti-économique », PUF.

Carbonnier, J. (1995), « Essais sur les lois », Rép. Defrénois, p. 284.

Challe, O. (2002), « Enseigner le français de spécialité », *Economica*.

Challe, O. (2008), « Les cultures en discours, trame de fond du français de spécialités », *Apprendre une langue de spécialité : enjeux culturels et linguistiques*, Éditions Polytechnique, pp. 79-84.

Charaudeau, P. (2008), *La médiatisation de la science dans les médias d'information*, De Boeck.

Mounin, G. (1974), « Dictionnaire de la linguistique », PUF.

Guyon, Y. (2002), « Droit des affaires », tome 1, *Droit commercial général et sociétés*, *Economica*, p. 645

Bouquin, H. (1994), « Les fondements du contrôle de gestion », PUF.

Jeandidier, W. (2000), « Droit pénal des affaires », Précis Dalloz.

Lehmann D. et al. (1979), « Lire en français les sciences économiques et sociales », Didier.

Moirand, S. (2007), « Les discours de la presse quotidienne : observer, analyser, comprendre », PUF.

Neuer, J.-J. (2004), « Vers un droit des faillites plus souple », *La Tribune du 2 juin*.

Pailhousseau, J. (1985), « Du droit des faillites au droit des entreprises en difficulté », *Mélange Houin*.

Perochon, F. (2001), « Entreprises en difficulté », LGDJ.

Philippe, S. (1977), « Le Monde et le pouvoir », Paris, *Les presses d'aujourd'hui*.

Soinne, B. (2004), « Réalisme et confusion (à propos du projet de loi réformant le droit des procédures collectives) », Dalloz.

Sourieux, J.-L., Lerat, P. (1983), « L'euphémisme dans la législation récente », *Dalloz chron.* 221 s.

Terré, F. (2007), "Droit du langage". In Odile Challe (dir.), « Langue française spécialisée en droit », *Economica*, pp. 1-6.

Comment modéliser des concepts en rapprochant un langage orienté objet et deux normes terminologiques orientées concept ?

Hendrik J. Kockaert

Lessius, Département de linguistique appliquée, Anvers, Belgique
Hendrik.Kockaert@lessius.eu

Bassey E. Antia

University of Maiduguri, Department of Languages & Linguistics
Borno State, Nigeria

Résumé :

Cette présentation se réfère à un nouveau rapport technique de l'ISO (RT ISO 24156 : *Guidelines for using UML notation in terminology work*) et vise à appliquer le langage UML à une modélisation orientée concept. Dans l'objectif d'établir un maximum de correspondances entre les modélisations UML et ISO, cette présentation démontrera les cas où une extrapolation du langage UML à la modélisation de concepts ISO est exécutable et lancera en même temps certains défis pour les cas de non-existence de compatibilité entre le langage UML et les schémas ISO.

Mots-clés : Modélisation, terminologie, UML, orienté concept

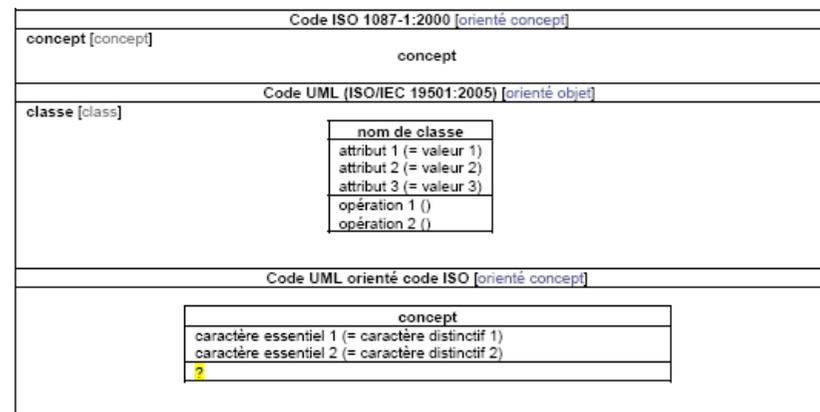
1. Introduction

Commençons par une brève description du langage UML. Son acronyme anglais UML est la forme contractée de *Unified Modeling Language* qui peut se traduire en français par *langage unifié pour modélisation*. C'est un langage de programmation informatique pour modéliser des objets. Afin de représenter et de visualiser les composants d'un système constitué d'objets, le langage UML se base sur une "sémantique précise et sur une notation graphique expressive" [Fournier 2002]. Il permet de modéliser la structure et le comportement d'un système, indépendamment de toute méthode ou de tout langage de programmation.

Nous basant sur les objectifs de cette présentation, il convient de chercher à juxtaposer la modélisation UML et notre proposition de modélisation ISO telle qu'elle figurera dans le rapport technique RT ISO 24156. Nous nous référons aux tableaux ci-dessous qui se rapprochent du Tableau figurant dans l'Annexe informative du RT ISO 24156. Dans ces tableaux, la première rangée mentionne les notations du document ISO 1087-1, alors que la seconde mentionne les notations UML telles qu'elles figurent dans le document ISO/IEC 19501. La troisième rangée visualise comment nous envisagerions de représenter le code UML en l'intégrant dans le code ISO. Les tableaux se voient escortés par des commentaires au sujet des parallélismes ISO-UML existants, ainsi que de certaines notations telles que nous les proposerions pour les cas de non-existence de compatibilité ISO-UML.

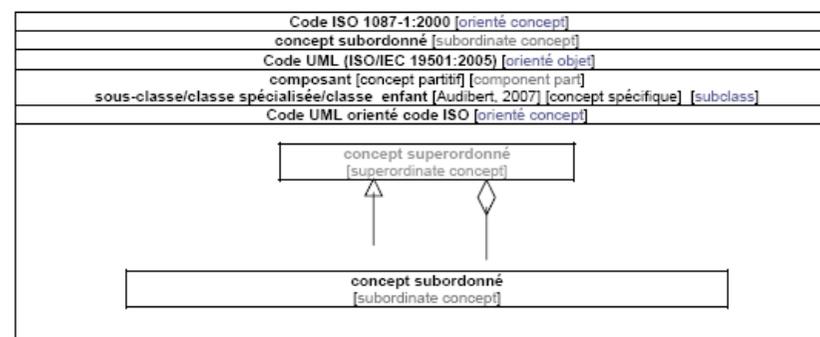
2. Concepts

Le Tableau 1 reprend pour la modélisation ISO la représentation UML pour visualiser une classe. Une classe est représentée par un rectangle, où le premier compartiment contient le nom de la classe et le deuxième les attributs et les valeurs (couples attribut-valeur). Le troisième compartiment, qui est facultatif, contient les opérations. Pour le modèle ISO, nous avons fait correspondre les attributs aux caractères essentiels et les valeurs aux caractères distinctifs (voir infra). S'il y a lieu d'ajouter des opérations, celles-ci figurent dans un nouveau compartiment. Une modélisation du concept ainsi que des caractères fait défaut en ISO 1087-1. Nous proposerions de combler cette lacune par la modélisation ci-dessous (voir Tableau 1) où *nom de classe* devient *concept*, *attributs* devient *caractères essentiels*, et *valeurs* devient *caractères distinctifs*.



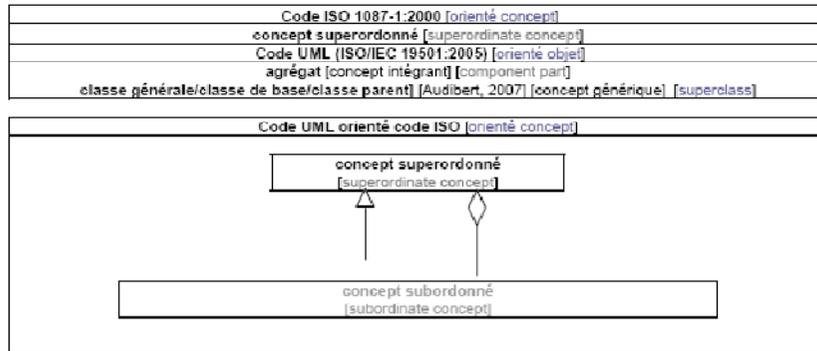
Tab. 1 : Concept

Pour le concept subordonné, il faut observer que l'ISO 1087-1 donne une définition par extension de ce concept, selon laquelle un concept subordonné est soit spécifique soit partitif. En UML, une sous-classe est l'équivalent du seul concept spécifique ISO. L'UML ne définit pas la notion d'une sous-classe comme classe spécialisée ou composant. C'est sous la mention *aggregation* (agrégation) que l'UML fait indirectement mention d'un component part (composant). Pour l'ISO modélisable, nous opterons par conséquent pour la modélisation du concept subordonné comme dans le Tableau 2.



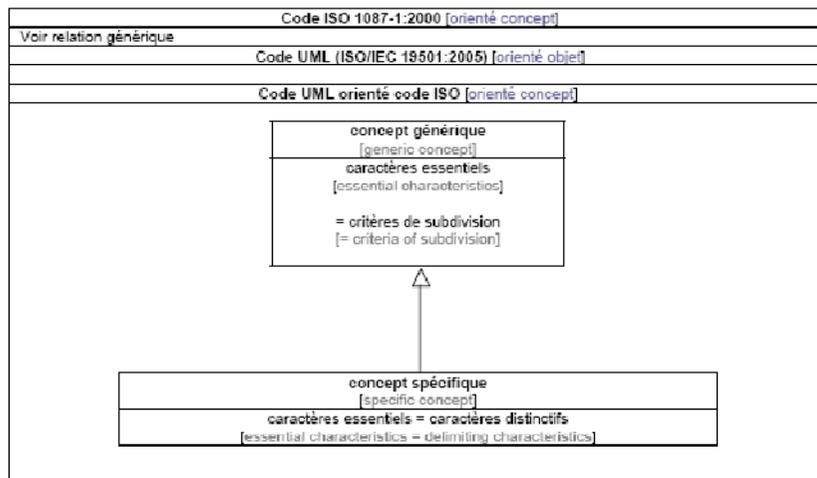
Tab. 2 : Concept subordonné

Parallèlement, l'ISO définit un concept superordonné comme un concept intégrant ou un concept générique. En UML, la notion de classe générale (classe de base/classe parent) est l'équivalent du seul concept générique. L'agrégat est l'équivalent du concept intégrant. Parallèlement au schéma du concept subordonné, nous opterons par conséquent pour la modélisation du concept superordonné suivante (voir Tableau 3).



Tab. 3 : Concept superordonné

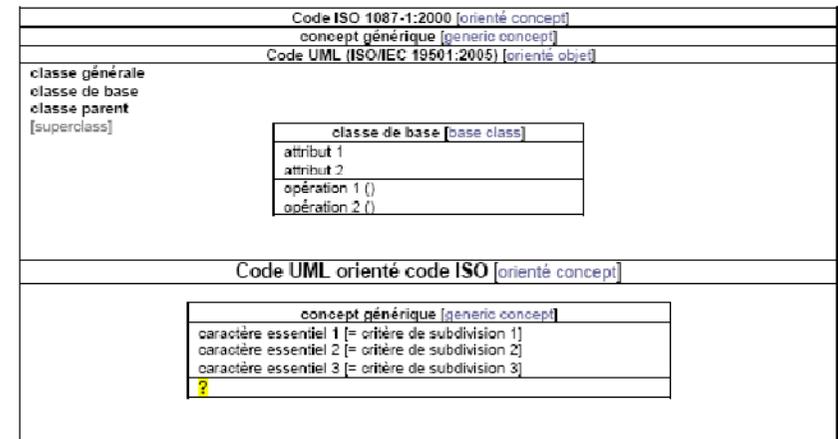
Le concept spécifique sera modélisé comme un rectangle depuis lequel part un trait orné d'une flèche vide fermée vers le concept générique (voir Tableau 4).



Tab. 4 : Concept spécifique

Nous référant à la discussion au sujet des critères de subdivision, le concept générique ISO adoptera la forme du rectangle classique modélisant tout concept, partitionné en deux rangées dont la première hébergera les caractères essentiels, qui eux renvoient aux critères de subdivision reliant le concept générique en question aux concepts spécifiques qui équiperont ces critères de subdivision des caractères distinctifs appropriés.¹

Le concept générique est représenté par un rectangle (et ses relations) au-dessus de l'extrémité de la flèche sous forme d'un triangle fermé vide.



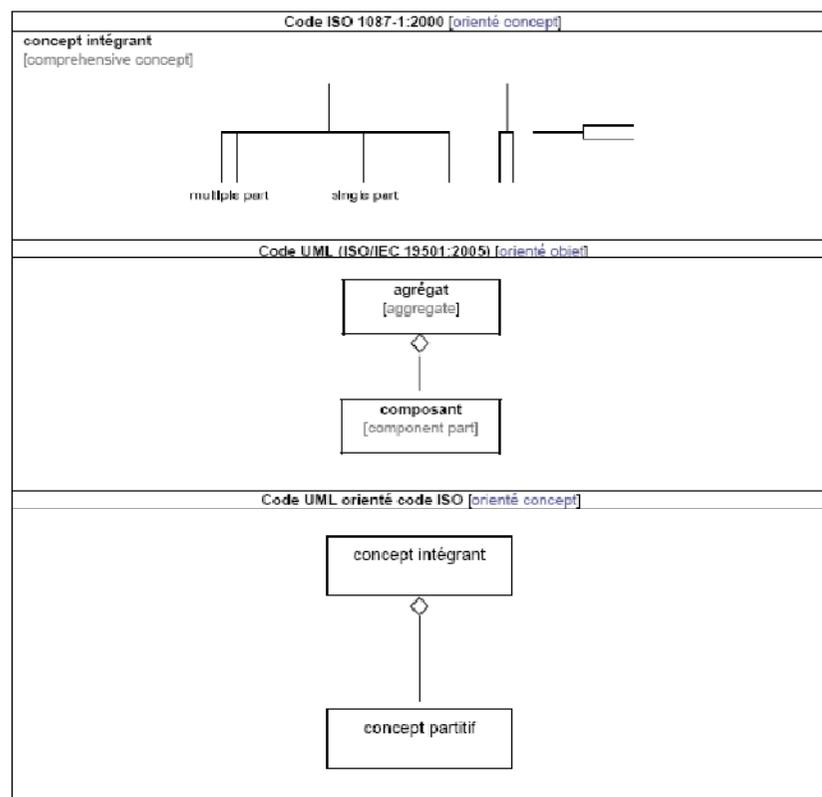
Tab. 5 : Concept générique

En UML, l'agrégat est représenté comme une classe se trouvant à l'extrémité d'une flèche ornée d'un losange vide (agrégation simple) ou d'une flèche ornée d'un losange plein (agrégation composite). C'est au-dessus d'une fourchette représentant la relation partitive que figure le concept intégrant en ISO. Reprendre la solution UML pour le schéma orienté concept nous semble bien

¹ Une modélisation des opérations UML dans un système modélisé orienté concept fera l'objet d'une autre étude.

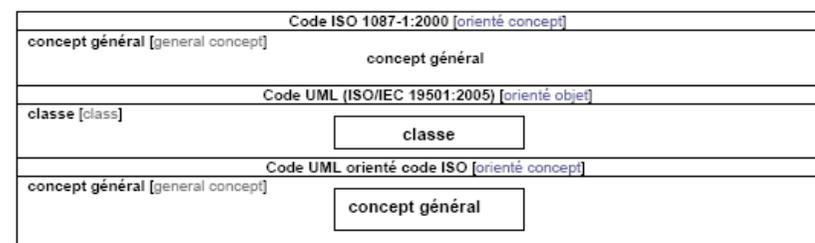
pratique. En même temps le concept partitif reprend la position qu'occupe le composant en UML.

En ISO, le concept partitif est défini comme le concept subordonné dans une relation partitive considéré comme l'une des parties constituant le tout. Le nom de composant [Audibert, 2007: 152] ne figure pas explicitement dans ISO/IEC 19501, ni n'est doté d'une définition dans l'UML [2003]. Cependant, le glossaire d'ISO/IEC 19501 mentionne *component part* sous *aggregation*, et définit ainsi indirectement la notion du *component part*. Le symbole du composant est celui de la classe en fin du symbole de l'agrégation (voir Tableau 6).

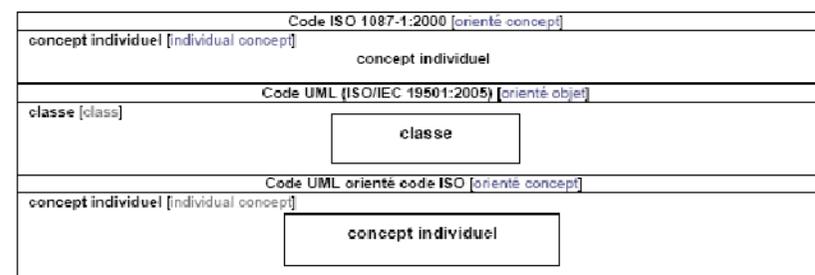


Tab. 6 : Concepts intégrant et partitif

Le code UML ne distingue pas les concepts individuels des concepts généraux, comme le font les normes ISO. Nous proposerions d'ajouter ces deux types de concepts à notre liste d'items modélisables orientés concept. Dans ce contexte, nous adopterons la représentation du concept (Tableau 1) pour modéliser tous les types de concepts (concept individuel, général, générique, spécifique, intégrant, partitif, superordonné et subordonné) (voir Tableau 7 et 8).



Tab. 7 : Concept général



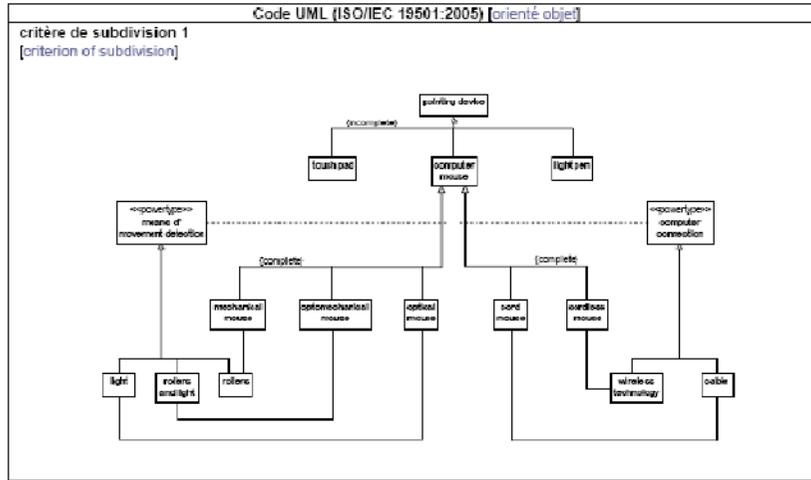
Tab. 8 : Concept individuel

3. Caractères et critères de subdivision

Afin de mieux comprendre comment le concept pourra abriter caractères essentiels et distinctifs en appliquant le schéma UML "attributs-valeurs", nous nous référons aux schémas suivants où figurent trois modes de représentation.

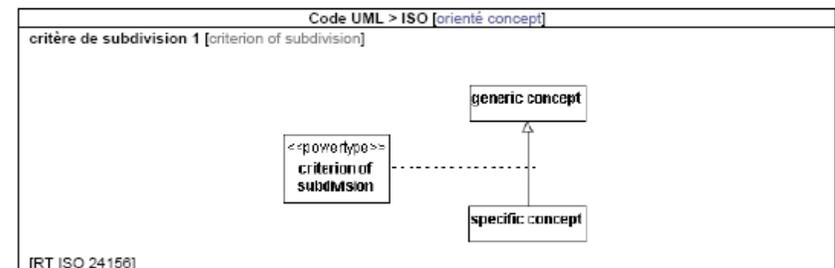
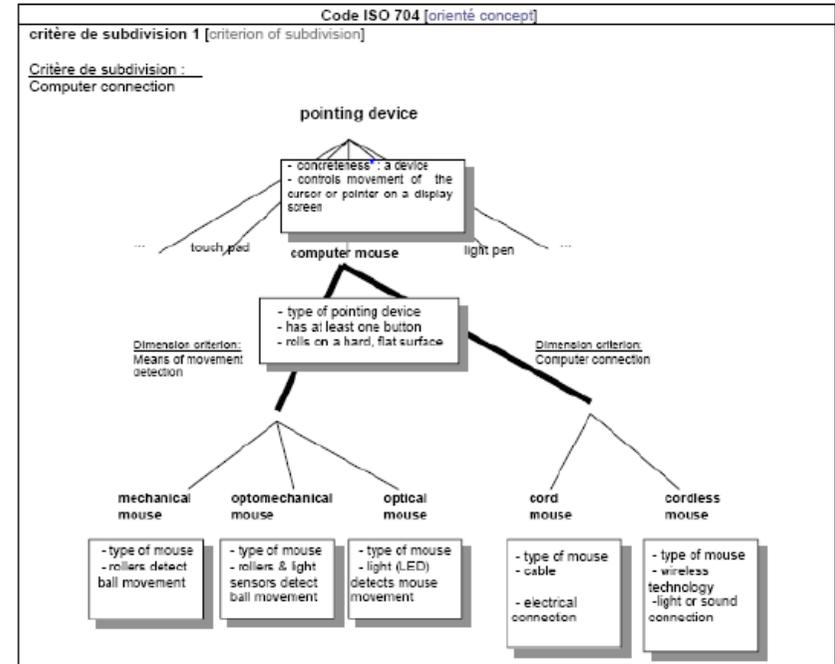
Le premier mode de représentation UML consiste en une classe dont la mention <<power type>> figure au-dessus du critère de subdivision. Une telle classe s'apparente en fait à une classe association [Audibert, 2007: 56] qui héberge le critère de subdivision reliant les classes de base aux classes spécialisées au

moyen d'un trait pointillé en perpendiculaire par rapport au symbole de généralisation. La classe <<powerType>> ne mentionne pas de couples attributs-valeurs de la classe de base. Le Tableau 9 visualise les représentations telle qu'elles figurent dans le RT ISO 24156 et l'ISO 704 respectivement. Si nous superposons les deux schémas, nous pouvons en déduire que la mention UML <<powerType>> (*means of movement detection; computer connection*) correspond à la mention *Dimension criterion* (= critère de subdivision) en ISO.



Tab. 9 : Critère de subdivision 1 (UML)

Parallèlement, la représentation ISO met les critères de subdivision *means of movement detection* et *computer connection* au-dessus d'un trait horizontal le long du trait reliant concepts superordonnés et subordonnés. Sous ce trait figure le nom du critère de subdivision (voir Tableau 10).



Tab. 10 : Critère de subdivision 1 (ISO)

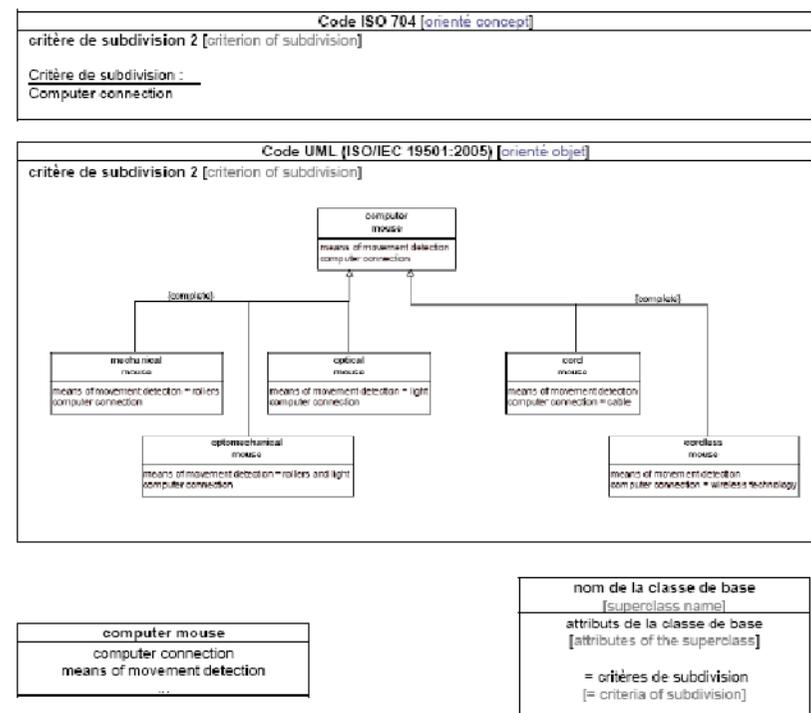
Ayant observé que cette représentation ne pourra guère adopter sans modification le code UML, nous avons modélisé le premier critère de subdivision comme dans le schéma ci-dessus (voir Tableau 10). Il est important de constater que les caractères distinctifs ne sont pas encore représentés, ce qui reflète évidemment un vide dans le modèle ISO par rapport aux couples attributs-

valeurs dans l'homologue UML. En d'autres termes, nous ne verrions plus les caractères distinctifs dans cette représentation, ce que nous ne préfererions pas.

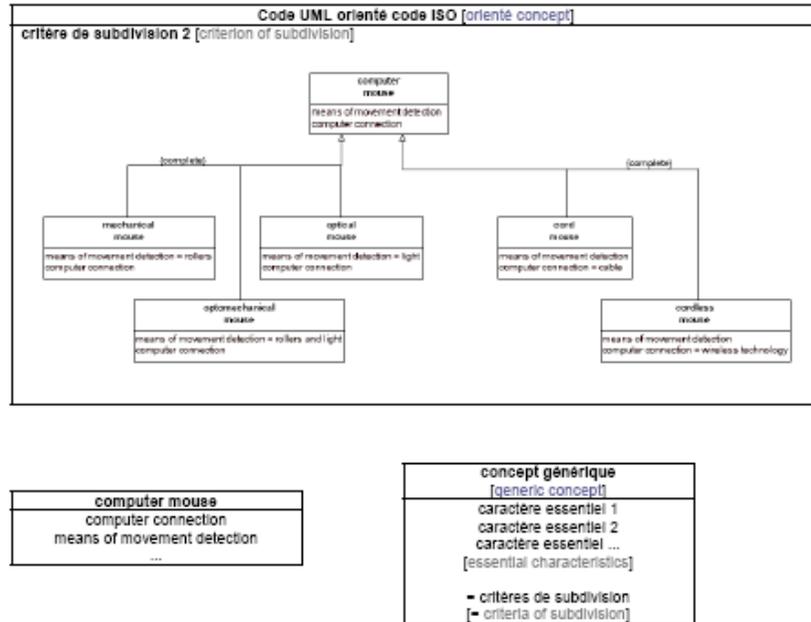
Dans un deuxième mode de représentation, les critères de subdivision réfèrent aux caractères (essentiels) du concept générique. C'est ainsi que nous pourrions maintenir que les critères de subdivision *means of movement detection* et *computer connection* figurent en tant que caractères essentiels dans un schéma qui est adaptable à l'ISO.

Cela nous permettra d'inclure où figurent les caractères essentiels et distinctifs dans un schéma modélisable par le modèle UML. Les critères de subdivision sont représentés comme des attributs de la classe de base, tout en sachant que ces attributs ne réfèrent pas aux caractères distinctifs du concept générique. Cependant, dans les classes UML spécialisées (sous-classes), ces attributs sont dotés de valeurs, figurant après le "=". Ce sont ces couples attributs-valeurs qui peuvent être convertis en des couples caractères essentiels = caractères distinctifs des concepts spécifiques dans le schéma ISO.

Dans le schéma que nous avons développé pour le RT ISO 24156, le schéma suivant (voir Tableau 11) pourra servir d'instrument compatible avec les codes UML et ISO respectivement.

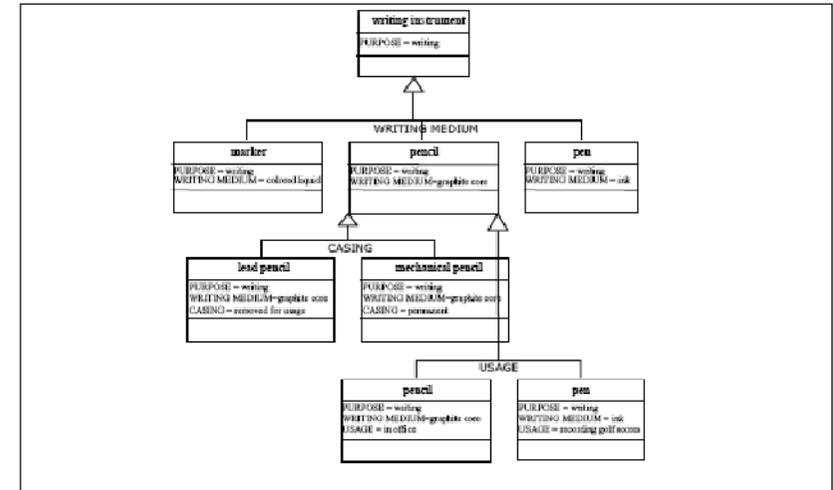


Tab. 11 : Critère de subdivision 2 (UML)



Tab. 11 : Critère de subdivision 2 (ISO)

C'est à l'aide d'un *discriminator* que Madsen & Thomsen [2008] envisagent de représenter les critères de subdivision ou caractères. Ci-dessous (voir Tableau 12), nous nous référons au schéma de ce troisième critère de subdivision.



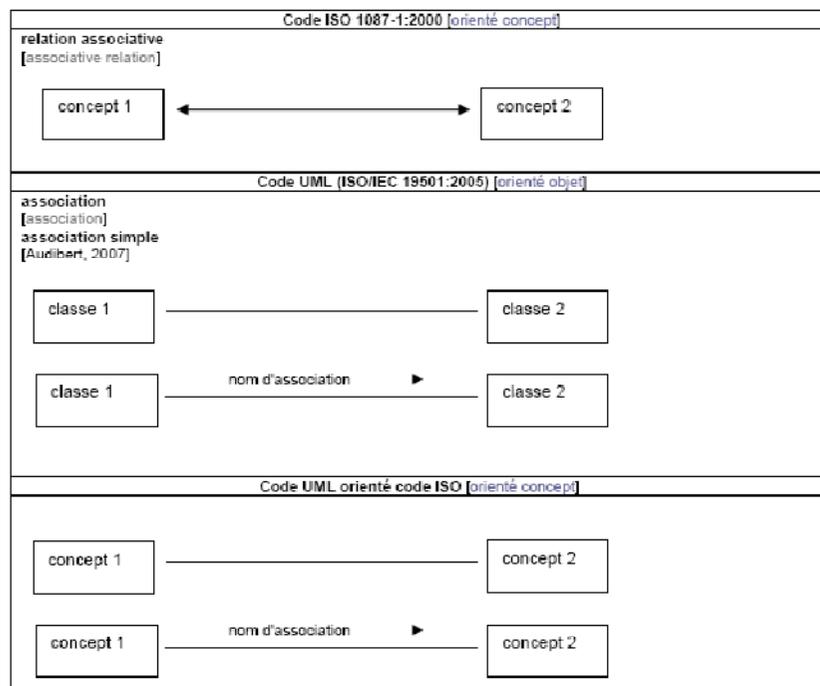
Tab. 12 : Critère de subdivision 3

S'il y a lieu de choisir un critère de subdivision, nous opterions pour le deuxième critère en raison de sa capacité de représenter tous les caractères, y inclus les caractères distinctifs, et de sa compatibilité avec le modèle UML quant à représenter les concepts dans un système de concepts.

4. Relations entre les concepts

Les associations sont utilisées pour connecter les classes dans le diagramme de classe ; dans ce cas, la terminaison de l'association (du côté de la classe cible) est généralement une propriété de la classe de base.

En ISO, les relations associatives sont systématiquement bidirectionnelles, alors que l'UML les représente comme une ligne continue connectant deux classes pour indiquer l'association générale entre deux classes. La ligne peut être renseignée par un nom d'association, indiquant la nature de l'association, ornée d'une pointe de flèche pleine indiquant le sens dans lequel la relation doit être comprise. Le langage UML peut représenter des associations unidirectionnelles ce qui est impossible en terminologie. Les concepts ISO peuvent entrer dans les deux types de relations associatives, à savoir uni- et bidirectionnelles, ce qui donne lieu à un ajout de modalités directionnelles dans les normes ISO.



Tab. 13: Relation associative

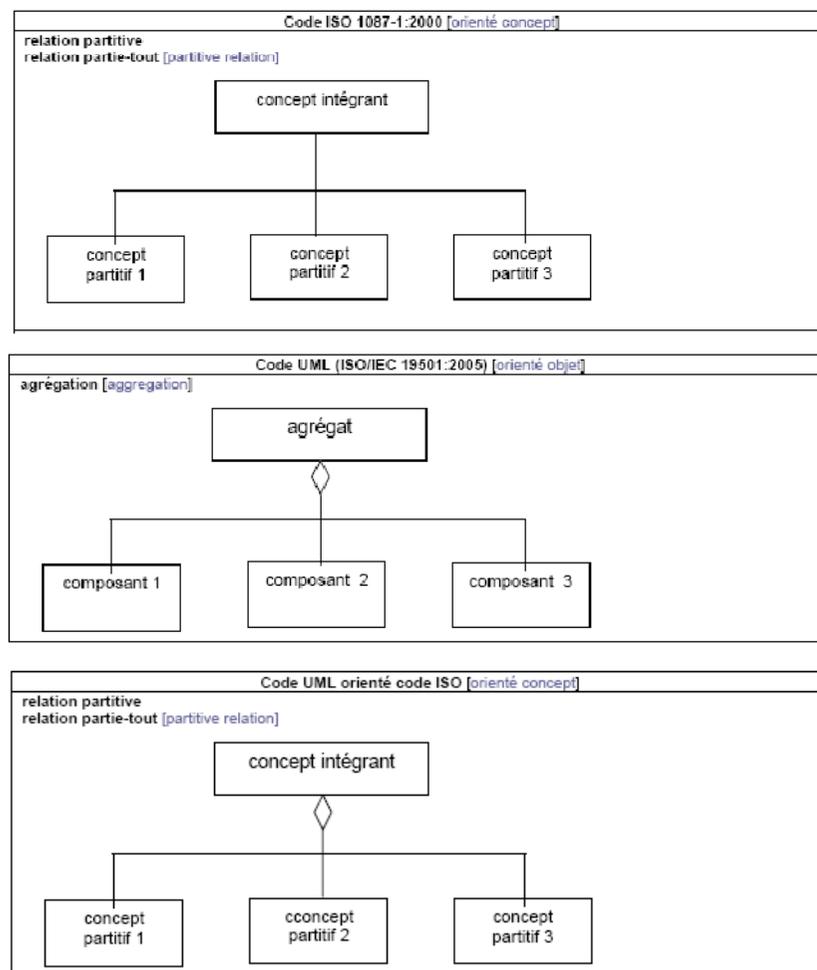
En ISO 1087-1 une relation hiérarchique est toute relation entre deux concepts qui est soit une relation générique, soit une relation partitive. Le document UML ne mentionne pas ce type de relation. La définition ISO est une définition par extension :

relation hiérarchique : relation entre deux concepts qui est soit une relation générique, soit une relation partitive [ISO 1087-1:2000, p. 4]

Or, en UML il s'agit respectivement d'une relation de généralisation (relation générique) ou d'une agrégation (relation partitive). En UML, rédigé en anglais, il est fait mention de *relationship* et non de *relation*. L'UML n'explicite par conséquent pas de concept superordonné pour la généralisation et l'agrégation.

L'UML considère une agrégation comme une association qui représente une relation d'inclusion structurelle ou comportementale d'un élément dans un ensemble. L'ISO 1087-1 considère la relation partitive comme une relation hiérarchique, et non comme une association (relation associative).

En dépit de ces différences entre les deux métamodèles, il nous semble justifié de faire correspondre l'agrégation UML à la relation partitive ISO. Par conséquent nous modélisons la relation partitive comme une agrégation, et nous considérons l'agrégation comme une relation hiérarchique à la ISO. D'ailleurs, dans le RT ISO 24156, nous symbolisons l'agrégation comme une relation hiérarchique. C'est ainsi que nous arrivons à la modélisation de la relation partitive telle qu'elle figure au Tableau 14.



Tab. 14 : Relation partitive

Lorsque l'on souhaite modéliser une relation partie-tout où une classe constitue un élément plus grand (tout) composé d'éléments plus petits (partie), il faut utiliser une agrégation à la UML.

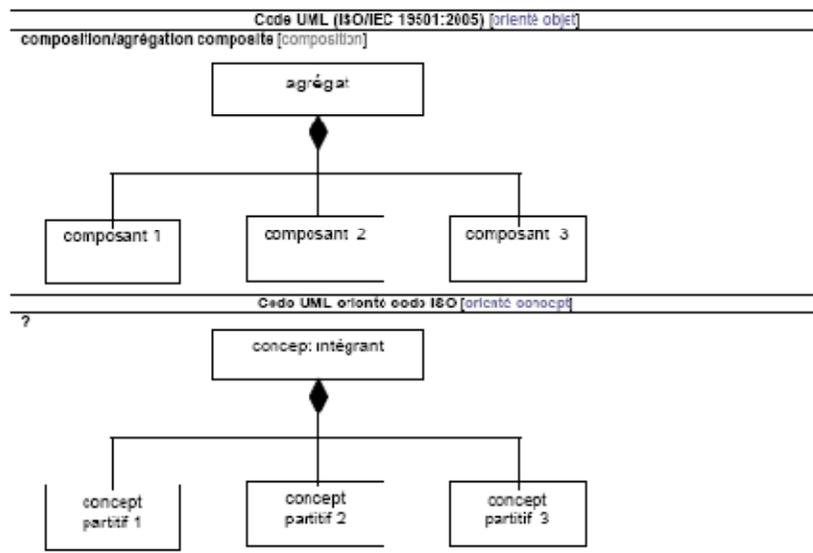
En UML, une agrégation est représentée par un losange vide pointant vers l'agrégat. Contrairement à une association simple, l'agrégation est transitive. La

signification de cette forme simple d'agrégation est uniquement conceptuelle. Elle ne contraint pas la navigabilité ou les multiplicités de l'association. Elle n'entraîne pas non plus de contrainte sur la durée de vie des parties par rapport au tout [Audibert, 2007: 59].

La relation partitive ISO est définie comme une relation entre deux concepts dans laquelle le concept intégrant constitue le tout et le concept partitif une partie de ce tout.

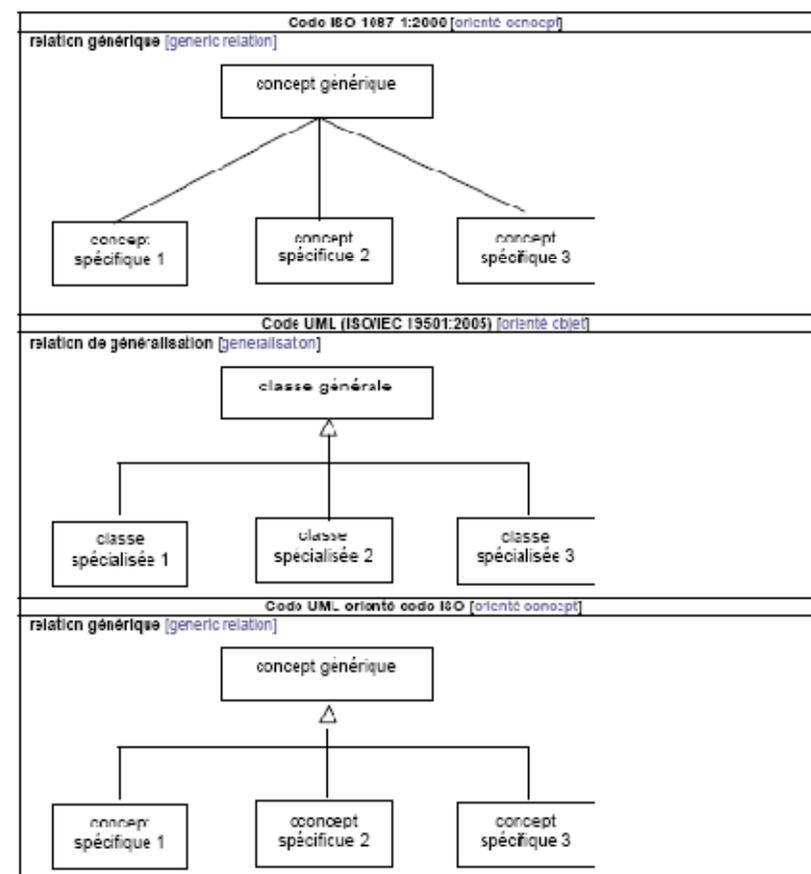
Toujours au sujet de la relation partitive ISO, nous devons signaler que le modèle UML offre une nuance supplémentaire en ce qu'il fait une distinction entre une agrégation (simple) et une agrégation composite ou composition. La composition décrit une contenance structurelle entre le tout et les parties. Ainsi, la destruction de l'objet composite implique la destruction de ses composants. Une instance de la partie appartient toujours à une instance de l'élément composite [Audibert, 2007: 60]. L'ISO ne fait pas cette distinction et grâce au modèle UML, il sera désormais possible d'ajouter la composition aux concepts ISO, tout en lui dotant une modélisation directement empruntée de l'UML. Voir le Tableau 15² qui montre un losange plein pointant vers l'agrégat.

² Le point d'interrogation signale notre incertitude à l'égard d'une mention équivalente de *composition* dans le modèle ISO.



Tab. 15 : Composition

Pour représenter une modélisation orientée concept d'une relation générique à l'aide du langage UML, nous nous référons au Tableau 16. Nous adopterons à cet effet le symbole UML, c'est-à-dire une flèche sous forme de triangle fermé pour visualiser cette relation.



Tab. 16 : Relation générique

Quant à aboutir à une représentation modélisable et orientée concept du schéma conceptuel, nous avons opté pour le modèle ISO, un schéma homologue faisant défaut en UML. Le diagramme de classes UML modélise les *concepts* du domaine d'application [classes] sans pour autant être capable de représenter une partie du modèle. Le champ de classes et le système de classes correspondent au champ conceptuel et le système de concepts respectivement (voir Tableau 17).

Code ISO 1087-1:2000 [orienté concept]	
schéma conceptuel [concept diagram]	
Code UML (ISO/IEC 19501:2005) [orienté objet]	
diagramme de classe [schéma de classes] [class diagram]	
Code UML orienté code ISO [orienté concept]	
schéma conceptuel [concept diagram]	
Code ISO 1087-1:2000 [orienté concept]	
champ conceptuel [concept field]	
Code UML (ISO/IEC 19501:2005) [orienté objet]	
champ de classes [set of classes]	
Code UML orienté code ISO [orienté concept]	
champ conceptuel [concept field]	
Code ISO 1087-1:2000 [orienté concept]	
système de concepts [concept system]	
Code UML (ISO/IEC 19501:2005) [orienté objet]	
terminological concept model [système de classes]	
Code UML orienté code ISO [orienté concept]	
système de concepts [concept system]	

Tab. 17 : Schéma conceptuel, champ conceptuel, système de concepts

5. Résultats et conclusion

Grâce aux schémas ci-dessus, nous espérons avoir pu modéliser la plupart des items issus des normes ISO pour que ceux-ci puissent être échangés sur la base d'un langage tel l'UML. Cette expérience nous a en même temps montré que certains items ISO n'entrent pas sans modification dans le modèle UML et vice-versa. C'est dans ce contexte que la modélisation ISO n'a pas su héberger les opérations, la navigabilité d'associations, ou encore la composition.

En revanche, cet exercice comparatif pourra constituer une base sur laquelle le modèle UML pourrait envisager l'inclusion, après conversion ISO >> UML des items suivants: concepts superordonnés, concepts subordonnés, ou encore relation hiérarchique.

Bibliographie

Audibert, L. UML 2. Édition 2007-2008, Institut Universitaire de Technologie de Villetaneuse – Département Informatique, 2007.

[<http://www-lipn.univ-paris13.fr/~audibert/pages/enseignement/cours.htm>]

Fournier, J.-P. & Glorioso, M. Guide de l'utilisateur du LDI, Laboratoire didactique informatique, Université de Genève, 2002.

[<http://www.infeig.unige.ch/support/se/lect/uml/web.html>]

ISO 1087-1, Travaux terminologiques -Vocabulaire - Partie 1 : Théorie et application. Genève : International Standards Organisation, 2000.

ISO 704, Terminology work - Principles and methods. Genève : International Standards Organisation, 2000.

ISO/IEC 19501, Information technology - Open Distributed Processing - Unified Modeling Language (UML) Version 1.4.2. Genève : International Standards Organisation, 2005.

ISO/PDTR 24156.6b, Guidelines for using UML notation in terminology work. ISO copyright office, Case postale 56, CH-1211 Genève 20, 2008.

Madsen, B. N. & Thomsen, H. E. "Terminological Concept Modelling and UML Diagrams", 2008 [à paraître].

OMG, Unified Modeling Language: Superstructure, version 2.0, août 2003.

Wüster, E. Einführung in die allgemeine Terminologielehre und terminologische Lexikographie. Copenhagen: LSP Centre, Copenhagen School of Economics, 1985 (1979).



APPLICATIONS

TA statistique à petits corpus pour de petits sous-langages

Najeh Hajlaoui, Christian Boitet

Laboratoire LIG, GETALP – Université Joseph Fourier,
385 rue de la bibliothèque, BP n° 53,
38041 Grenoble, Cedex 9, France
Najeh.Hajlaoui@imag.fr, Christian.Boitet@imag.fr

Résumé :

Nous avons appliqué un système de TA statistique au "portage linguistique" de l'arabe au français de CATS, un système traitant le contenu de brefs messages spontanés en langue naturelle (SMS). Il s'agit d'un "sous-langage" très restreint. Nous ne disposons que d'un très petit corpus parallèle, augmenté d'un dictionnaire bilingue assez complet lié à l'application choisie (petites annonces en occasion automobile). Bien que la TA statistique soit réputée ne fonctionner assez bien que si l'on dispose de très grands corpus parallèles, le système que nous avons construit avec Pharaoh a produit des résultats satisfaisants, au sens où les descripteurs de contenu obtenus sont assez proches de ceux obtenus à partir des SMS correspondants en arabe. Il semble donc qu'on puisse se passer de très grands corpus pour utiliser efficacement la TA statistique sur des "sous-langages" très restreints : les traductions ne sont pas très "fluides", mais elles sont "adéquates", et ce même si les deux "langues-mères" des deux sous-langages considérés sont assez distantes.

Mots-clés : sous-langage, langue générale, langue standard, énoncés spontanés, traduction statistique, extraction de contenu.

1. Introduction

Les chercheurs du groupe TAUM à l'UdM (Université de Montréal) furent les premiers à se rendre compte de la relative facilité de construction de certains systèmes de TALN, et de leur grande qualité, quand on pouvait les limiter à des « sous-langages ». Après avoir connu une « bonne surprise » avec la traduction de

bulletins météo (système TAUM-météo)¹, le groupe TAUM a cherché longtemps (en vain d'ailleurs²) d'autres sous-langages aussi « faciles » pour la méthode employée (programmation « experte » reposant sur une étude précise du sous-langage en question et sur la mise en œuvre d'heuristiques adaptées).

Cela conduisit les linguistes du groupe TAUM (surtout R. Kittredge et J. Lehrberger) à approfondir la notion de sous-langage, introduite par Z. Harris en 1968, pour la rendre opérationnelle. Beaucoup de chercheurs les ont suivis dans cette voie, et ont montré l'importance de la notion de sous-langage dans le traitement du texte d'un langage naturel amélioré ou simplifié par l'utilisation de restrictions lexicales, syntaxiques ou sémantiques spécifiques (Kittredge and Lehrberger 1982a), (Grishman and Kittredge 1986), (Slocum 1986), (Biber 1993), (Sekine 1994). Dans ce dernier article, intitulé « A New direction for Sublanguage NLP », Satoshi Sekine montre de façon convaincante que la restriction (explicite ou implicite) à des sous-langages « assez restreints » conduit en général au succès : on arrive à construire des systèmes très performants avec un investissement très raisonnable en temps humain de spécialistes et en ressources de calcul (temps, place). Il cite lui aussi le cas du système TAUM-METEO.

Nous présentons dans la première partie quelques définitions possibles du terme *sous-langage* et un exemple de sous-langage réel. Dans la deuxième partie, nous décrivons quelques méthodes de portage linguistique d'applications traitant des énoncés spontanés en langue naturelle dont nous détaillons, dans la dernière partie, le portage par TA statistique, et son efficacité, au moins dans un cas de sous-langage très petit et restreint à une tâche, même si on ne dispose que d'un dictionnaire bilingue assez complet et d'un petit corpus parallèle.

¹ Ce système fut construit par le groupe TAUM de l'UDM en 1975-76 (Isabelle 1984), (Chandioux 1988). Il fut mis en service opérationnel à Environnement Canada le 24 mai 1977 par la société J. Chandioux Conseil. C'est un système de traduction automatique qui marche extrêmement bien pour le sous-langage des bulletins météo (mais pas pour ceux des situations ou des avertissements météo !). Il traduit environ 20 M mots/an d'anglais en français et 10 M mots/an dans l'autre sens, avec une qualité liée à la tâche de plus de 97 % (moins de 3 opérations d'édition pour 100 mots traduits).

² NTT a trouvé une application de ce type, la traduction en anglais des brèves ("flash reports") du Nikkei (bourse de Tokyo), et développé pour cela le système ALTFlash, totalement automatique, de grande qualité, et « bimoteur » (système à patrons avec secours une version spécialisée du système général ALT/JE).

2. Sous-langage naturel

2.1. Selon Zellig Harris

Plusieurs définitions pour le terme « sous-langage » ont été données. Il semble que la première a été proposée par Zellig Harris (Harris 1968) : « *Certain proper subsets of the sentences of a language may be closed under some or all of the operations defined in the language, and thus constitute a sublanguage of it.* »

Autrement dit,
« Un sous-ensemble strict d'une langue peut être fermé pour un sous-ensemble des opérations définies dans la langue, et ainsi en constituer un sous-langage. »

Cette définition semble à première vue incorrecte, car les phrases d'un « sous-langage » ne sont souvent pas des phrases (correctes) de la « langue standard », dont on suppose que parle un linguiste, et alors on ne pourrait pas parler de « sous-ensemble » au sens usuel.

Par exemple, il est acceptable dans un article de biochimie de dire « *The polypeptides were washed in hydrochloric acid* », mais pas « *hydrochloric acid was washed in polypeptides* ».

Comme Z. Harris savait parfaitement ce qu'est un sous-ensemble d'un ensemble, nous sommes conduits à admettre qu'il entendait par le terme « langue » une extension du terme « langue standard ». Nous utiliserons donc le terme « langue standard » pour désigner l'ensemble des énoncés d'une communauté linguistique formés d'une façon « correcte » par rapport à la grammaire et au vocabulaire usuels, tels qu'enseignés dans les cours de langue, et nous appellerons « langue générale » l'union d'une langue standard et de toutes ses variantes (jargons, langues de spécialité, parlars régionaux, langages « techniques », et langages « secrétés » par des contextes socioprofessionnels).

Dans la définition précédente, assez générale, Harris ne dit pas de quelles opérations il parle. Mais il propose ensuite une définition « inductive » plus précise : un sous-langage SL est le plus petit ensemble contenant une base B et fermé (stable) par un ensemble de règles R.

SL = <B, R>, où

- la base B est un ensemble « noyau » d'énoncés ou schémas d'énoncés observés ;

- les règles R sont des règles de transformation comme la passivation, l'extraposition, l'interrogation, la mise au passif, à l'impersonnel, à l'interrogatif, ou simplement à un autre temps ou un autre mode, etc.

Un énoncé du sous-langage est donc dans le "noyau", ou bien il résulte d'un énoncé du sous-langage par une transformation de R. Par exemple, si « *The enzyme activated the process.* » est dans le sous-langage, et si la passivation est une des transformations permises, « *The process was activated by the enzyme.* » le sera aussi.

Cette définition est difficilement utilisable en pratique, car elle ne fournit pas de moyen opérationnel pour identifier le noyau et les règles caractérisant un sous-langage observé.

2.2. Définition selon l'usage

Une deuxième définition a été donnée par Bross et autres (Bross, Shapiro et al. 1972) :

« *Informally, we can define a sublanguage as the language used by a particular community of speakers, say, those concerned with a particular subject matter or those engaged in a specialized occupation.* »

Autrement dit, un sous-langage est l'ensemble des énoncés susceptibles d'être prononcés par une communauté (de communication) en certains temps et certains lieux.

Grishman et Kittredge (Grishman and Kittredge 1986), puis Deville (Deville 1989), définissent aussi un sous-langage comme une forme spécialisée d'une langue naturelle employée dans un domaine ou un thème particulier.

Cette définition est observationnelle et expérimentale, et prend directement en compte un contexte d'usage particulier. C'est celle qui a été utilisée dans le projet TAUM-METEO (1972-1973) et pour des manuels de maintenance d'avions dans le cadre du projet TAUM-AVIATION (1974-1981) et du PN-TAO (Projet National de TAO, 1982-87) en France.

À titre d'exemples de sous-langages, on peut citer les bulletins METEO, les manuels de maintenance d'un avion, les articles scientifiques concernant la pharmacologie, les rapports de radiologie, les annonces immobilières, etc.

Un sous-langage est alors caractérisé par un vocabulaire spécialisé, une sémantique restreinte, et dans beaucoup de cas une syntaxe spécialisée. Ainsi, les

prépositions et articles normalement obligatoires peuvent être omis. Exemple : « *trappe visite réservoir avant gauche* », « *vent fort lac Saint-Jean* », « *orienté objet* ».

Cette définition a été précisée par Kittredge de la façon suivante.

Un sous-langage est un sous-ensemble d'une langue :

- qui fait référence à un *domaine particulier* ou à une famille de domaines liés,
- dont l'ensemble des phrases et des textes reflète *l'usage d'une communauté* de personnes ayant en commun des connaissances élaborées du domaine,
- qui a les propriétés fondamentales d'un *système linguistique* : consistance, complétude, économie d'expression, etc.,
- qui est *maximal* par rapport au domaine (il n'y en a pas de plus grand qui possède ces propriétés).

Type de langue Genre de langue	Langue générale	Langue standard	Sous-langage
Textes	Énoncés corrects spontanés	Énoncés corrects	Énoncés spontanés
Grammaire	Usuelle + spécifique	Usuelle	Spécifique
Vocabulaire	Usuel + restreint	Usuel	Restreint

Tableau 1 : types de langues et caractéristiques associées

Le Tableau 1 résume les trois différents types de langue : langue générale, langue standard, et sous-langage.

1.1. Exemple : sous-langage de l'arabe des SMS en occasion automobile

CATS est une application de e-commerce déployée en Jordanie sur le réseau FastLink (Daoud, 2006). Elle traite des petites annonces envoyées par SMS et concernant l'occasion automobile (Cars), l'immobilier à Amman (RealEstate), l'emploi (Jobs), et autres (Misc). Elle permet de "poster" des petites annonces et de mettre en contact les personnes susceptibles d'être intéressées (Daoud, 2005). Voici quelques exemples de tels SMS, avec une traduction en français.

مطلوب سياره هونداى موديل 97 والسعر ما بين 3500 الى 3750	Recherche voiture Honda, modèle 97, prix entre 3500 et 3750
مطلوب سياره سبور	Recherche voiture sport
اريد سياره مرسيدس موديل 82 لون ابيض	Je veux une voiture Mercedes modèle 82 couleur blanche

Tableau 2: Exemples de SMS arabe

Pour le domaine de l'automobile (Cars), la taille du vocabulaire utilisé est d'environ 638 entrées principales. Comme il comprend des mots étrangers translittérés, éventuellement de plusieurs façons, on y ajoute des variantes, dites entrées secondaires. Voici quelques exemples d'entrées.

Entrée arabe	Principale /secondaire	UW (notation du concept dans Cars)	en français
الفا روميو	P	ALFA ROMEO(country>Italy,country>europe)	Alfa Romeo
الفا روميو	S	ALFA ROMEO(country>Italy,country>europe)	Alfaromeo
روميو	S	ALFA ROMEO(country>Italy,country>europe)	Romeo
A3	P	A3(country>germany,country>europe,make>AUDI)	A3
اي3	S	A3(country>germany,country>europe,make>AUDI)	a3

Tableau 3 : Entrées du dictionnaire de CATS (arabe)

Les énoncés sont très simples et très courts (zone sombre dans la figure suivante). Si l'on parcourt un corpus de tels SMS, on observe une convergence grammaticale très rapide, mais une convergence lexicale moins rapide, à cause des nouveaux motifs qui peuvent apparaître après un certain temps.

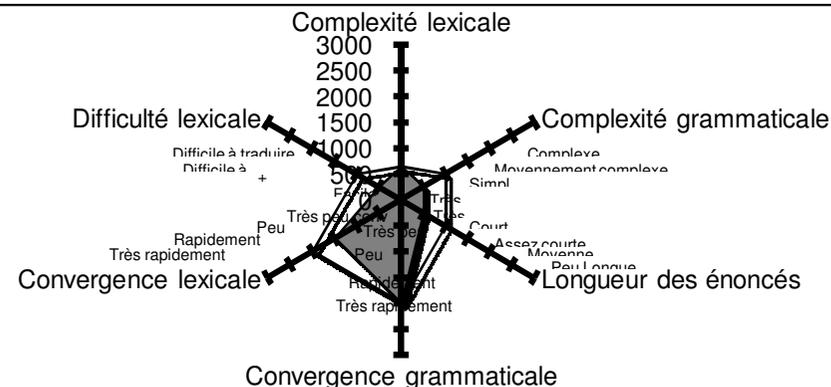


Figure 1 : analyse du sous-langage de l'automobile dans CATS

On peut aussi avoir des phrases simples et courtes (zone claire dans la figure précédente avec une convergence grammaticale et une convergence lexicale très rapides.

3. Multilinguïstation d'applications traitant des énoncés spontanés en langue naturelle

3.1. Problème et solutions possibles

Nous nous intéressons plus généralement à la multilinguïstation, ou "portage linguistique" (plus simple que la localisation) des services de gestion de contenu traitant des énoncés spontanés en langue naturelle, souvent bruités mais contraints par la situation. Tout service de ce type (soit App) est muni d'un extracteur de contenu (EC-App) produisant une forme interne spécifique (CRL-App) à partir de la langue "native" L1. Nos étapes sont les suivantes :

- Choix de l'application à porter et choix des couples des langues.
- Étude de la situation traductionnelle.
- Choix de la ou des méthodes de portage linguistique possibles, en fonction du résultat de l'étape précédente.
- Réalisation du portage linguistique.
- Évaluation du portage.

Nous avons illustré cette approche par le portage de la partie *Cars* de CATS. L'étude de la situation traductionnelle associée nous a menés à définir et expérimenter trois stratégies de portage pour ce couple de langues assez distant (arabe-français) : (1) *localisation "interne"*, i.e. adaptation à L2 de l'extracteur de contenu (EC) donnant EC-App-L2 ; (2) *localisation "externe"*, i.e. adaptation d'un EC existant pour L2 au domaine et à la représentation de contenu de App (EC-X-L2-App); (3) *traduction* des énoncés de L2 vers L1.

Le choix de la stratégie est contraint par la situation traductionnelle : types et niveau d'accès possibles (accès complet au code source, accès limité à la représentation interne, accès limité au dictionnaire, et aucun accès), ressources disponibles (dictionnaires, corpus), compétences langagières et linguistiques des intervenants pour la multilinguisation des applications.

Les trois stratégies ont été expérimentées et ont donné de bons résultats sur le portage de l'arabe au français de la partie *Cars* de CATS (Hajlaoui, 2007).

3.2. Localisation interne

SMS en français	CRL-CATS obtenue
recherche	[S]
voiture	wan(saloon:0A, wanted:00)
OPEL	mak(saloon:0A, OPEL(country>germany,country>europe):0I)
VECTRA	mod(saloon:0A, Vectra(country>germany,country>europe,make>OPEL):0N)
	[/S]
recherche à l'achat	[S]
NISSAN	wan(saloon:00, wanted:00)
Sunny modèle	mak(saloon:00, NISSAN(country>japan):0L)
92 à 95	mod(saloon:00, Sunny(country>japan,make>NISSAN):0S)
	yea(saloon:00, 92:16)
	yea(saloon:00, 95:1C)
	[/S]

Tableau 4 : exemples de résultat de la localisation interne

En localisation interne, la partie grammaticale a été très faiblement modifiée, ce qui prouve que, malgré la grande distance entre l'arabe et le français, ces deux sous-langages sont très proches l'un de l'autres, une nouvelle illustration de l'analyse de R. Kittredge. Le Tableau 4 quelques résultats de SMS français.

Le Tableau 5 montre la répartition de l'effort pour le portage interne en terme de temps de travail et de pourcentage du code modifié ou ajouté.

Adaptation de EC-CATS	Dictionnaire	Règles
Temps de travail (H)	100	45
% du code modifié	90	5

Tableau 5: Répartition de l'effort pour le portage interne

3.3. Localisation externe

La localisation externe a été expérimentée sur une deuxième application de recherche de musique (IMRS) (Kumamoto 2007) qui traite des énoncés spontanés en japonais en adaptant le même extracteur de contenu du français construit initialement par H. Blanchon (Blanchon 2003) pour le domaine du tourisme, en restant dans la même langue, puis en changeant de langue (anglais).

Pour IMRS (Kumamoto 2007), nous avons obtenu une représentation interne (IF-Musique pour le français et IF-Music pour l'anglais) qui contient chacune un vecteur composé de dix composants. Chaque composant correspond à un axe parmi dix. La valeur d'un composant est un nombre réel entre 0 et 7 qui correspond à sept degrés de l'échelle associée à l'axe en question. Le symbole « nil » veut dire « *don't care* ». Par exemple, l'axe « *Happy – Sad* » est caractérisé par sept valeurs intermédiaires, « *very happy*, » « *happy*, » « *a little happy*, » « *medium*, » « *a little sad*, » « *sad*, » et « *very sad*, » qui correspondent respectivement aux valeurs 7.0, 6.0, 5.0, 4.0, 3.0, 2.0, et 1.0.

SMS en français	(IF-CATS → CRL-CATS) obtenue
recherche voiture OPEL	S
VECTRA	wan(saloon, wanted)
	mak(saloon, OPEL(country>germany,country>europe))
	mod(saloon, Vectra(country>germany,country>europe,make>OPEL))
	/S
recherche à l'achat	S
NISSAN Sunny modèle	wan(saloon, wanted)
92 à 95	mak(saloon, NISSAN(country>japan))
	mod(saloon, Sunny(country>japan,make>NISSAN))
	yea(saloon, 95)
	/S

Énoncé en français	IF-Musique obtenue
je veux un morceau de musique calme et très solennel	{c:give-information+disposition+service(disposition=(desire, who=i), service=music, musique-spec=(nil 6,0 nil nil 7,0 nil nil nil nil nil))}
je veux un morceau de musique assez fort et clair	:{c:give-information+disposition+service(disposition=(desire, who=i), service= music, musique-spec=(3,0 nil nil 6,0 nil nil nil nil nil nil))}
Énoncé en anglais	IF-Music obtenue
I want a calm and very solemn music	{c:give-information+disposition+service(service=music, musique-spec=(nil 6,0 nil nil 7,0 nil nil nil nil nil))}
I want a little noisy and bright music	{c:give-information+disposition+service(service=music, musique-spec=(3,0 nil nil 6,0 nil nil nil nil nil nil))}

Tableau 6 : exemples de résultats obtenus par portage externe

Adaptation de FR-IF CATS	Dictionnaire	Règles
Temps de travail (H)	90	140
% du code modifié/ajouté	20	15
IMRS		
Temps de travail (H) (Fr ; En)	(20 ; 30)	(10 ; 20)
% du code modifié/ajouté (Fr ; En)	(3 ; 6)	(2 ; 4)

Tableau 7 : Répartition de l'effort pour le portage externe

Nous présentons et évaluons dans la section suivante le portage de CATS de l'arabe (langue source pour le portage) vers le français par traduction statistique des énoncés du français vers l'arabe (langue cible pour la traduction) et utilisation de l'extracteur de contenu original.

4. Portage par TA statistique avec un petit corpus

Le but n'est pas de produire des traductions parfaites, mais de produire des traductions permettant à l'extracteur de contenu "natif" d'extraire l'information pertinente.

La question est de savoir si cela est possible étant donné la très petite taille du corpus d'apprentissage disponible, et si oui quelle est la taille minimale suffisante d'un tel corpus.

4.1. Corpus arabe-français

Il s'agit toujours ici du domaine Cars du système CATS. Nous avons utilisé dans cette expérience des données d'entraînement et des données de développement.

Les données d'entraînement sont constituées d'un corpus parallèle, composé d'un corpus français obtenu par traduction manuelle (transtac_train.fr) d'un corpus arabe original (transtac_train.ar), et de ce corpus, dont une partie servira de référence pour les tests. Nous avons commencé l'expérience avec une très petite taille, 100 SMS, en calculant les scores BLEU et NIST obtenus pour la traduction des données de développement non utilisées dans l'étape d'entraînement. Nous avons progressivement augmenté la taille jusqu'à l'obtention d'une légère stabilité de la mesure.

Le Tableau 8 résume les informations concernant les données d'entraînement.

La comparaison entre la taille en mots du corpus français et arabe ne peut être définitive. Ici, la longueur moyenne d'un SMS français est plus élevée (9,93) que celle de l'arabe (8,75). Dans d'autres blocs de données, on peut observer l'inverse. Cela peut varier en fonction de la nature des données sources utilisées et de l'utilisation des variantes lexicales.

Une explication possible est que les données sources sont des données réelles rédigées par de vrais utilisateurs en Jordanie, et qu'aucune règle ne les empêche d'écrire « VOLKSWAGEN » en un seul mot ou en deux mots « VOLKS WAGEN », ou de même « LANDROVER » et « LAND ROVER » etc.

Taille du corpus d'entraînement	Taille du corpus arabe en mots	Nombre de mots par SMS arabe	Taille du corpus français en mots	Nombre de mots par SMS français	Taille du corpus arabe en octets	Taille du corpus français en octets
100	860	8,60	1010	10,10	8353	6362
200	1788	8,94	2051	10,26	17191	12828
300	2696	8,99	3052	10,17	25794	19004
400	3575	8,94	4006	10,02	33890	25015
500	4424	8,85	4954	9,91	42280	31150
600	5299	8,83	5907	9,85	50537	37222
700	6090	8,70	6809	9,73	58095	43087
800	6867	8,58	7742	9,68	65792	49040
900	7729	8,59	8811	9,79	74378	55990
1000	8685	8,69	9961	9,96	83999	63360

1100	9446	8,59	10803	9,82	91152	68716
Moyenne		8,75		9,93		

Tableau 8 : taille des données d'entraînement

4.2. Traduction statistique

La figure suivante présente l'adaptation de l'architecture générale d'un système de traduction statistique à notre cas. Étant donnée un SMS français f , nous cherchons la traduction arabe \hat{e} qui maximise $p(e|f)$, la probabilité qu'un SMS e soit la traduction de f , et la soumettons à CATS.

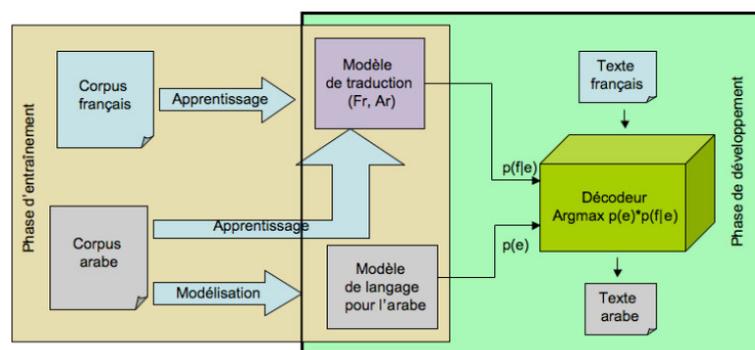


Figure 2 : Architecture générale d'un système de traduction statistique

Nous avons effectué un prétraitement des données avant de les transmettre au décodeur Pharaoh (Koehn, 2004). En particulier, nous avons aligné les données en utilisant l'outil GIZA++ (Och and Ney 2000).

Un modèle de langage pour la langue cible est nécessaire. Dans notre cas, l'arabe est la langue "native" de l'application et donc la langue cible pour la traduction. Nous devons donc construire un modèle de langage pour l'arabe. C'est une langue pour laquelle très peu de ressources sont disponibles et gratuites.

Nous avons trouvé un modèle de langage pour l'arabe, mais construit à partir du Coran, ce qui n'est pas du tout adapté au cas du sous-langage traité. Nous avons donc dû construire un modèle de langage pour le sous-langage *Cars* en nous basant sur ce que nous avons comme données. Nous avons utilisé pour cela le générateur de modèles de langage de Stolcke, disponible gratuitement sur le Web (Stolcke 2002) (<http://www.speech.sri.com/projects/srilm/>).

Nous avons construit notre modèle de langage en utilisant le même corpus d'entraînement que celui utilisé pour l'entraînement du décodeur de traduction. De la même façon, nous sommes partis d'une taille de corpus minimale et nous avons augmenté la taille au fur et à mesure jusqu'à obtenir des résultats satisfaisants.

4.3. Évaluation des résultats

4.3.1. Exemples

La Figure 3 montre quelques résultats obtenus pour une taille de corpus d'entraînement limitée à 400 SMS.

Langue référence (arabe originale)	Langue source (français)	Langue cible (TA statistique Fr → Ar)
مطلوب نيسان صني عادي موديل 93 إلى 97	recherche NISSAN Sunny manuelle modèle 93 à 97	مطلوب نيسان صني عادي موديل 93 إلى 97
مطلوب رينو عادي موديل 95 إلى 2000	recherche RENAULT Clio manuelle modèle 95 à 2000	مطلوب رينو عادي موديل 95 إلى 2000
مطلوب سيارة هونداي	recherche voiture HYUNDAI	مطلوب سيارة هونداي
مطلوب سيارة ميتسوبيشي	recherche voiture MITSUBISHI	مطلوب سيارة ميتسوبيشي
أبحث عن سيارة بي إم دبليو موديل 92	je cherche une voiture BMW modèle 92	أبحث عن سيارة بي إم دبليو موديل 92
مطلوب دايو لانوس	recherche DAEWOO	مطلوب دايو لانوس
مطلوب نيسان صني موديل 93 إلى 95	recherche NISSAN Sunny modèle 93 à 95 manuelle	مطلوب نيسان صني موديل 93 إلى 95 عادي
مطلوب هوندا سيفيك موديل 94	recherche HONDA Civic modèle 94 toutes options ma	مطلوب هوندا سيفيك موديل 94 في اوشن عادي
مطلوب كيا سيبيا	recherche KIA Sephia	مطلوب كيا سيبيا
مطلوب دايو لانوس 95 إلى 97	recherche DAEWOO Lanos 95 à 97 manuelle	مطلوب دايو لانوس 95 إلى 97 عادي

Figure 3 : exemple de résultats obtenus par traduction statistique (Pharaoh)

En tenant compte de la complexité et de la richesse de la langue arabe, le résultat obtenu pour cette taille est encourageant. En effet, très peu de mots sont inconnus, comme HYUNDAI.

Nous avons utilisé le même corpus d'évaluation que celui utilisé dans l'évaluation de la version originale de CATS (200 SMS, 100 d'achat et 100 de vente).

Dans ce qui suit, nous évaluons d'abord la traduction par les deux méthodes NIST et BLEU, classiquement utilisées en TA statistique. Pour mesurer l'adéquation à la tâche, nous évaluons ensuite l'extraction d'information par une mesure de rappel et de précision.

4.3.2. Évaluation de la traduction par NIST et BLEU

L'évaluation par NIST et BLEU suppose d'avoir au moins une traduction de référence et une traduction candidate pour chaque énoncé source (Papineni, Roukos et al. 2002).

Dans notre cas, la traduction de référence est le SMS arabe original, la traduction candidate est le résultat produit par le système Pharaoh, et l'énoncé source est le SMS du corpus d'évaluation en français.

Le Tableau 9 présente les différents scores NIST et BLEU obtenus pour le corpus d'évaluation en fonction de la taille du corpus d'entraînement.

On observe que ces scores n'augmentent presque plus à partir de 500 SMS, et qu'ils sont très faibles par rapport aux scores obtenables avec de très gros corpus. Mais il est possible que les résultats soient malgré cela utilisables pour en extraire une information correcte.

Taille du corpus d'entraînement	NIST	BLEU
100	3,52	0,14
200	4,23	0,20
300	4,42	0,21
400	4,64	0,23
500	4,95	0,25
600	5,00	0,25
700	5,04	0,25
800	5,05	0,26
900	5,01	0,25
1000	5,07	0,26
1100	5,05	0,26

Tableau 9 : scores NIST et BLEU en fonction de la taille du corpus d'entraînement

La courbe de la Figure 4 montre une croissance très faible du score BLEU à partir de la valeur 0,26 qui correspond à une taille du corpus d'entraînement égale à 800 SMS. À partir de cette même taille de corpus, la courbe de la Figure 5, représentant la mesure NIST, croit aussi très faiblement à partir de la valeur 5,05. Cela veut dire qu'une augmentation de la taille du corpus d'entraînement ne modifie presque pas la valeur du score BLEU.

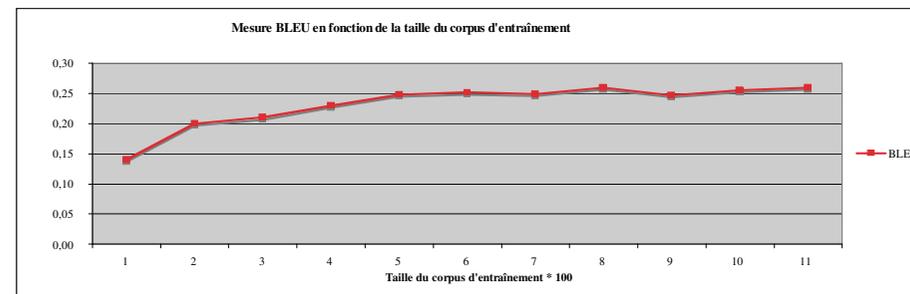


Figure 4 : BLEU en fonction de la taille du corpus d'entraînement

Nous ne garantissons pas que les mesures ne peuvent s'améliorer après une certaine augmentation de la taille du corpus d'entraînement ou l'ajout d'autres ressources et outils comme un analyseur morphologique pour le français. Mais, rappelons-le, notre objectif était de proposer des solutions de multilinguisation simples, et applicables sur le terrain avec le coût le plus faible possible. Or, un examen rapide des résultats nous indique qu'il semble n'y avoir que peu ou pas de perte d'information. Nous vérifierons ce point plus loin, en appliquant l'extracteur de contenu à ces résultats, et en comparant les CRL-CATS obtenus avec ceux obtenus à partir des SMS originaux.

Notons qu'on arrive approximativement au même score BLEU que d'autres expériences sur le couple de langues anglais—arabe. Ainsi, d'autres chercheurs de notre équipe sont arrivés à la valeur 0,25 pour BLEU, mais cela leur a demandé une taille de corpus d'environ 42000 phrases car il s'agissait d'une variante de l'arabe plus complexe et plus générale que la nôtre (dialecte irakien, dialogues informationnels) (Besacier, 2007).

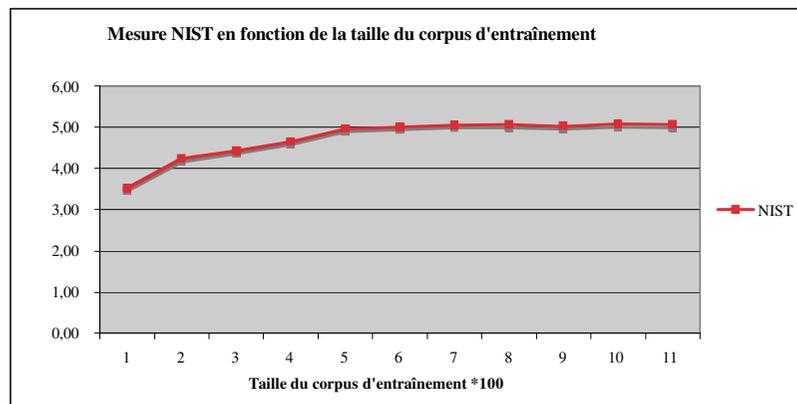


Figure 5 : score NIST en fonction de la taille du corpus d'entraînement

4.4. Évaluation de l'extraction d'information

Les résultats d'extraction de contenu de la version arabe obtenue par TA statistique des SMS français et ceux obtenus à partir de la version originale (arabe) sont regroupés dans le tableau suivant, pour les propriétés les plus importantes. Les pourcentages de portage par rapport à la version originale varient entre 85% et 98%, avec une moyenne de 93 %.

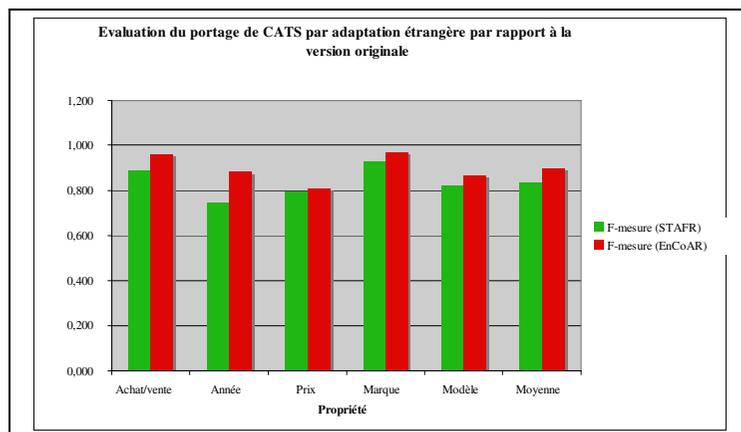


Figure 6 : comparaison entre F-mesure (par rapport à la version originale)

L'avantage de cette méthode est qu'elle ne nécessite aucun accès aux ressources de l'application-mère. La Figure 6 permet de mieux visualiser la comparaison entre les valeurs de F-mesure trouvées pour chacune des versions du système.

Propriété	SMT-FR			EnCo-AR			% portage
	Précision	Rappel	F-mesure (SMT-FR)	Précision	Rappel	F-mesure (EnCo-AR)	
Achat/vente	1,000	0,800	0,889	0,956	0,970	0,963	92
Année	0,753	0,740	0,746	0,817	0,960	0,883	85
Prix	0,883	0,726	0,797	0,800	0,822	0,811	98
Marque	0,964	0,901	0,931	0,978	0,963	0,970	96
Modèle	0,957	0,718	0,820	0,901	0,837	0,868	95
Moyenne	0,912	0,777	0,837	0,890	0,910	0,899	93

Tableau 10 : Comparaison entre les résultats d'extraction de contenu pour 200 SMS en arabe obtenus par traduction statistique, par rapport aux SMS de référence

Conclusion

Nous avons présenté une application de la traduction automatique statistique (SMT) au "portage linguistique" de l'arabe au français de CATS, un système traitant le contenu de brefs messages spontanés en langue naturelle (SMS). Il s'agit d'énoncés réels, car CATS est une application déployée sur le réseau FastLink en Jordanie. Nous avons travaillé sur la partie "occasion automobile" (Cars), où il s'agit d'un "sous-langage" très restreint.

Nous avons préalablement expérimenté deux autres méthodes, l'une demandant un accès au code de l'extracteur de contenu "natif", et l'autre consistant à adapter un extracteur de contenu du français existant. Il nous avait suffi pour cela de construire un très petit corpus parallèle, augmenté d'un dictionnaire bilingue assez complet lié à l'application choisie (petites annonces en occasion automobile), et nous nous sommes limités à ces ressources pour construire avec Pharaoh un système de TA statistique français-arabe pour des SMS en français et évaluer la faisabilité d'un portage de CATS en français par cette méthode.

Bien que la TA statistique soit réputée ne fonctionner assez bien que si l'on dispose de très grands corpus parallèles, ce système a produit des résultats satisfaisants, au sens où les descripteurs de contenu produits par l'extracteur de contenu de CATS sont très proches de ceux produits à partir des SMS de référence correspondants en arabe, en termes de rappel et de précision, alors même que les scores BLEU et NIST sont assez mauvais.

Il semble donc qu'on puisse se limiter à de très petits corpus pour utiliser efficacement la TA statistique sur des "sous-langages" très restreints, du moment qu'on a un dictionnaire bilingue assez complet : même si les traductions ne sont pas très "fluides", elles peuvent être "adéquates", même si les deux "langues-mères" des deux sous-langages considérés sont assez distantes.

On a ici une illustration de la validité de l'affirmation de Kittredge selon laquelle deux sous-langages qui se correspondent dans deux langues différentes sont très proches entre eux, et souvent plus proches entre eux qu'ils ne le sont chacun de leur langue-mère respective, ce qui permet de les considérer et de les traiter comme des variantes l'un de l'autre.

Bibliographie

Besacier, L. (2007). *Transcription enrichie de documents dans un monde multilingue et multimodal*. Grenoble, Université Joseph Fourier. HDR, 300 p.

Biber, D. (1993). *Using register-diversified corpora for general language studies (Special issue on using large corpora)*: 219-241. MIT Press Cambridge, MA, USA.

Blanchon, h. (2004). *Comment définir, mesurer et améliorer la qualité/l'utilisabilité et l'utilité des systèmes de TAO de l'écrit et de l'oral. Une bataille contre le bruit, l'ambiguïté, et le manque de contexte*. Grenoble, Université Joseph Fourier. HDR, 380 p.

Bross, I. D. J., P. A. Shapiro, et al. (1972). *How information is carried in scientific sub-languages*. *Science*, pp. 1303-1307.

Chandioux, J. (1988). *10 ans de METEO. Traduction Assistée par ordinateur. Actes du séminaire international sur la TAO et dossiers complémentaires*, OFIL, A. Abbou, ed. Paris.

Daoud, D. M. (2006). *It is necessary and possible to build (multilingual) NL-based restricted e-commerce systems with mixed sublanguage and content-oriented methods*. GETA - CLIPS. Grenoble, Université Joseph Fourier, Thèse, 296 p.

Daoud, D. M. (2005). *Building SMS-based System using Information Extraction Technology*. ACIDCA-ICMI, Tozeur, Tunisia. 7 p.

Deville, G. (1989). *Modelization of Task-oriented Utterances in a Man-Machine Dialogue System*, University of Antwerpen, Belgique. PhD, 200 p.

Grishman, R. and R. Kittredge (1986). *Analyzing language in restricted domains*, Hillsdale NJ.

Hajlaoui, N. and C. Boitet (2007). *Portage linguistique d'applications de gestion de contenu*. TOTh Conférence sur la Terminologie & Ontologie: Théories et Applications, Annecy France, 13 p.

Harris, Z. (1968). *Mathematical structures of language*. New York, Wiley-Interscience.

Kittredge, R. (1982b). *Variation and Homogeneity of Sublanguages*. in *Sublanguage - Studies of Language in Restricted Semantic Domains*. Walter de Gruyter. Berlin / New York.

Koehn, P. (2004). *Pharaoh: a Beam Search Decoder for Phrase-Based SMT*. 6th AMTA, Washington.

Kumamoto, T. (2007). *A Natural Language Dialogue System for Impression-based Music-Retrieval*. CICLING 07 (Computational Linguistics and Intelligent Text Processing), Mexico.

Och, F. J. and H. Ney (2000). *Improved statistical alignment models*. The 38th Annual Meeting of the Association for Computational Linguistics. pp. 440-447.

Papineni, K., S. Roukos, et al. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia.

Slocum, J. (1986). *How one might automatically identify and adapt to a sublanguage*. Book section « *Analyzing language in restricted domains* », pp. 195-210.

Stolcke, A. (2002). *SRILM - an Extensible Language Modeling Toolkit*. ICSLP, Denver, USA.

Sekine, S. (1994). *A new direction for sublanguage NLP*. International Conference on New Methods in Language Processing, 8 p.

La place de la modélisation sémantique dans la méthodologie d'entreprise

Dominique VAUQUIER

Senior Architect, AXA Group IS – Président du *Praxeme Institute*

21, chemin des Sapins

93160 NOISY-LE-GRAND

dvau@praxeme.org

http://www.praxeme.org

Résumé :

Dans les pratiques actuelles comme dans les méthodes (*Enterprise Architecture*, urbanisation de SI), la modélisation du métier se fait en termes de processus. Or, les processus véhiculent de nombreuses hypothèses d'organisation, des choix locaux qui rendent difficile le partage du modèle et qui limitent l'innovation. La modélisation sémantique fait abstraction des contingences organisationnelles et techniques et vise à exprimer l'essentiel du métier. Le modèle sémantique est, donc, plus facile à partager et il fournit le point de départ pour construire le cœur du système d'information. Par son effort d'abstraction, il permet aussi de repenser le métier, dans ses fondements.

Le procédé de modélisation sémantique est un apport essentiel de la méthode publique Praxeme.

Mots-clés : Modélisation, connaissances, entreprise, méthodologie, Praxeme, UML, architecture de SI, MDA

1. Introduction

1.1. Contexte académique

Praxeme est la méthodologie d'entreprise issue de l'initiative pour une méthode publique. Son ambition est de couvrir tous les aspects de l'entreprise afin d'articuler les expertises nécessaires à la transformation de l'entreprise. Praxeme, soutenue par des entreprises privées et des organismes publics (SAGEM, SMABTP, armée de terre, Caisses d'Allocations familiales, etc.) est

portée par le *Praxeme Institute*, association de loi juillet 1901, sans but lucratif, qui en garantit le caractère public et l'ouverture.

L'objectif de l'initiative pour une méthode publique est de fournir au marché une méthode de référence, à l'instar de Merise dans les années 80. Au principe d'un financement public, se substitue celui de la mutualisation des investissements entre des entreprises et organismes qui ressentent le même besoin d'une méthode de référence. Praxeme ne cherche pas à concurrencer les référentiels disponibles (par exemple : TOGAF pour l'architecture d'entreprise, UP pour le développement de logiciel), mais à les appuyer sur une base théorique et à les compléter par des procédés de modélisation précis.

1.2. Contexte industriel

Plusieurs références illustrent, à grande échelle, l'application de cette méthodologie. Plus particulièrement, le procédé de modélisation sémantique a été employé :

- pour élaborer le Référentiel Métier du groupe Celesio (répartiteur pharmaceutique),
- pour guider la refonte de l'informatique des stations de contrôle dans les systèmes de drone (SAGEM Défense),
- en préalable à la refonte du SI en SOA¹ à la SMABTP (assurance),
- pour modéliser la production chez EDF (Direction des Opérations Amont Aval Trade),
- pour clarifier la notion de « *customer centrivity* » dans le cadre de la stratégie du groupe AXA.

1.3. Contenu de l'article

Le procédé de modélisation est décrit dans le « Guide de l'aspect sémantique » de Praxeme (cf. [Vauquier 11/2006]). Nous présenterons, d'abord, le cadre général de la méthodologie d'entreprise. Puis, nous soulignerons les enjeux, pour l'entreprise, d'une approche sémantique. La modélisation

¹ SOA (*Service Oriented Architecture*) est défini par Praxeme comme un style d'architecture logique, c'est-à-dire une certaine façon de structurer les systèmes informatiques, en l'occurrence comme un mécano de composants fournissant des services. Voir le bilan du projet de refonte de la SMABTP : [Bonnet, 2007].

sémantique sera ensuite abordée sous l'angle des activités (l'approche), puis sous l'angle du produit (le modèle sémantique).

2. La Topologie du Système Entreprise

L'ambition de Praxeme est d'articuler les expertises nécessaires à l'étude et à la transformation des entreprises. La spécialisation des discours engendre fatalement l'isolement. Les expertises se succèdent, également légitimes : celle du stratège, celle de l'audit, de l'organisateur, du marketing, de l'informatique, etc. En l'absence d'un cadre global dans lequel ces expertises pourraient s'insérer et harmoniser leurs effets, nous assistons à une grande déperdition d'énergie. Plus grave encore, dans le bruit ambiant, les entreprises manquent des occasions d'amélioration ou d'innovation.

À partir de ce constat, Praxeme propose un cadre de référence qui identifie huit aspects du « Système Entreprise ». Nous nommons « Système Entreprise » l'objet complexe qu'est l'entreprise, quand elle se perçoit elle-même et s'analyse dans un mouvement d'autoréflexion. Les aspects sont constitutifs de ce système. Leur identification a reçu une justification théorique² et des preuves empiriques à travers les applications de la méthode.

Chaque aspect est un ensemble cohérent d'informations et de décisions, susceptible d'être modélisé, c'est-à-dire de recevoir une expression formelle. Tout, dans l'entreprise, ne peut pas recevoir une expression formalisée et il faut laisser un espace pour les expressions floues. C'est le rôle de la pré-modélisation que de recueillir et d'organiser ces expressions. Il s'agit des objectifs (de la stratégie aux opérations), des exigences, des vocabulaires. Cet ensemble d'éléments alimente le « cadrage » ; les travaux ultérieurs de modélisation viennent y puiser pour justifier leurs éléments et décisions de représentation. Des chaînes de traçabilité courent de ces termes de cadrage jusqu'au déploiement en passant par les éléments de modélisation répartis dans les aspects.

Recenser les aspects du Système Entreprise permet de se doter d'une grille de lecture canonique qui aide à maîtriser l'ensemble des informations et décisions. Ceci ne suffit pas. Il faut encore articuler ces informations et décisions. La Topologie du Système Entreprise est un schéma qui articule les aspects, en

² Cf. [Vauquier 10/2006].

résumant des règles précises. Ce schéma respecte le standard UML³ : les flèches doivent se comprendre comme des relations d'utilisation, des références. Ainsi l'aspect logique se construit à partir des aspects sémantique et pragmatique... Le tableau ci-dessous donne les définitions des aspects⁴.

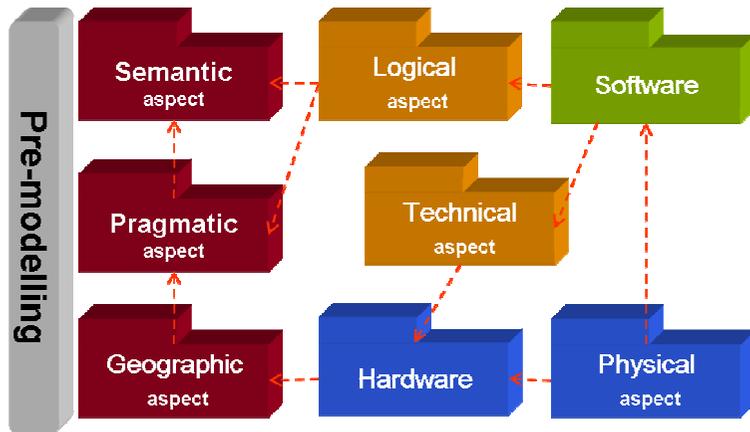


Fig. 1 : La Topologie du Système Entreprise, fondement de la méthodologie Praxeme

Aspect	Termes équivalents	Définitions
Sémantique	Conceptuel, essentiel, « Cœur de métier »	L'aspect sémantique ne retient que les objets au cœur de l'activité. On décrit le noyau fondamental indépendant de la manière de mener l'activité.
Pragmatique	Organisationnel	L'aspect pragmatique réunit les choix relatifs à la manière de mener l'activité : acteurs, responsabilités, actions sur les objets, processus situations de travail.
Géographique	«Communication», «Contexte»	L'aspect géographique est celui de la localisation des objets et des actions. Il fait apparaître les notions de sites, d'emplacements et de besoins de communication.
Logique	«Fonctionnel»	Aspect intermédiaire permettant de fixer les grandes décisions de structuration du système d'information, indépendamment des choix techniques.
Technique	Technologique	L'aspect technique est celui des choix de technologies et des façons de les mettre en œuvre.
Matériel	Logistique	L'aspect matériel du système est l'ensemble des machines physiques composant le système, avec leurs propriétés (capacité...).
Logiciel	Applicatif, informatique	L'aspect logiciel couvre l'ensemble des composants logiciels qui automatisent une partie des actions du système.
Physique	Déploiement	À travers l'aspect physique, on décrit la localisation des composants logiciels (bases de données comprises) sur les matériels.

Fig. 2 : La définition des aspects du Système Entreprise

³ UML : *Unified Modelling Language*. Standard de l'*Object Management Group*.

⁴ Identifier des types de modèles et les articuler, c'est l'esprit du standard MDA. MDA, *Model Driven Architecture*, est un autre standard de l'OMG qui réactive l'idée selon laquelle il nous faut plusieurs modèles et que ces modèles se lient les uns aux autres par dérivation.

L'aspect sémantique est donc le premier à apparaître dans l'effort de formalisation appliqué à la réalité de l'entreprise.

3. La modélisation sémantique et ses enjeux

3.1. Définitions

L'**interopérabilité sémantique** est la capacité, pour les systèmes sociotechniques, d'échanger de l'information porteuse de signification. La signification d'une information se conforme à l'usage qu'en font les acteurs du système (le métier). L'interopérabilité suppose que l'intégrité des informations est parfaitement assurée.

Le **modèle sémantique** exprime, formellement, les fondamentaux du métier. Ceci exclut les choix d'organisation qui se montrent dans les processus. Ce modèle n'est pas un modèle de l'informatique : il représente la connaissance du métier et, plus précisément, à l'intérieur de cette connaissance, la part invariante, indépendante des pratiques et des modes opératoires. Le modèle sémantique a donc vocation à l'universel.

L'**orientation client** (*customer centrivity*) est un élément de la culture et des pratiques de l'entreprise qui l'amène à considérer prioritairement le client.

On peut distinguer deux interprétations de l'orientation client :

1. La première, classique, consiste à placer le client au centre de la perspective et à exploiter au maximum l'information que l'entreprise possède sur lui (voir figure ci-dessous).
2. La seconde interprétation, radicale, consiste à adopter le point de vue du client, plutôt que celui de l'entreprise. Elle entraîne de grands changements, à commencer par la désignation de l'acteur central lui-même : en effet, le « client » ne se vit pas comme client, mais comme personne.

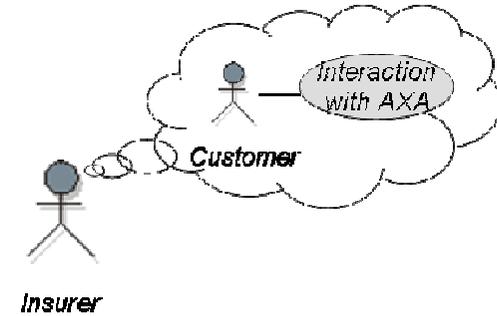


Fig. 3 : Première interprétation de l'orientation client

Parler du « client », c'est situer la personne uniquement ou principalement dans son rapport à l'entreprise qui le fournit. C'est donc occulter une partie de la réalité :

- celle de la personne dans ses relations avec d'autres fournisseurs ou dans ses aspirations ;
- celle des personnes qui ne sont pas clients, qui sont ou pourraient être en interaction avec l'entreprise.

Nous évoquons, ici, ces approches culturelles car leur influence se lira directement dans le modèle sémantique. L'approche retenue colore le modèle : ce n'est pas seulement le choix des termes qui est en jeu, mais surtout la structuration du modèle. Ce point est illustré dans la section 5.

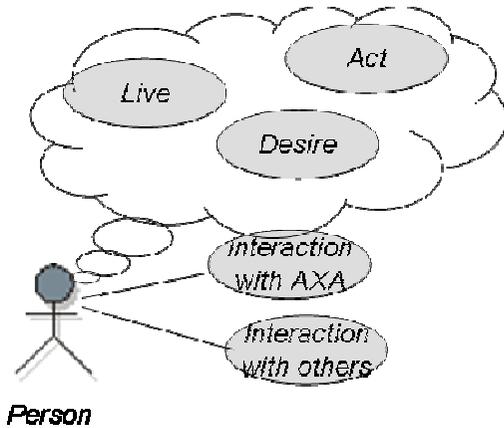


Fig. 4 : Deuxième interprétation de l'orientation client

3.2. Retombées

Le modèle sémantique, étendu à l'entreprise, devient un instrument de communication et d'action. Ses retombées sont multiples et touchent le métier autant que l'informatique.

Compréhension commune : Le modèle sémantique s'abstrait des variantes liées aux choix d'organisation ou aux solutions informatiques. De cette façon, il donne une perception simplifiée, en disant l'essentiel. Il recueille la connaissance du métier et en donne une forme suffisamment rigoureuse pour se plier à plusieurs exploitations, notamment en vue de l'informatisation. La connaissance ainsi capitalisée peut être partagée entre les acteurs du métier (nouveaux employés, différentes compagnies...) ou entre ces acteurs et les informaticiens.

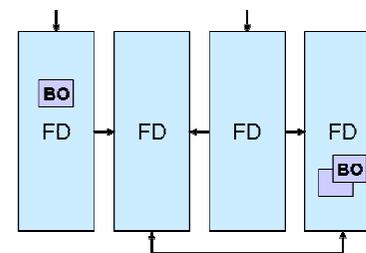
Perception enrichie : La modélisation sémantique a pour but d'exprimer toute la sémantique attachée aux objets et concepts du métier. Elle ne se borne pas aux données. Notamment, elle prend en charge des informations dérivées et des calculs tels que les indicateurs pour le pilotage ou le marketing. Sous l'unité de l'objet « métier », le modèle sémantique intègre donc des dimensions habituellement disjointes. Il enrichit la perception que nous avons des objets et notions du métier et sert de base pour mieux articuler les différentes solutions.

Partage et consolidation des données : De par son niveau d'abstraction, le modèle sémantique aide au partage de la connaissance métier ainsi que des

solutions informatiques. Des éléments de variabilité peuvent tout de même toucher ce modèle : réglementations nationales qui contraignent le comportement de certains objets, caractéristiques spécifiques valables dans certaines compagnies... Le procédé de modélisation comporte des dispositions pour isoler ces facteurs de variation et préserver le caractère générique du modèle. Ainsi, le modèle sémantique fournit le point de départ pour élaborer une description des données qui pourra être partagée par plusieurs systèmes, sous la forme d'un « langage pivot ». Ce langage unique est une des conditions pour échanger et consolider des données.

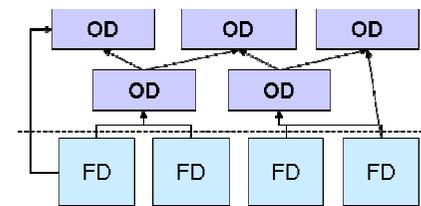
Structuration des systèmes : S'il est bien intégré à la chaîne de construction des systèmes d'information, le modèle sémantique a un impact significatif sur leur structure. À l'échelle de l'entreprise, il est nécessaire de décomposer le modèle sémantique. Pour cette décomposition, nous adoptons un autre critère que celui des domaines fonctionnels. La modélisation sémantique introduit la notion de « domaines d'objets » centrés sur les principaux objets du métier. Cette pratique conduit à revisiter de fond en comble la structure optimale des systèmes d'information. Elle permet de dégager un noyau stable, peuplé de services à fort contenu et hautement réutilisables.

Fig. 5 : La structure en silos (les objets métier sont perdus dans les domaines fonctionnels, d'où redondance ou couplage)



FD = *functional domain*
BO = *Business object*

Fig. 6 : La nouvelle physionomie des SI (les domaines d'objets permettent de factoriser les services sur les objets métier et de les mettre en commun)



OD = *objects domain*

Le changement d'approche : Quand il s'agit de données, les modèles existants sont surtout de nature logique ou physique. La pauvreté des moyens d'expression (limitation aux données ; associations binaires, essentiellement) convient pour les modèles logiques de données (entités-relations ou tabulaire). Cette pratique s'inscrit dans l'approche traditionnelle par projets, approche qui a conduit à la réalisation des silos applicatifs. Or, aujourd'hui, nous nous devons de penser à un modèle qui soit transverse et valable pour toute l'entreprise. Cette ambition nécessite un changement d'approche et une révision de nos pratiques. Notamment, les nouveaux besoins exprimés réclament un lien plus direct entre les solutions pour le marketing et le décisionnel, d'un côté, et le fonctionnement en mode transactionnel, de l'autre.

Par ailleurs, le changement d'approche se caractérise par l'adoption de la logique « objet ». Cette nouvelle perception du réel, si elle est intériorisée, modifie en profondeur les modèles et finit par changer la physionomie des systèmes. Ces changements se jouent à un niveau profond dans la culture des modélisateurs, des architectes et des informaticiens. Il s'agit d'une réelle transition méthodologique qui doit surmonter bien des résistances culturelles.

3.3. Positionnement

Le modèle sémantique : Les méthodologies actuellement bien implantées négligent souvent ce niveau d'abstraction. Que ce soit dans le courant *Enterprise Architecture* ou dans les processus de développement (RUP, UP...), la perception la plus élevée que l'on donne du métier s'exprime en termes d'activités : processus, cas d'utilisation, fonctions⁵. Le modèle sémantique se positionne en amont de cette approche fonctionnaliste. Il fait abstraction des acteurs et des pratiques pour exprimer l'essentiel du métier, en termes d'objets. Cette position en amont de la chaîne fait que le modèle sémantique peut servir de point de départ pour d'autres activités de réflexion ou de transformation de l'entreprise et de ses systèmes.

Les processus : Le modèle sémantique associe aux principaux objets « métier », un automate à états qui formalise leur cycle de vie. Dès lors, la conception des processus peut adopter un procédé innovant : plutôt que de partir de la description des activités – en prenant le risque de rester collé à l'existant – le

concepteur de processus ou l'organisateur peut simplement considérer que le meilleur des processus est celui qui se « contente » d'accompagner le cycle de vie de l'objet. C'est une utilisation possible du modèle sémantique. Il fournit un autre angle d'attaque pour une conception innovante des processus.

Le « langage pivot » : Le modèle sémantique n'est pas un modèle de données : il dit plus que cela, en exprimant toute la sémantique attachée aux objets et concepts du métier. Bien sûr, il incorpore un modèle conceptuel des données, puisqu'il décrit les informations portées par les objets. En conséquence, il est facile d'en dériver des structures de données. De ce point de vue, la méthode établit deux filières de dérivation à partir du modèle sémantique :

- l'une, pour le modèle logique des données qui prépare la conception des bases de données ;
- l'autre, pour les structures de données échangées au sein des systèmes.

Cette deuxième filière de dérivation revêt une importance particulière dans une perspective d'interopérabilité. Elle produit le « langage pivot » dans lequel les échanges de données devront s'exprimer. Le langage pivot est spécifié au niveau logique, en utilisant par exemple les *data types* de la notation standard UML. Le format retenu est important puisque cette spécification devra être incorporée dans plusieurs outils, sur les projets. Il reste ensuite à établir le support technique dans lequel cette spécification logique va se concrétiser.

Les services SOA : Une autre filière de dérivation s'accroche au modèle sémantique : celle qui permet d'identifier et de structurer une partie des services, au sens SOA. La dérivation du modèle sémantique permet de peupler la strate « Métier » de l'architecture logique. Dans cette démarche, le concepteur découvre les services « par en haut », c'est-à-dire par le métier et, même, par ce qu'il y a de plus invariant dans le métier. Ces services offrent donc un fort contenu fonctionnel et se révèlent hautement réutilisables. SOA est un style d'architecture logique. Dans le détail, ce style fait l'objet d'une négociation entre :

- l'architecture logique, soucieuse de la bonne structure du système ;
- l'architecture technique, garante de la convertibilité du modèle logique en logiciel.

L'approche SOA porte tous ses fruits quand elle se situe dans le cadre complet, de l'amont à l'aval. Indépendamment des autres enjeux indiqués ici, la modélisation sémantique compte parmi les principales conditions de succès des

⁵ Dans TOGAF, le premier plan de représentation est la *Business Architecture*, parfois plus explicitement nommée *Business Process Architecture*.

projets SOA. La portée du modèle sémantique, conçu à l'échelle de l'entreprise, augmente la réutilisabilité des services dérivés.

4. L'approche

Cette section évoque les actions clefs qui impliquent le modèle sémantique⁶.

4.1. La collecte

Afin de préparer le modèle, le justifier et lui donner rapidement une substance significative, deux types d'entrées se présentent :

- les modèles de données existant dans les compagnies ;
- les modèles et standards présents sur le marché.

Ces entrées devraient être incorporées dans des dictionnaires et tracées vers les termes du modèle final, de sorte que :

- les éléments du modèle sémantique soient « justifiés » par l'existant ou les standards du marché ;
- les correspondances préparent les conversions de données.

4.2. La pré-modélisation

La pré-modélisation prolonge la collecte. Elle recueille des éléments d'information, des exigences, des objectifs... tout ce qui a de la valeur mais ne peut pas s'exprimer dans les termes formels d'un modèle. En ce qui concerne le modèle sémantique, ce sont les « vocabulaires » qui sont au cœur de nos préoccupations. Pour un terme « normalisé » retenu par le modèle, il peut y avoir plusieurs termes en usage (sans compter le multilinguisme). De plus, des concepts représentés par des entités ou des tables dans l'existant, peuvent se trouver restitués dans le modèle par des moyens plus subtils. Il est donc important, pour faciliter l'exploitation du modèle, d'établir des dispositifs qui permettront de passer des vocabulaires vers le modèle. Le thesaurus est ce dispositif. Il sert de sas d'entrée entre la perception ordinaire et le modèle. C'est donc un instrument de communication essentiel pour assurer la réception du modèle.

⁶ Le plan du chantier (référence « CIS-02 ») détaille les actions et la démarche.

4.3. La structuration du modèle

Une des premières décisions fortes sur le modèle sémantique consiste à le décomposer en domaines d'objets. Il s'agit d'un acte d'architecture lourd de conséquences puisqu'il déterminera la physionomie des systèmes futurs ou, au moins, les orientations d'évolution. Cet acte doit, donc, impliquer plusieurs personnes et faire l'objet d'une large validation. La décomposition en domaines d'objets devrait être transportée dans l'architecture des systèmes informatiques (à travers l'architecture logique). C'est un point qui devrait être examiné lors des grandes étapes de l'évolution d'un SI (*appropriation request, roadmap*).

4.4. La modélisation sémantique

L'activité de modélisation sémantique peut se dérouler selon plusieurs modes :

- mode activité permanente : c'est un travail de consolidation de longue haleine ;
- mode projet : les projets fournissent les occasions pour apporter les derniers détails à une portion du modèle de référence et, réciproquement, pour diffuser l'approche et les éléments qui ont été capitalisés dans le modèle commun.

Le mariage entre ces deux modes permet d'avancer en combinant les deux approches *top-down* et *bottom-up*. L'approche *top-down* de la modélisation sémantique est le chemin de l'innovation : elle permet de revisiter de fond en comble la définition du métier. L'approche *bottom-up* garantit le réalisme nécessaire pour mobiliser les ressources et concrétiser le modèle.

4.5. L'exploitation du modèle

Il existe plusieurs façons – non exclusives – d'exploiter un modèle sémantique, particulièrement quand il prend l'ampleur d'un Référentiel Entreprise :

- En tant qu'expression formelle des fondamentaux du métier, il relève de la gestion des connaissances.
- Puisqu'il exprime les cycles de vie des objets « métier », le concepteur de processus peut l'utiliser comme un nouveau point de départ, dans une approche innovante des processus.

- La dérivation du modèle sémantique vers le modèle logique guide une partie des décisions de l'architecte logique et aide à découvrir des services à fort contenu (SOA).
- D'autres dérivations produisent le « langage pivot » et les modèles logiques de données.

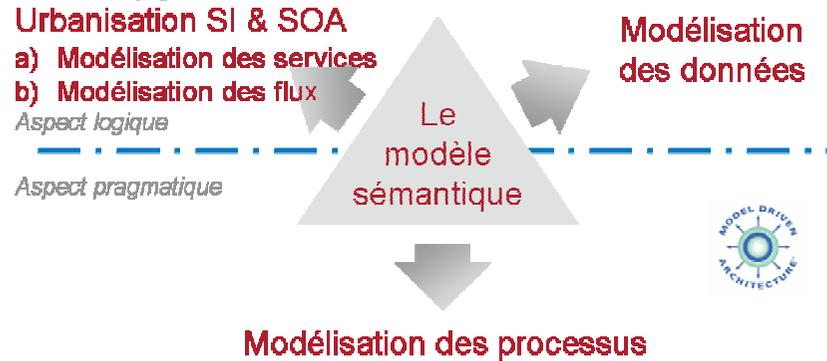


Fig. 6 : Récapitulatif des filières de dérivation à partir du modèle sémantique

4.6. La lecture du modèle

Le modèle sémantique peut être conçu à destination de plusieurs types d'acteurs. Il faut alors veiller aux différents niveaux de lecture et à la production de « vues » différenciées. Notamment, il intéresse :

- les acteurs du « Business » (au premier titre, la direction Marketing) ;
- les organisateurs et concepteurs de processus ;
- les stratèges (qui retrouvent dans le modèle sémantique les objectifs de marché, inscrits sous la forme d'objets pertinents qui portent les indicateurs) ;
- l'informatique, qui, dans son souci d'alignement de l'informatique sur le métier, exploite le modèle sémantique pour construire le noyau applicatif, stable et partageable.

4.7. La cohabitation

C'est une chose d'exprimer les connaissances du métier en faisant abstraction de l'organisation et des solutions informatiques ; c'en est une autre de

concilier cette représentation avec les systèmes existants. En outre, l'établissement d'un Référentiel Entreprise est un chantier qui s'inscrit dans la longue durée. Dès lors, se pose la question de la cohabitation entre la vision classique (celle des systèmes existants) et la vision nouvelle. SOA, justement, est une approche progressive de la refonte des systèmes. Ses procédés fournissent des réponses opératoires pour la cohabitation avec l'existant ou les ERP.

5. Le modèle sémantique

5.1. Nature

Le modèle sémantique décrit les fondamentaux du métier, abstraction faite des contingences organisationnelles et techniques. Il se situe sur l'aspect le plus en amont dans la succession des modèles, surplombant même les modèles de processus. Ceci le dote d'un avantage énorme : en écartant les variations dues à l'organisation et aux diverses adaptations, le modélisateur a plus de chance de dégager l'essentiel. Le modèle sémantique est donc plus simple et a vocation à l'universel.

Pour éviter les travers connus de l'approche fonctionnelle, le modèle sémantique ne décrit pas l'entreprise en tant qu'ensemble d'activités, mais à travers les objets et les notions qu'elle manipule. La modélisation sémantique adopte, donc, une approche orientée objets, en veillant à préserver le pouvoir d'expression de ses outils.

Le modèle sémantique n'est pas un modèle du logiciel ou du système informatique. C'est une représentation formelle des connaissances fondamentales d'un métier.

5.2. Procédé

La modélisation sémantique est un procédé de représentation formelle qui véhicule plusieurs exigences.

Un procédé de modélisation
appliqué à l'aspect « sémantique »

La modélisation sémantique a pour objectif de décrire, rigoureusement, les fondamentaux du métier

Sous forme d'un modèle « exécutable »

abstraction faite des contingences organisationnelles et techniques

Fig. 7 : Définition de la modélisation sémantique

Dire que la modélisation ne prend en charge que l'aspect sémantique, c'est évacuer toutes les considérations des autres aspects qui risqueraient d'alourdir le modèle et de réduire son domaine d'application. Le modélisateur recherche les solutions les plus génériques. Ceci le conduit à opter pour des solutions parfois très différentes des pratiques courantes ou des systèmes existants.

5.3. Illustration

La comparaison entre les deux diagrammes de classes, ci-dessous, révèle l'impact de l'approche retenue sur la structure du modèle.

L'approche classique de la modélisation des données

Le premier diagramme ci-dessous résulte d'une approche classique de modélisation. Cette approche se caractérise par :

- les moyens d'expression, réduits à des associations binaires ;
- la perception focalisée sur les activités et le vocabulaire des acteurs « métier ».

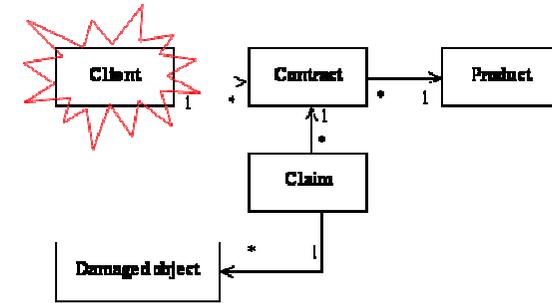


Fig. 8 : Approche classique de la modélisation des données, illustration de la première interprétation de l'orientation client

Les conséquences sont :

Le masquage des dépendances entre concepts : Par exemple, la notion de contrat est purement relative : un contrat n'existe que comme une relation entre un client et un produit. Cette nature relative du contrat n'apparaît pas instantanément, dans cette représentation.

Une perte de la sémantique : Beaucoup d'associations ne sont pas nommées. Quand elles le sont, la réduction des relations conceptuelles à des associations binaires rend difficile la recherche d'un nom expressif. Beaucoup d'associations sont nommées – quand elles le sont – par des expressions telles que « gérer... », « concerne... », « avoir », etc. Un modèle de cette sorte est trop peu expressif et échoue à capturer toute la sémantique du métier.

La focalisation sur des artifices plutôt que sur les objets réels : C'est le cas, ici, avec la classe Client. Le client n'est pas un objet réel mais un rôle que joue une *personne* dans la relation avec un fournisseur. Nous risquons de multiplier ces notions artificielles qui conduisent à compliquer le système et à dupliquer l'information. La même personne peut être, à la fois, client, salarié, responsable, etc. Dans cette approche, l'information de la personne risque d'être enregistrée plusieurs fois, peut-être sans liens apparents.

L'approche de la modélisation sémantique

La modélisation sémantique exploite toutes les possibilités d'expression de la notation UML. Bien utilisé, ce langage de modélisation devient un bon outil pour

exprimer formellement la connaissance du métier. Le diagramme de classes, ci-dessous, illustre une partie des changements :

- recours à des types d'associations plus sophistiqués, permettant de révéler les relations conceptuelles ;
- meilleure capture du vocabulaire métier.

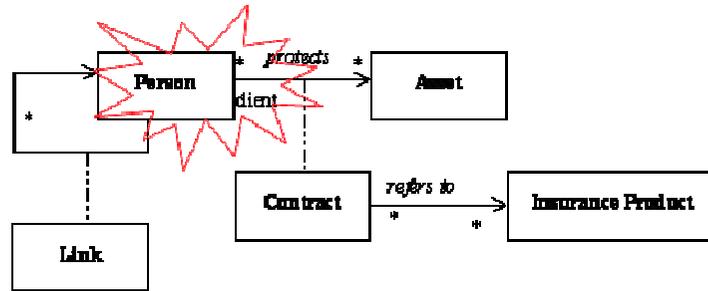


Fig. 9 : Approche de la modélisation sémantique, illustration de la deuxième interprétation de l'orientation client

Commentaire du diagramme

Dans cette version, le modèle est centré sur la classe Personne. Il restitue la notion de client sous la forme d'un rôle sur l'association « *protects* ». Cette solution est beaucoup moins coûteuse que la précédente, si l'on considère le déploiement dans le système informatique et la maintenance.

Le contrat est représenté par une classe associative, c'est-à-dire une classe attachée à une association. Ce moyen exprime la nature relative du contrat : il n'existe que pour un couple (une personne, un asset).

Le concept de personne recèle une sémantique riche. Le diagramme montre comment sont traités les liens entre personnes.

L'exemple du « *client dashboard* »

Les tableaux de bord affichent des données consolidées ou statistiques. Par exemple :

- nombre de nouveaux clients ;

- nombre moyen de polices par client ;
- évolution du portefeuille des clients « *high value* »...

Dans une approche classique, de telles informations sont considérées à part, à l'occasion de projets dédiés au pilotage. Cette façon de faire contredit l'exigence d'une meilleure intégration avec l'informatique transactionnelle, exigence exprimée par la Direction Marketing. Elle introduit une cassure dans les systèmes.

Pourtant, ces informations appartiennent à la sémantique des objets « métier », même si elles résultent d'algorithmes de navigation et de calcul. Le modèle sémantique les absorbe. Dans certains cas, le modélisateur recourt à un détail de la notation : les propriétés de portée « classe ». Le nombre total de personnes, par exemple, est un attribut de portée classe, inscrit sur la classe Personne⁷.

Ainsi le modèle dit tout de la sémantique attachée aux objets « métier ». Les règles de dérivation vers le modèle logique de style SOA prévoient la reprise de ces propriétés par des services attachés à des machines logiques dites ensemblistes. En conséquence, les informations statistiques ne sont plus traitées comme un système à part, mais intégrées, au moins au niveau logique.

5.4. Les automates à états

Les cycles de vie des objets constituent une des sources majeures de complication dans les systèmes classiques. Ils obligent, en effet, à alourdir les programmes avec de longues structures conditionnelles, souvent dupliquées à travers les systèmes.

L'approche objet nous dote d'un outil parfait pour ce type de besoin : les automates à états (ou machines à états), représentés en UML par le diagramme d'états (cf. [Desfray]).

Prenons l'exemple de l'objet (au sens de l'objet matériel). Très souvent, on trouve dans les systèmes une table pour les objets assurés, une autre pour les objets endommagés. Or, un objet est toujours un objet et est décrit par les mêmes données, quel que soit son état. La modélisation sémantique reprend les états (souvent exprimés par des adjectifs) dans un automate : assuré, déclaré,

⁷ Dans les cas plus difficiles, la propriété s'exprime sous la forme d'une opération, ce qui permet d'introduire des paramètres.

évalué, expertisé, endommagé, réparé, etc. Entre ces états, il y a des transitions valides et des relations qui peuvent être complexes (concurrence, synchronisation...).

Il est important d'y réfléchir en ne considérant que le métier. L'automate à états exprime le cycle de vie de l'objet métier, bien mieux qu'une suite de conditions. Il pourra être traduit, mécaniquement, en termes informatiques : les solutions de dérivation des automates sont connues⁸.

5.5. Structure

À l'échelle de l'entreprise, le modèle sémantique devient le Référentiel Entreprise ou Référentiel Métier. Il en couvre tous les domaines et considère tous les objets. Il comporte plusieurs centaines de classes. Il est donc nécessaire de le décomposer.

Dès qu'il est question de décomposer, l'architecte se doit d'accorder une attention particulière au critère qu'il va utiliser. Le critère du domaine fonctionnel s'applique à l'aspect pragmatique (organisationnel) mais ne convient pas pour l'aspect sémantique. En effet, les mêmes objets « métier » sont impliqués dans plusieurs activités, dans différents domaines fonctionnels. Il est nécessaire d'introduire une nouvelle notion, propre à l'aspect sémantique, pour bien « ranger » les objets.

La décomposition du modèle sémantique se fait à base de « domaines d'objets ». Un domaine d'objets s'obtient par voisinage étendu autour d'un objet principal. Il y a une dizaine d'objets principaux, permettant de couvrir l'activité d'une entreprise. Dans notre cas : objet réel, personne, produit, contrat, sinistre... Autour de ces objets se dessinent les domaines d'objets. La règle est qu'un objet, une information, une action sur un objet, une transition d'état... sont localisés à un et un seul endroit du modèle et exprimés sous une seule forme.

La figure ci-dessous est un exemple de décomposition du modèle sémantique en domaines d'objets. On notera, par exemple, le domaine « Réalité » qui contient la sémantique des objets extérieurs à l'entreprise. C'est là que se joue, en partie, l'orientation client.

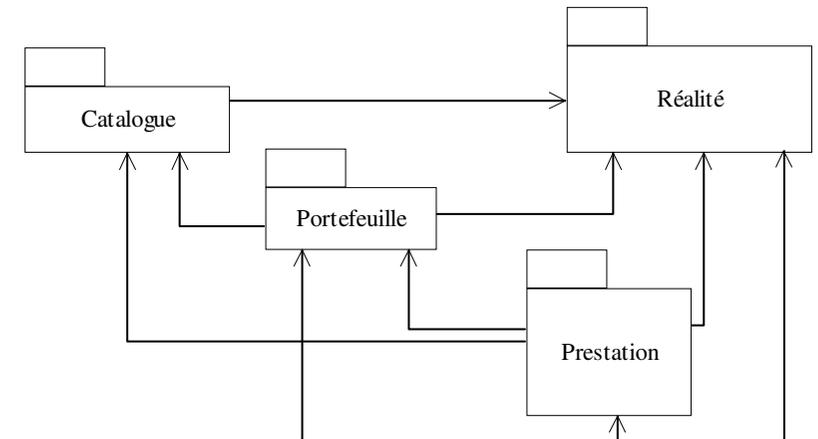


Fig. 10 : Exemple de décomposition en domaines d'objets

Sur ce diagramme en notation UML, les domaines d'objets sont représentés par des paquetages⁹. Les flèches expriment les dépendances entre ces domaines. Elles sont nombreuses. Dans le modèle *sémantique*, ce niveau de couplage peut être toléré, jusqu'à un certain point, car le but est de représenter la connaissance. En revanche, le graphe d'architecture *logique* reprendra les domaines d'objets sous la forme de constituants logiques mais réduira le couplage.

5.6. Qualité du modèle

Un modèle sémantique obéit à plusieurs exigences qui en font un outil exploitable. Notamment :

- Il ne se borne pas à la dimension statique (les données). Il dit tout des objets : leurs informations (dont les données calculées), leurs actions, leurs transformations (cycle de vie des objets).
- Certaines représentations peuvent être synthétiques, mais la modélisation sémantique n'est terminée que lorsque tous les détails ont été formulés, tout le savoir exprimé. « Sémantique » (ou

⁸ Il en existe six. C'est un point de la négociation logique/technique qui se joue entre l'architecte logique et l'architecte technique.

⁹ Le paquetage est un mécanisme, en UML, pour ordonner des éléments de modélisation.

« conceptuel ») ne signifie pas « général » comme opposé à « détaillé ». Par exemple, selon l'effort consenti, les opérations du modèle pourront présenter une signature parfaitement définie et un algorithme formellement exprimé.

- Un des préceptes de la modélisation sémantique garantit l'économie du modèle : un même terme ne doit apparaître qu'une seule fois¹⁰.
- La discipline impose l'encapsulation des règles de gestion. La structure du modèle n'est stabilisée que quand les règles de gestion et les contraintes ont été incorporées. Cette exigence conduit à ajuster la structure du modèle. Elle fournit aussi un des moyens pour dégager les opérations à valeur sémantique.

6. Conclusion

La modélisation sémantique est une technique de représentation formelle pour capturer et conserver la connaissance des fondamentaux du métier. Elle répond à des enjeux importants pour l'entreprise, à la fois pour protéger son patrimoine intellectuel et pour revisiter ses pratiques.

La méthode publique Praxeme propose :

- un cadre global qui situe cette approche parmi les autres représentations de l'entreprise ;
- un procédé détaillé de modélisation sémantique.

Son exigence et sa rigueur excèdent les pratiques et les compétences actuellement disponibles. La communauté Praxeme espère un retour sur les fondamentaux de la modélisation afin de préparer les compétences dont les entreprises ont besoin pour mener correctement leurs projets. Le monde de l'enseignement a un rôle éminent à jouer dans la restauration des compétences de modélisation.

La méthode Praxeme se veut ouverte. Pour la modélisation des connaissances, elle promeut une approche orientée objets et outillée par le standard UML. Cette approche, appuyée sur le standard MDA, facilite l'exploitation du modèle sémantique, notamment pour le développement logiciel.

Toutefois, il est tout à fait envisageable de la compléter par d'autres contributions (ontologie, approche par agents...), tant que celles-ci se conforment au cadre général de Topologie du Système Entreprise.

Bibliographie

Bonnet P., Detavernier J.-M., Vauquier D., *Le système d'information durable*, Hermes, Novembre 2007.

Vauquier D., *Guide général*, www.praxeme.org, octobre 2006.

Vauquier D., *Guide de l'aspect sémantique*, www.praxeme.org, novembre 2006.

Philippe Desfray, *Object Engineering, The Fourth Dimension*, 1994, Addison & Wesley

¹⁰ Lors d'un projet international, pas moins de 27 champs avaient été identifiés pour stocker des prix. Le modèle sémantique ne retint qu'un seul attribut et révéla les combinaisons de structure qui permettent de définir tous les prix.

Modélisation des connaissances métiers du domaine de la génétique humaine en situation d'aménagement terminologique

Josée Di Spaldro, Pierre Auger,
Maryvonne Holzem, Jacques Ladouceur

Résumé :

L'aménagement terminologique visant les milieux de travail autorise la modélisation des connaissances métiers. C'est ce que nous nous efforcerons de démontrer dans le cadre de cette communication qui porte sur un point crucial de cet aménagement, l'implantation de la terminologie française, ici du domaine de la génétique humaine. Cet article met en jeu deux aspects de l'axe terminolinguistique : la théorisation heuristique de l'élaboration d'une méthode assistée par ordinateur contribuant à la francisation de la terminologie, qui s'adresse aux praticiens au sein d'une sphère d'activité de métiers et de professions, à travers leur démarche de communication, dans le dessein de transmettre leurs savoirs ou connaissances; l'application d'une partie de la méthode auprès de sept catégories socioprofessionnelles de spectre étendu, du technicien au scientifique, au nombre de 14 participants. Première étape vers une démarche contrastive France/Québec des mêmes sphères d'activité, dans le cadre d'une enquête socioterminologique que nous menons conformément à notre projet de doctorat.

Mots-clés : français, génétique, implantation, ordinateur, terminologie, utilisateur

1. Introduction

S'il est vrai que les progrès techniques et scientifiques dans les pays industrialisés et en voie d'industrialisation ont progressé au 19^e siècle et explosé au 20^e siècle, la science de la génétique, notamment une de ses branches, la génomique aurait atteint une certaine stabilité en ce début de 21^e siècle [Gibson et Muse, 2004 : VII].

Mais qu'en est-il des termes utilisés par les praticiens du domaine pour exprimer les connaissances techniques et scientifiques lorsqu'ils communiquent

entre eux, avec l'extérieur? Cette « stabilité » épistémologique a-t-elle transcendé les notions et dénominations utilisées?

Il semble que non. La mondialisation des marchés, la langue anglaise dans laquelle elle s'investit, et la frénésie qu'elle suscite contribueraient à gêner l'implantation des terminologies françaises en milieu de travail.

Le volet de notre enquête menée en France montre une anglicisation non négligeable du milieu de travail des praticiens : 1) utilisation d'emprunts intégraux « homeobox », « gap », « heat shot », « HPLC » pour *high-pressure liquid chromatography*, « LTR » pour *long terminal repeat*, « nick translation », « PCR » pour *polymerase chain reaction*, « polylinker », « random priming », « siRNA » pour *small interfering RNA* « western blot »; utilisation d'emprunts hybrides « activation bystander », « blaster », « clustering hiérarchique », « méthode SAGE » pour *serial analysis of gene expression*, « virus helper »; 2) praticiens qui songent à effectuer les rencontres d'équipes hebdomadaires en anglais; 3) chercheur-professeur qui déplore le devoir exiger de ses étudiants qu'ils sachent et parlent l'anglais, aux fins de communications ou de publications.

Dans le but de démontrer que l'aménagement terminologique visant les milieux de travail autorise la modélisation des connaissances métiers, cet article s'attachera à exposer un plan d'aménagement terminologique, celui du Québec, les études sur l'implantation des terminologies françaises qui s'en sont suivies et la problématique qui en est ressortie. Nous présenterons ensuite l'axe terminolinguistique qui devrait rendre possible cette démonstration, par le biais de la théorisation heuristique de l'élaboration—d'une méthode assistée par ordinateur contribuant à la francisation de la terminologie, qui s'adresse aux praticiens au sein d'une sphère d'activité de métiers et de professions, à travers leur démarche de communication, dans le dessein de transmettre leurs savoirs ou connaissances; et de l'application d'une partie de la méthode auprès de sept catégories socioprofessionnelles de spectre étendu, du technicien au scientifique, au nombre de 14 participants. Première étape vers une démarche contrastive France/Québec des mêmes sphères d'activité, dans le cadre d'une enquête socioterminologique que nous menons conformément à notre projet de doctorat. Subséquemment, nous exposerons les résultats attendus, conclurons et évoquerons les perspectives d'avenir.

2. Aménagement terminologique

En 1961, sous la pression des intellectuels, et de citoyens, un Office de la langue française est institué par l'Assemblée nationale du Québec. Son instauration est attribuable à la situation diglossique qui sévit, depuis la conquête des Britanniques en 1760, dans le bassin francophone le plus considérable du Canada, la province de Québec (anciennement la Nouvelle-France). Il aura pour mandat de « veiller, sous la direction du ministre, à la correction et l'enrichissement de la langue parlée et écrite » (Loi créant le ministère des Affaires culturelles, art. 14). Pour lui donner plus de pouvoir, la loi 63 est promulguée en 1969. Elle sera enrichie et remplacée par la loi 22 (1974), puis par la loi 101 (1977).

À partir de la version actuelle de la loi 101 ou *Charte de la langue française*¹, nous avons observé qu'en matière de terminologie, l'OQLF a pour obligations : la conduite de l'officialisation de la langue, de la terminologie et de la francisation des entreprises, entre autres (art. 159); la conduite de programmes de francisation dans les entreprises, dont la nécessité d'utiliser une terminologie française (art. 141.6); la possibilité d'instituer des comités linguistiques pour relever les lacunes terminologiques ou les termes et expressions problématiques, et les corriger par les mécanismes d'intervention de la normalisation ou de la recommandation (art. 116); la publication des termes recommandés ou normalisés dans la *Gazette officielle du Québec* (art. 116.1); la nécessité d'assurer le respect de la présente loi (art. 159), en ce qui concerne, notamment, l'obligation d'employer les termes et expressions normalisés dans des circonstances précises (art. 118); la nécessité de s'assurer, entre autres, que la langue normale et habituelle du travail soit le français (art. 161). En outre, mentionnons les rôles que jouent le Comité d'officialisation linguistique et le Comité de suivi de la situation linguistique qui permettent à l'Office de remplir son mandat d'officialisation, en soumettant des propositions et des avis à ce dernier (art. 165.11).

Bien qu'on ait introduit la composante sociolinguistique dans le plan d'aménagement terminologique québécois dès le début des années 1970, l'approche de la théorie wüsterienne préconisée, théorie classique de la terminologie qui admet une relation de biunivocité entre notion et

¹ Charte de la langue française. *L.R.Q., c. C-11*, [Québec] : Éditeur officiel du Québec, 2003.

dénomination², écarte les relations de synonymie et de polysémie. Parallèlement, on élève au rang de dogme le principe d'évitement de l'emprunt à l'anglais [Darbelnet 1966 : 24; Depecker 2001], de manière indifférenciée, voire des emprunts intégraux non intégrés à la langue française aux emprunts sémantiques de type calque morphologique intégrés à la langue française. Ce faisant, on néglige, entre autres, de prendre en compte les pratiques terminologiques réelles comme la variation terminologique et le calque technoscientifique à partir de l'anglais. C'est ainsi que Daoust³, cité par Maurais [1987 : 406], soulignera que l'implantation terminologique serait vraisemblablement le point faible du plan d'aménagement terminologique québécois : « Il est notoire que l'Office de la langue française a produit des lexiques pendant près de deux décennies sans réussir à les implanter. »

Sept périodes viennent ponctuer les activités d'implantation de la terminologie française au Québec : 1) 1961-1969 – Correction et enrichissement de la langue [Bouchard 1995; René 2001]; 2) 1969-1974 – Début d'implantation en milieu de travail [Bouchard 1995; René 2001]; 3) 1974-1977 – Programme de francisation [Bouchard 1995; René 2001]; 4) 1977-1989 – Certification et diffusion intenses [Bouchard 1995; René 2001]; 5) 1989-1996 – Remise en question des stratégies [Bouchard 1995; René 2001]; 6) 1996-2001 – Considération des besoins de diffusions [René 2001; Célestin *et al.* 2003]; 7) 2001-2008 – Essai de valorisation socioterminologique [Célestin *et al.* 2003 : 9; Bergeron 2004].

La dernière période portant sur l'essai de valorisation socioterminologique est visé directement par nos travaux; nous considérons les pratiques terminologiques réelles des usagers que nous situons comme élément central du plan d'aménagement terminologique.

2.1. Apport théorique des études sur l'implantation terminologique

Des études ont été menées sur et autour du sujet de l'implantation de la terminologie française, et permis des avancées considérables - au Québec, avec

² De sorte que pour une notion il existe une seule dénomination et une dénomination pour une unique notion.

³ Traduction de Daoust, dans Daoust, D. « Francization and Terminology Change in Quebec Business Firms », 1984, dans Bourrhis, R.Y. *Conflict and Language Planning in Quebec*, Clevedon : Multilingual Matters Ltd, p. 81-113, p. 94.

Auger [1983⁴, 1999⁵], Maurais [1984⁶], Brent [1986⁷], Loubier [1993⁸], Martin [1993⁹, 1998¹⁰], Bouchard [1995] et Quirion [2003¹¹], en France, avec Chansou

⁴ Auger [1983: 28-32] a présenté, dans le cadre du colloque *Aménagement de la terminologie : diffusion et implantation*, une procédure en neuf phases pour réaliser un aménagement terminologique de manière cohérente et rigoureuse, avec le dessein de pallier les stratégies lacunaires du plan d'aménagement terminologique québécois selon lequel il est difficile de faire passer la terminologie dans l'usage.

⁵ Auger [1999] s'est appliqué à mesurer le degré de l'implantation des formes normalisées (officialismes) des noms de commerce en langue française des poissons à potentiel commercial, par l'administration fédérale et provinciale, l'industrie des pêcheries, le commerce de gros et de détail du poisson et des produits de la mer au Québec ainsi que dans une moindre mesure, par les consommateurs.

⁶ Maurais s'est attaché à analyser l'évolution de la qualité et du statut de la situation terminologique du domaine de la publicité des chaînes d'alimentation. Il fera le constat d'une amélioration de la qualité du français, l'émergence d'un français québécois standard, les questions relatives à l'acceptation ou le refus par l'Office de formes en usage, la tendance lourde à rejeter les emprunts intégraux au « bénéfice » des calques sémantiques faux amis et les cas de variation qui sont acceptables ou inacceptables. Au cours de la même année, cet auteur publiera un ouvrage sur une enquête sur la langue des affiches publicitaires. Elle constitue la première véritable étude ayant abordé l'évaluation de l'implantation des termes normalisés par l'Office. Inédite, elle fera ressortir que « des dix termes étudiés, seulement trois peuvent être considérés comme ayant été implantés avec succès (c'est-à-dire que leur taux d'occurrence dans le corpus dépasse les 50 %). [...] » [Maurais 1994 : 446].

⁷ Brent [1986] a rédigé un guide d'implantation de la terminologie française dans l'entreprise. Inédit, ce document visait à nourrir la réflexion et à donner des indications pratiques à l'attention des gestionnaires et professionnels.

⁸ Loubier [1993] s'emploiera à examiner le processus de changement (termino)linguistique planifié, pour parvenir à l'implantation du français dans la langue du travail au Québec, en suggérant un modèle de processus d'implantation qui permettrait l'implantation du français réelle et durable.

⁹ Martin [1993] s'appliquera à élaborer un nouveau cadre théorique pour faciliter l'implantation terminologique, en présentant le phénomène de la dynamique sociale de l'usage et de la pénétration des terminologies, sur la base de la théorie de la diffusion sociale des innovations établie par Rogers et ses collaborateurs⁹.

(audiovisuel et publicité), Fossat et Rouges-Martinez (télé-détection aérospatiale), Gouadec (informatique), Guespin *et al.* (génie génétique) et Thoiron *et al.* (santé et médecine) publiés dans Depecker et Mamavi [1997] -.

Néanmoins, force est de constater, d'une part, l'ignorance de l'état actuel de la situation terminologique française en milieu de travail et d'autre part, la nécessité de déployer des efforts continus pour assurer une présence actualisée et durable à la terminologie française des utilisateurs francophones, eu égard à l'hégémonie de la langue anglaise sur les autres langues, ici la langue française.

2.2. Notre apport théorique et pratique

2.2.1 Constatation d'une problématique de l'implantation terminologique française

Suivant l'examen des études sur l'implantation terminologique française, nous avons remarqué une problématique portant sur quatre aspects : la production, la diffusion, la mesure de l'implantation et la mesure de suivi.

On ne peut que constater une lacune terminologique évidente en raison d'« un retard néologique qui ne fait que s'accroître » [DCMCC, 1996 : 236], ce que nos résultats de mémoire de maîtrise [Di Spaldro 2006] ont également suggéré : 33 % des termes candidats de notre corpus portant sur des textes québécois et français couvrant les années 1985 à 2004 sont absents des banques de terminologie et des dictionnaires terminologiques consultés en 2005. Le retard à la francisation constituant une cause de l'implantation durable de l'anglais [Depecker et Mamavi 1997 : XXVII].

On remarque des défauts quant à la diffusion de la terminologie normalisée par les dispositifs de régulation terminolinguistiques, de sorte que « les avis, les travaux, l'existence même de la Commission de terminologie sont méconnus »

¹⁰ Martin [1998] a effectué une enquête sur l'implantation de la terminologie normalisée par l'Office du domaine de l'éducation au Québec à partir des travaux de la Commission de terminologie de l'éducation du Québec (CTE).

¹¹ Quirion [2003] a établi un protocole de mesure de l'implantation des terminologies (terminométrie) normalisées du domaine des transports au Québec, en cherchant à l'universaliser par le biais d'une méthode scientifique, avec pour but ultime de fournir des données pour amorcer des recherches et réflexions sur les causes de succès ou d'échec des implantations.

[Martin, 1998 : 190], et « en général, on ne distingue pas un avis de normalisation d'un avis de recommandation » [Martin, 1998 : 190], voire de proposition. Constatations également dégagées par Depecker [Depecker et Mamavi 1997 : XXIX] au sujet des Commissions ministérielles de terminologie. Nommément, par Thoiron *et al.* [dans Depecker et Mamavi 1997 : 66] qui signalent qu'« environ la moitié des généralistes interrogés ne connaît pas l'existence des arrêtés et la majorité d'entre eux refuserait de toute façon de les respecter eux-mêmes, de les diffuser à leurs confrères ou de les imposer à leurs subordonnés » et que les spécialistes connaissent mieux l'existence des arrêtés, « mais de façon encore insuffisante ». À titre d'illustration, le terme *marqueur de clonalité* est absent des banques de terminologie, ce qui constitue une menace à la terminologie française, étant donné que seul le terme anglais, langue des publications technoscientifiques, est connu, *clonality marker*.

Aucune mesure d'implantation terminologique n'est instaurée afin d'évaluer les terminologies utilisées (terminométrie) : « [...] nous ne possédons pas de données sur l'utilisation réelle et effective des terminologies françaises en milieu de travail [...] » [Martin 1996 : 9]. Théoriciens, praticiens et chercheurs réclament pourtant une telle mesure, certains depuis plus de 30 ans, entre autres, Guilbert [1975 : 3592], Auger [1983 : 36], Brent [1986 : 37-39], Loubier [1993 : 124-126], Martin [1993 : 43], Quirion [2003 : 168-169], etc. Le fait qu'aucune mesure d'implantation terminologique ne soit instaurée rend ardue et arbitraire la détermination du degré d'utilisation de la terminologie française par les utilisateurs, et son évolution dans le temps. En outre, on ne peut associer certificat de francisation et utilisation réelle du français : « [u]n fait est certain pourtant : le niveau de francisation générale n'est pas nécessairement en corrélation directe avec le niveau de francisation de la terminologie » [Daoust – Blais 1981 : 2]. Il est plutôt « un préalable à un fonctionnement en français, à la vie en français » [Bouchard 2002 : 90]. On ignore donc l'état de la terminologie française.

Aucune mesure de suivi n'est en place pour contrôler la terminologie diffusée. Notamment, on ne peut s'assurer que la terminologie *officialisée* a été transmise aux travailleurs, qu'elle a été comprise ni que ceux-ci n'aient été convaincus du bien-fondé de son utilisation. La normalisation ou la recommandation terminologiques semblent peu déterminantes, autour de 30 %, selon les études conduites au Québec par Maurais [1984], Martin [1998] et Auger [1999], avec des taux similaires relevés en France dans Depecker et Mamavi

[1997 : XXIX]. On ignore ainsi l'impact du français sur la terminologie en milieu de travail francophone.

2.2.2 Théorisation heuristique de l'élaboration d'une méthode assistée par ordinateur pour la francisation de la terminologie

En vue de contribuer à résoudre la problématique de l'implantation terminologique décrite ci-dessus, nous avons exploré une piste de solution, celle de l'élaboration d'un modèle théorique prototypique de système logiciel.

La méthode vise à ce que l'utilisateur ait accès en tout temps à des termes français. Elle a pour fondement une métagrille élaborée dans le cadre du mémoire de maîtrise (Master 2 en France), un outil d'aide à la délimitation des termes et à leur viabilité. Cette métagrille a été créée à partir de critères d'acceptabilité de l'OQLF 1998, ©2001-2003-2004, de Loubier 2003 et de critères d'acceptabilité [Guilbert 1970 : 117, 120, 123; Kocourek 1991, ©1982 : 151; Depecker 2001, etc.] et de choix terminologiques [Guilbert 1970 : 122-123; Auger 1979, etc.] distinguant unités lexicales et de discours, retenus par ou de Di Spaldro 2006.

Cette méthode devrait contribuer à l'avancement des deux branches de la discipline de la terminologie que sont l'aménagement terminologique et la socioterminologie. Elle légitime l'amorce d'une réflexion sur l'élaboration d'une plate-forme informatique pour l'implantation d'une terminologie française en milieu de travail. Spécialement, par l'application d'une partie de la méthode à l'aide de trois questions épiterminologiques s'adressant à notre échantillon de praticiens dans la deuxième partie du questionnaire d'enquête Qualtrics¹², test que nous expliciterons au point 2.2.3.

Cette méthode fait du praticien au sein d'une sphère d'activité de métiers et de professions l'élément central du plan d'aménagement terminologique, tandis que Prairie [1986 : 118]¹³, cité par Maurais [1987 : 387], déplore que « dans leur ensemble, les travailleurs et les travailleuses n'auraient joué qu'un rôle assez

¹² Logiciel d'enquête en ligne (<http://www.qualtrics.com/>).

¹³ Il a effectué une enquête auprès de 35 représentants de travailleurs aux comités de francisation, entre décembre 1982 et janvier 1984.

secondaire dans la francisation des entreprises »¹⁴, alors que Loubier [1993 : 90] exhorte de « [...] placer ou replacer toute l'activité et les activités reliées au processus d'aménagement [...] [terminologique] dans leur contexte réel en tenant compte des acteurs sociaux que sont les locuteurs », et que Trousson¹⁵, cité par Martin [1998 : 201], déplore le fait que « les locuteurs sont à ce jour les grands absents de l'intervention officielle sur la langue ». Cet outil laisserait également place à la variation terminologique, tandis que Depecker, au sujet des cinq études d'implantation menées pour le compte de la DGLF(LF), souligne la remise en cause de la notion même d'usage, en raison de la « fragmentation discursive » observée¹⁶.

Parce qu'il produira lui-même, à l'aide de la méthode, la terminologie française non accessible ou critiquée, le praticien pourrait éviter le retard à la francisation, « une des causes de l'implantation durable de l'anglais » [Depecker et Mamavi 1997 : XXVII].

2.2.2.1 Contenu de la plate-forme informatique

La plate-forme informatique sera composée d'une base de données qui servira à l'enrichissement du stock terminologique français. Elle sera constituée de quatre éléments que nous verrons plus en détail ci-dessous : 1) une série d'automates pour reconnaître les principales structures syntagmatiques du domaine de la génétique; 2) un automate de reconnaissance des critères de sélection de choix terminologiques; 3) une base pour stocker les termes du domaine de la génétique contenus dans les banques de terminologie du GDT, de la DGLFLF et de TermiumPlus; 4) un formulaire démographique.

Les automates permettront de reconnaître les termes constitués d'une base suivie d'une ou plusieurs expansions, de sorte que tant le terme simple, complexe ou surcomplexe seront pris en considération. Un exemple d'automate pour le

¹⁴ Il faut entendre par *travailleurs*, une personne qui « travaille », que son statut occupationnel relève de l'ouvrier, du technicien, du spécialiste ou du scientifique.

¹⁵ Trousson, M. *Une politique terminologique pour la Communauté française de Belgique : bilan et réflexion prospective*, mémoire de stage de secrétaire d'administration, Bruxelles, [Bruxelles] : ministère de la Culture et des Affaires sociales/Service de la langue française, 1996.

¹⁶ Depecker, L. « Introduction », dans Depecker, L. et G. Mamavi. *La Mesure des mots : cinq études d'implantation terminologique*, [Rouen], Publications de l'Université de Rouen, p. VII-XXXVII, p. XXXII, 1997.

terme amorce d'hybridation au hasard sera construit selon la formule syntagmatique **substantif + préposition + (substantif + article contracté + substantif)**.

Le terme-clé entré devra satisfaire deux des trois critères suivants : 1) bien formée selon le modèle linguistique français; 2) permet de comprendre le concept; 3) peut servir de modèle pour créer d'autres termes (gène, génome, génomique, etc.). Ces critères devront être automatisés, afin que le terme-clé produise une réponse immédiate.

Il est prévisible que le stock de termes du domaine de la génétique des banques de terminologie du GDT, de la DGLFLF et de TermiumPlus ne soit pas renouvelé faute de ressources.

Le logiciel s'activera à la condition de remplir quatre cases, à l'aide de choix de réponses, ayant trait à l'occupation, au département occupé, au sexe et à l'âge, ce qui devrait prendre tout au plus 10 secondes. De cette façon, il sera possible de mesurer la fréquence d'utilisation en plus de recueillir des données pour fin d'études. Une cinquième case permettra l'inscription de tout commentaire, permettant ainsi de « collecter régulièrement les commentaires, analyses et propositions des experts, réagissant par rapport aux problèmes terminologiques rencontrés sur le terrain » [dans Depecker et Mamavi 1997 XXXIII, citant Rouges-Martinez¹⁷].

2.2.2.2 Fonctionnement de la méthode par l'intermédiaire de la plateforme informatique (ou logiciel)

Le principe fondateur de la méthode consiste en l'utilisation d'une norme de traduction française littérale adaptée de la dénomination anglaise du terme ou néoterme recherchée, sur le modèle de la composition syntagmatique nominale (calque technoscientifique ou CTS).

C'est que les terminologies de pointe, parmi lesquelles la génétique humaine, sont généralement publiées dans la langue anglaise [Lagueux 1988 : 94]. En anglais, le vocabulaire technoscientifique construit à l'aide d'une base et d'une expansion suit l'ordre déterminant-déterminé et serait « spécifique de la typologie de l'allemand » [Guilbert 1970 : 123]. En langue française, il « se réalise par déterminations successives selon le développement linéaire [de gauche à droite] »

¹⁷ Rouges-Martinez, J. *L'implantation terminologique dans le domaine de la télédétection aérospatiale*, Université de Toulouse-le-Mirail : Toulouse-le-Mirail, 1992.

[Guilbert 1970 : 117] et selon l'ordre déterminé-déterminant [Benveniste 1966 : 91; Guilbert 1970 : 117; Auger 1979 : 15; Dubuc 1992 : 43; Pruvost et Sablayrolles 2003 : 105, etc.].

À titre d'illustration, pour clonality marker, nous aurions marqueur de clonalité. D'où le calque technoscientifique. Il faut savoir toutefois que le CTS peut admettre erronément un même champ notionnel [Tardivel 1880¹⁸, cité par Bouthilier et Meynaud 1972 : 207, Sournia 1974 : 25, 28, Depecker 2002 : 35, 58] générant un calque sémantique de type faux ami, par exemple, *librairie de gènes au lieu de bibliothèque de gènes pour gene library.

Puis, une interface de dialogue interactive, disponible en tout temps, sera générée, dans laquelle l'utilisateur entrera un terme issu de la langue anglaise, traduit en français. Ce procédé devrait ainsi produire un CTS tel marqueur de clonalité. Il y aura alors renvoi d'une réponse terminologique immédiate acceptant le terme tel quel ou proposant un autre terme.

La méthode sera utile lorsque l'équivalent français pour une dénomination anglaise est inconnu, non accessible ou écarté. En entrant le terme-clé, le système indiquera si ce terme est acceptable et viable, d'après une liste d'automates du domaine, la satisfaction de deux critères d'acceptabilité et de viabilité sur trois et les termes de la génétique dans les banques de terminologie du GDT, de la DGLFLF et TermiumPlus.

La base de données de la plate-forme informatique indiquera alors la dénomination telle quelle avec sa définition si cette dernière est déjà stockée dans le logiciel, ou elle en affichera la forme variante ou synonymique. Il pourra arriver qu'exceptionnellement un avis soit affiché mentionnant que la dénomination sera traitée dans les 24 prochaines heures, car elle ne correspond jusqu'alors à aucun des automates créés et qu'il faut par conséquent ajouter ce nouvel automate à la liste.

Le terme-clé produira une réponse automatisée immédiate, positive ou négative. La satisfaction de deux critères sur trois engendrerait une réponse positive, alors que la satisfaction d'un seul critère produirait une réponse négative. Une explication serait donnée quant au résultat. Le logiciel garderait en mémoire

¹⁸ Tardivel J.-P. *L'anglicisme, voilà l'ennemi!*, Cercle catholique de Québec, 17 décembre 1879, Québec : Imprimerie du Canadien, 1880.

toutes les réponses, pour fin d'études, mais ne diffuserait que les réponses positives.

2.2.3 Application d'une partie de la méthode

La partie de la méthode éprouvée a porté sur la deuxième section du questionnaire concernant l'usage effectif. Elle comportait vingt-deux notions, dont soixante-huit dénominations. Le corpus de référence de quelque 70 000 mots a été formé à partir de trois textes de trois sources différentes retenus par les participants, pour un total de quelque 5 000 mots ou une vingtaine de pages, tirés de sites internet donnant accès à des textes publics représentatifs de la terminologie actuelle utilisée dans leur milieu de travail. L'analyse du texte pour en extraire des termes a été effectuée par le biais du logiciel indexeur textuel *Linguistica*¹⁹, mais aussi manuellement. Comme la majorité des textes étaient de langue anglaise, nous avons traduit les termes, afin qu'ils correspondent à un calque littéral adapté. Nous les avons par la suite confrontés à l'usage effectif dans le moteur de recherche Google. Seuls ceux retrouvés dans trois sites internet fiables et de trois différentes sources ont été retenus et insérés dans le questionnaire.

Le test a été conçu de la façon suivante. Pour chacune des dénominations d'une notion donnée issue du corpus de référence, assortie de son équivalent anglais, le répondant avait pour tâche de cocher les cases correspondant aux trois critères susmentionnés qu'il estimait satisfaits (voir ci-dessous le tableau 1).

Tableau 1 : Test de la méthode à partir de trois questions épiterminologiques

Dénomination	Bien formée selon le modèle linguistique français	Permet de comprendre le concept	de	Peut servir de modèle pour créer d'autres termes
candidat gène/candidat gene		✓		✓
virus à ARN/RNA virus	✓	✓		✓
transcription par l'ARN polymérase/RNA polymerase transcription	✓	✓		✓

3. Résultats

3.1. Résultats attendus

Aucune étude d'implantation, à notre connaissance, n'a établi une méthode assistée par ordinateur contribuant à la francisation terminologique, afin de résoudre la problématique de l'implantation de la terminologie française.

La plate-forme informatique de la méthode n'a pas encore été créée; le but de la réflexion étant surtout de tenter d'en formuler la théorisation, ensuite d'en examiner l'utilité, puis la faisabilité. Son application mettait essentiellement en jeu les trois critères de sélection des choix terminologiques assurant une certaine acceptabilité et viabilité terminologique.

3.2. Résultats obtenus

Le test a permis d'évaluer une partie de la méthode assistée par ordinateur pour l'implantation d'une terminologie française.

Nous avons remarqué que le critère 1, *Dénomination bien formée selon le modèle linguistique français*, a récolté des réponses significatives. Néanmoins, tant du point de vue de l'axe syntagmatique que de l'aspect cognitif de la formation terminologique, il n'a pu fournir que des réponses aléatoires et non-explicites. Le critère 2, *Dénomination permet de comprendre le concept*, a présenté également des

¹⁹ Mis au point par Jacques Ladouceur.

réponses significatives. Toutefois, il autorisait des contradictions possibles avec le critère 1 : le critère 1 appelant une dénomination générique; le critère 2, une dénomination spécifique. Le critère 3, *Peut servir de modèle pour créer d'autres termes (gène; génome; génomique)*, a autorisé peu de réponses significatives. Voir le tableau 2 ci-dessous.

Tableau 2 : Exemple des réponses des trois questions épiterminologiques dans le questionnaire Qualtrics

#	Question	Bien formée selon le modèle linguistique français	Permet de comprendre le concept	Peut servir de modèle pour créer d'autres termes (gène, génome, génomique, etc.)	Responses
1	antioncogène	5	3	2	10
2	anti-oncogène	8	7	1	16
3	gène suppresseur de tumeur	10	7	1	18
4	gène suppresseur de tumeurs	5	5		10
5	gène suppresseur tumoral	3	3		6
6	oncosuppresseur	6	4	2	12
7	onco-suppresseur	5	4		9
8	autre(s)				
9	commentaire(s)				

autre(s) commentaire(s)

4. Conclusion

La théorisation heuristique de l'élaboration d'une méthode assistée par ordinateur pour la francisation de la terminologie suggère que la méthode se révèle un outil de production, de diffusion, de mesure, de mise à jour, de collecte et d'analyse de données efficace.

En tant qu'outil de **production** de l'expression des connaissances, la méthode autorise la fabrication de termes et néotermes acceptables, en palliant l'absence de termes - officialisés dans les banques de terminologie ou sites

internet gouvernementaux²⁰, ou attestés dans les dictionnaires terminologiques ou ouvrages de référence spécialisés -, ou en remédiant à l'ignorance de leur existence ou à leur rejet. Ce sont les utilisateurs eux-mêmes qui entérineront termes et néotermes. Le cautionnement de l'acceptabilité terminologique de ces derniers sera facilité, eu égard au principe néologique qui s'adresse non seulement aux littérateurs et académiciens, mais à la « nation entière » [Mercier 1801 : lxxi], à un écart minime entre « l'usage établi et l'usage désiré » [Leblanc 1994 : 517], à l'implication des utilisateurs dans le processus de francisation des terminologies²¹ [Loubier 1991 : 13] et à la reconnaissance de marges de décision laissées au sujet » [Gaudin 1990 : 152].

Comme outil de **diffusion** des connaissances en langue française, elle favorise l'accroissement de manière substantielle de la diffusion des terminologies. De plus, par l'exploitation de la base de données de la plate-forme, ces termes pourront non seulement être diffusés automatiquement et s'afficher sur une interface de type logiciel libre (Wikipédia), mais également rendre possible l'utilisation des matériaux produits aux organismes de régulation terminolinguistiques étatiques dans le cadre de leurs travaux d'*officialisation*. Ils pourraient ainsi les légitimer, les promouvoir, les corriger dans le cas de production de termes fautifs (faux amis, quasi-synonymes). Ultimement, ils pourraient verser ces termes et néotermes dans le système logiciel ou dans leur base de données terminologiques, dans le but d'offrir, notamment des synonymes aux emprunts intégraux à l'anglais ou de proposer des termes appropriés aux termes inappropriés tels les faux amis et les quasi-synonymes.

En tant qu'outil de **mesure** de l'implantation de la terminologie, la méthode rend possible la comptabilisation du pourcentage de déclaration d'utilisation des termes, par le remplissage du formulaire démographique.

²⁰ Par exemple, le Grand dictionnaire terminologique (GDT), organe de diffusion officiel de l'OQLF; TermiumPlus, organe de diffusion officiel du ministère canadien des Travaux publics; la Délégation générale de la langue officielle et autres langues de France (DGLFLF), organe de diffusion officiel des arrêtés publiés par le Journal officiel, etc.

²¹ À l'opposé, l'enquête de Prairie révélera que les travailleurs et les travailleuses n'auraient joué qu'un rôle assez secondaire dans la francisation des entreprises. Voir : Prairie, M. *La francisation des entreprises : l'expérience vécue par des travailleurs et des travailleuses de la CSN et de la FTQ*, Montréal : UQAM/CSN/FTQ, 1986.

Comme outil de **mise à jour**, elle permet la sauvegarde des termes et néotermes retenus, et potentiellement, l'utilisation d'une interface de type logiciel libre (Wikipédia) ainsi que le versement dans la base de données des termes et néotermes traités par les organismes de régulation terminologiques étatiques.

En tant qu'outil de **collecte et d'analyse de données**, la programmation de la méthode telle que décrite pourrait servir à des études de type ethnoterminologiques, en raison des questions préalables posées à l'ouverture du système, habilitant ainsi la collecte de données pour des fins de recherche. Par conséquent, elle pourrait faciliter la prise de décisions judicieuses des stratégies à mettre en place quant au plan d'aménagement terminologique, à l'évolution des procédés terminogènes employés et des types d'usage en cours (officialisme, variation terminologique, calque technoscientifique, anglicismes, langue de travail), voire même, aux rapports entre notions et désignations, glissements de sens, changements terminologiques, aux réseaux interactionnels des milieux de travail et aux distinctions entre pratiques des sciences, des techniques et des industries terminologiques.

Enfin, l'application d'une partie de la méthode a pu être validée grâce au questionnaire faisant partie notre enquête terminologique, par l'intermédiaire de trois questions épiterminologiques représentant trois critères de sélection pour l'acceptabilité et la viabilité des termes.

Compte tenu des résultats obtenus, nous avons dû revoir nos critères. Au premier a été ajouté *et propre au domaine*, le second est demeuré intact; le troisième a été modifié pour *Peut servir à enrichir une famille de termes apparentés*, et les exemples assortis ont subi également des modifications, notamment, par l'ajout d'un exemple de terme complexe *séquence → de tête, séquence leader*. Voir le tableau 3 ci-dessous.

Tableau 3 : Test reformulé à partir des trois questions terminologiques

1. Occupation	Bien formée selon le modèle linguistique français et propre au domaine	Permet de comprendre le concept	Peut servir à enrichir une famille de termes apparentés
2. Département	(ex. : exposition sur phage/phage display)	(ex. : gène tardif : gène viral transcrit à la fin de l'infection cellulaire)	(ex. : gène → génome, génétique séquence → de tête, séquence leader)
3. Sexe			
4. Âge			
Dénomination x			

29

Ce nouveau test donnant lieu à une formulation différente devrait être éprouvé auprès de répondants québécois, lors du deuxième et dernier volet de l'enquête socioterminologique, de septembre à novembre 2008.

5. Perspectives

Cet article nous a permis de réfléchir au processus d'implantation des terminologies françaises, point crucial de l'aménagement terminologique, en lien avec la modélisation des connaissances métiers.

Dans le dessein d'offrir une application générant des solutions pratiques quant à la problématique de l'implantation de la terminologie française, compte tenu des lacunes de production de termes, de diffusion, de mesure et de suivi, nous avons théorisé une méthode assistée par ordinateur pour l'implantation de la terminologie française, afin que les praticiens aient accès en tout temps à la terminologie française, et en avons éprouvé une partie, à partir de notre métagrille d'aide à la délimitation des termes et à leur viabilité, réduite à trois critères et simplifiée au point de créer des critères très généralisés.

Cette théorisation ainsi que les résultats de son application partielle devraient permettre de faire progresser la discipline de l'aménagement terminologique, entre autres, en autorisant la modélisation des connaissances métiers. En outre, ces avancées pourraient possiblement consentir à des prises de décision éclairées par les dispositifs de régulation terminolinguistiques étatiques pour favoriser l'implantation terminologique en langue française, entre autres, en s'employant à décrire les changements linguistiques spontanés et planifiés des pratiques langagières [Gaudin, 2003].

Les bases d'une méthode d'implantation terminologique assistée par ordinateur pour l'implantation du français ont pu être jetées.

Par l'intermédiaire de cette méthode, il est possible de concevoir une certaine dématérialisation géographique informatique de données terminologiques. Elle se révèle non sans utilité en cette ère de mondialisation des marchés où la langue de communication est l'anglo-américain. L'exploitation d'une interface de type logiciel libre à l'aide de la base de données de la plateforme permet de concevoir l'émergence d'un réseau francophone de néologie technoscientifique; la méthode agissant comme une passerelle cognitive intergroupes socioprofessionnels et inter-organismes terminolinguistiques.

Par conséquent, les retombées de la méthode pourraient, d'un point de vue international, servir à enrichir les travaux d'aménagement terminologique québécois, canadien, catalan, gaélique, etc., en encourageant notamment les partenariats internationaux et les nouveautés technologiques préconisés par l'OQLF [Célestin *et al.* 2003 : 14] dans ses efforts de francisation et

d'enrichissement de la langue française; d'une perspective nationale, elles pourraient servir de compléments aux travaux de veille néologique de l'Observatoire de néologie du Québec (OBNEQ).

Comme l'application de la méthode assistée par ordinateur pour l'implantation de la terminologie française a porté sur l'épreuve d'une partie de la méthode et non sur la méthode elle-même, ni la programmation d'automates du domaine, ni l'automatisation des critères de sélection des choix terminologiques autorisant des termes acceptables et viables, non plus que l'exploitation du stock des termes du domaine de la génétique des banques de terminologie du GDT, de la DGLFLF et de TermiumPlus n'ont été réalisées. Par conséquent, il serait indiqué d'évaluer de manière réaliste la possibilité de créer une telle plate-forme informatique.

En matière de procédé terminogénique, il sera pertinent de comparer nos résultats aux travaux actuels de Quirion [2003] intéressant l'automatisation de la terminométrie et ceux de Picone [1991] concernant l'impulsion synthétique.

Quant à la formulation améliorée autorisant l'épreuve d'une partie de la méthode, elle pourra être évaluée en novembre 2008 auprès d'utilisateurs québécois. Il sera intéressant d'effectuer des études comparatives sur les perceptions et les attitudes des Français et des Québécois, la conscientisation de néologisation, particulièrement en ce qui a trait à la possibilité d'éprouver le critère 3 reformulé : *Peut servir à enrichir une famille de termes apparentés (ex. : gène → génome, génétique, séquence → de tête, séquence leader)*. Ces résultats pourront se révéler une nouvelle source d'études sur les pratiques terminologiques françaises.

Bibliographie

Auger P. « *La syntagmatique terminologique, typologie des syntagmes et limite des modèles en structure complexe* », dans *Commission de terminologie de l'AILA. Table ronde sur les problèmes de découpage du terme : actes du 5^e Congrès international de linguistique appliquée, AILA, Montréal, Québec, 20-26 août 1978, [Québec] : Office de la langue française, p. 9-26, 1979.*

Auger P. « *La problématique de l'aménagement terminologique au Québec* », dans *Office de la langue française/Société des traducteurs du Québec. Aménagement de la terminologie : diffusion et implantation : actes du 4^e Colloque OLF/STQ, Québec, p. 255-280, 1983.*

Auger P. *L'implantation des officialismes halieutiques au Québec : essai de terminométrie*, coll. *Langues et sociétés*, n° 37, [Montréal] : Office de la langue française, 1999.

Benveniste É. *Bulletin de la Société de Linguistique*, Paris, p. 88-106, 1966.

Bergeron M. « Le traitement de la variation terminologique dans les technologies de l'information à l'Office de la langue française », dans Bouchard, P. et R. Vézina. *La variation dans la langue standard : actes du colloque, Québec, 13-14 mai 2002, 70^e Congrès Acfas*, coll. *Langues et sociétés*, n° 42, [Montréal] : Office québécois de la langue française, p. 195-204, 2004.

Bouchard P. « L'implantation de la terminologie française au Québec : bilan et perspectives », *Présence francophone*, n° 47, p. 53-79, 1995.

Bouchard P. « La langue du travail : une situation qui progresse, mais toujours teintée d'une certaine précarité », *Revue d'aménagement linguistique*, p. 85-104, 2002.

Bouthillier G. et J. Meynaud. *Le choc des langues au Québec, 1760-1960*, Montréal : Presses de l'Université du Québec, 1972.

Brent E. *Guide d'implantation de la terminologie française dans l'entreprise*, document inédit, Montréal : Office de la langue française, 1986.

Célestin T., M. Bergeron, A. Galarneau et J. Maltais. « Le phénomène de la néologie technique et scientifique au Québec », *La néologie scientifique et technique : bilan et perspectives : actes du colloque, Réseau panlatin de terminologie, Rome, 28 novembre 2003, Groupe de travail sur la néologie [Québec] : Office québécois de la langue française*, p. 1 à 19, http://dtil.unilat.org/realiter_spip/spip.php?article222, [2003].

Daoust-Blais D. *Diffusion et utilisation de la terminologie technique de langue française dans douze entreprises québécoises*, synthèse de l'étude réalisée par Sorécom, [Québec] : OLF, 1981.

Darbelnet J. *Regards sur le français actuel*, Montréal : Beauchemin, 1963.

Depecker L. *L'invention de la langue : le choix des mots nouveaux*, Paris : A. Colin/Larousse, 2001.

Depecker L. *Entre signe et concept : éléments de terminologie générale*, [Paris] : Presses Sorbonne Nouvelles, 2002.

Depecker L. et Mamavi G. *La Mesure des mots : cinq études d'implantation terminologique*, n° 229, [Rouen] : Publications de l'Université de Rouen, 1997.

Direction des communications du ministère de la culture et des communications (DCMCC). *Le français langue commune : enjeu de la société québécoise, rapport du comité interministériel sur la situation de la langue française*, Gouvernement du Québec, 1996.

Di Spaldro J. *Les emprunts à l'anglais médical dans la langue française contemporaine, mémoire de maîtrise*, Université de Montréal, 2006.

Dubuc R. *Manuel pratique de terminologie*, 3^e éd., Brossard : linguattech éditeur, 1992.

Gaudin F. « De l'interaction à la terminologie : le travail scientifique », *Cahiers de linguistique sociale*, n° 17, *Linguistique et matérialisme : actes des rencontres de Rouen*, p. 149-160, 1990.

Gaudin F. *Socioterminologie : une approche sociolinguistique de la terminologie*, 1^{re} éd., coll. *Manuels champs linguistiques*, Bruxelles : De Boeck & Larcier et Duculot, 2003.

Gibson G. et Muse S. *Précis de génomique*, de Boeck et Larcier, 2004.

Guilbert L. « La dérivation syntagmatique dans les vocabulaires scientifiques et techniques », dans *Les langues de spécialité. Analyse linguistique et recherche pédagogique : actes du stage, Saint-Cloud, 23-30 mars 1967*, Strasbourg : Association internationale d'éditeurs de linguistique appliquée, p. 116-124, 1970.

Guilbert L. « La néologie », *Grand Larousse de la langue française*, Paris : Librairie Larousse, tome 4, p. 3584-3594, 1975.

Kocourek R. *La langue française de la technique et de la science : vers une linguistique de la langue savante*, 2^e éd., Auflage : Oscar Brandsetter Verlag GmbH & Co. KG/Wiesbaden, ©1982, 1991.

Lagueux P.-A. « La part des emprunts à l'anglais dans la création néologique, en France et au Québec », dans Darbelnet, J. et M. Pergnier (dir.), *Le français en contact avec l'anglais : en hommage à Jean Darbelnet*, coll. *Linguistique*, n° 21, Paris : Didier Éruditions, p. 91-111, 1988.

Leblanc B. « L'implantation terminologique en usine : ajustements nécessaires », dans *Office de la langue française/Université du Québec à Chicoutimi. Problématique de l'aménagement linguistique : actes du 8^e Colloque international de terminologie, OLF-UQAC, 5-7 mai 1993*, coll. *Langues et sociétés*, tome II, [Montréal] : Office de la langue française, p. 515-521, 1994.

Loubier C. *L'importance de l'activité terminologique dans le processus de francisation du Québec*, document inédit, Québec : Office de la langue française, 1991.

Loubier C. « L'implantation du français comme langue du travail au Québec : vers un processus de changement linguistique planifié », dans Martin, A. et C. Loubier. *L'implantation du français : actualisation d'un changement linguistique planifié*, coll. *Langues et sociétés*, [Montréal] : Office de la langue française, p. 56-133, 1993.

Martin A. « Théorie de la diffusion sociale des innovations et changement linguistique planifié », dans Martin, A. et C. Loubier. *L'implantation du français : actualisation d'un changement linguistique planifié*, coll. *Langues et sociétés*, [Montréal] : Office de la langue française, p. 9-55, 1993.

Martin A. « La production terminologique : un aménagement de la langue ou un aménagement de son statut? », *Terminogramme*, n° 79, p. 6-9, 1996.

Martin A. *Les mots et leurs doubles : étude d'implantation de la terminologie officialisée dans le domaine de l'éducation au Québec*, coll. *Langues et sociétés*, n° 36, [Montréal] : Office de la langue française, 1998.

Maurais J. *La langue de la publicité des chaînes d'alimentation. Étude sur la qualité de la langue et sur l'implantation terminologique*, coll. *Dossiers du CLF*, n° 18, Québec : Conseil de la langue française, 1984.

Maurais J. « L'expérience québécoise d'aménagement linguistique », dans Maurais, J. *Politique et aménagement linguistiques*, Québec/Paris : Conseil de la langue française/Le Robert, p. 359-416, 1987.

Maurais J. « Quelques aspects sociolinguistiques de l'implantation des décisions de normalisation terminologique », dans Office de la langue française/Université du Québec à Chicoutimi. *Problématique de l'aménagement linguistique : actes du 8^e Colloque international de terminologie*, OLF-UQAC, 5-7 mai 1993, coll. *Langues et sociétés*, tome II, [Montréal] : Office de la langue française, p. 441-453, 1994.

Mercier L.-S. *Néologie; ou, Vocabulaire de mots nouveaux à renouveler, ou pris dans des acceptations nouvelles*, Paris : Moussard/Maradan, 1801.

Office de la langue française/Direction des services linguistiques. *Énoncés de politique sur les critères d'officialisation*, dans *Répertoire des avis terminologiques et linguistiques*, 4^e éd., Sainte-Foy : Les publications du Québec, p. 277-280, 1998.

Office québécois de la langue française. *Politique de l'officialisation linguistique*, Québec : Les publications du Québec, <http://www.oqlf.gouv.qc.ca/>

[ressources/bibliotheque/officialisation/politique_officialisation_20040305.pdf](http://www.oqlf.gouv.qc.ca/ressources/bibliotheque/officialisation/politique_officialisation_20040305.pdf), ©2001, 2003, 2004.

Office québécois de la langue française. *Politique de l'emprunt linguistique, document inédit*, Québec : Gouvernement du Québec, 2003.

Picone M.D. « L'impulsion synthétique. Le français poussé vers la synthèse par la technologie moderne », *Le Français moderne*, 2(59), p. 148-163, 1991.

Prairie M. *La francisation des entreprises : l'expérience vécue par des travailleurs et des travailleuses de la CSN et de la FTQ*, Montréal : UQAM/CSN/FTQ, 1986.

Pruvost J. et Sablayrolles J.-F. *Les néologismes*, coll. *Que sais-je?*, n° 3674, Paris : Presses universitaires de France, 2003.

Quirion J. *La mesure de l'implantation terminologique : proposition d'un protocole, étude terminométrique du domaine des transports au Québec*, coll. *Langues et sociétés*, n° 41, [Montréal] : Office québécois de la langue française, 2003.

René N. « Le rôle de l'Office de la langue française en matière de diffusion terminologique », dans *Implantation terminologique : actes d'une conférence*, Vitoria, 13 février 2001, Pays basque, [Montréal] : Office de la langue française, p. 1-11, [http://www.olf.gouv.qc.ca/RESSOURCES/bibliotheque/conferences/](http://www.olf.gouv.qc.ca/RESSOURCES/bibliotheque/conferences/index.html)

[index.html](http://www.olf.gouv.qc.ca/RESSOURCES/bibliotheque/conferences/index.html).

Sournia J.-C. « La vie d'un langage : la maladie des emprunts étrangers », dans *Langage médical moderne*, Comité d'études des termes, Conseil international de la langue française médicaux français, Paris : Hachette, p. 23-29, 1974.

Une terminologie normée pour la maintenance des moyens de production hydrauliques

Anne Dourgnon-Hanoune

EDF R&D
6 quai Watier
BP 49
78401 Chatou cedex
anne.dourgnon@edf.fr
<http://www.edf.fr>

Philippe Rouard

EDF Centre d'Ingénierie Hydraulique
15 avenue Lac du Bourget Passerelles
Savoie Technolac
73373 Le Bourget du Lac Cedex
philippe.rouard@edf.fr
<http://www.edf.fr>

Marie Calberg-Challot

Ontologos corp.
P.A.E. du Levray
6, route de Nanfray
74 960 Cran Gevrier
marie.calberg-challot@ontologos-corp.com
<http://www.ontologos-corp.com>

Résumé :

Nous présentons une étude réalisée avec le Centre d'ingénierie hydraulique (CIH) d'EDF. Dans ce travail, nous nous intéressons plus particulièrement au sous-domaine des turbines hydrauliques. Après avoir présenté le secteur d'activité et les motivations d'un tel projet, nous étudierons l'apport de la terminologie dans la mise en place d'une représentation ontologique. Nous verrons alors, sur l'exemple des turbines hydrauliques,

l'intérêt de constituer une terminologie normée ou formelle distincte mais néanmoins liée aux usages et pratiques de diverses communautés.

Mots-clés : connaissances, ontologie, terminologie, ontoterminologie, modèle formel, turbine hydraulique.

1. Une terminologie de référence des matériels de production hydrauliques

La production énergétique d'origine hydraulique s'appuie, en France, sur des aménagements hydrauliques très divers, tant sur le plan de la puissance produite (de 50 à 1800 mégawatt électriques) que sur le plan technique. Il existe une grande diversité d'aménagements hydroélectriques en fonction de leur situation géographique, du type de cours d'eau, de la hauteur de chute, de la nature du barrage et de sa géographie.

Ces aménagements hydrauliques peuvent être fort anciens - le barrage de Cusset date de 1890 - ou plus récents - le barrage de Grand'Maison date de 1988 - mais le principe de fonctionnement est identique. L'eau, captée puis amenée par des conduites forcées, entraîne une roue. L'alternateur transforme l'énergie mécanique de la roue en énergie électrique qui passe ensuite par le transformateur avant d'être mise sur le réseau. Le schéma de la Fig. 1 illustre ce principe.

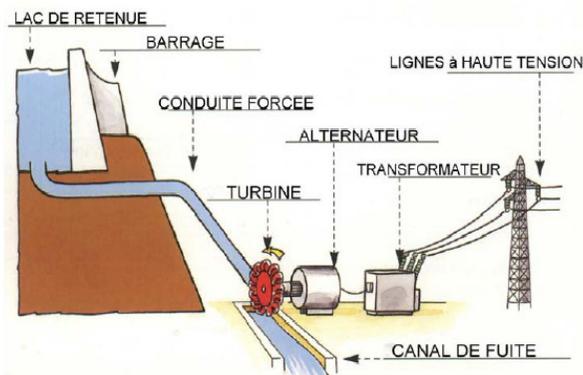


Fig. 7 : principe d'un aménagement hydraulique
(Source : www.musee-hydrelec.fr)

Ces aménagements hydrauliques sont télégérés depuis une trentaine d'années. Les personnels de conduite ne sont plus sur site et conduisent à distance les installations. Des équipes de maintenance et d'ingénierie interviennent également sur l'ensemble des aménagements. Des communautés d'exploitants, de techniciens et d'ingénieurs se sont constituées au fil des ans. Elles partagent un savoir mais les pratiques langagières peuvent différer d'une communauté de pratique à une autre, d'un site à un autre.

Si le principe est commun, les variations sont nombreuses et la gamme des matériels composant les centrales hydrauliques est très étendue.

La diversité des aménagements hydrauliques ainsi que les diverses communautés de pratique travaillant sur ces sites illustrent la diversité des pratiques langagières qui peuvent être rencontrées. La langue comprend des implicites et des variations d'usage que nous aborderons dans le but d'établir une terminologie commune pour les multiples matériels à maintenir qui composent une turbine hydraulique.

C'est donc une terminologie de référence au service de la Division Production et Ingénierie Hydraulique (DPIH) que nous nous proposons de réaliser dans l'objectif de pérenniser la connaissance des moyens de production hydrauliques.

2. Une terminologie au service de la production et de l'ingénierie Hydraulique

2.1. L'ingénierie hydraulique

Les aménagements hydrauliques exploités par EDF, également appelés « usines d'exploitation », sont aujourd'hui regroupés en cinq unités de production. Chaque unité dispose d'une ou plusieurs équipes opérationnelles de maintenance.

Le Centre d'ingénierie hydraulique (CIH) est l'un des deux centres d'ingénierie de la production hydraulique. Il assure, entre autres, les missions d'ingénierie pour le fonctionnement et la maintenance des matériels, le génie civil et diverses missions d'expertise. Dans le cadre du projet de modernisation de la maintenance, le CIH souhaite une terminologie cohérente des matériels maintenus. Cette terminologie doit correspondre aux matériels tels qu'ils ont été conçus, tels qu'ils sont exploités et tels qu'ils vont être maintenus. Elle s'appuie

donc sur une modélisation des moyens de production, depuis l'aménagement hydraulique jusqu'au niveau des matériels à maintenir.

Cette terminologie doit être un vecteur de compréhension et de transmission de la connaissance et du savoir entre les différentes communautés de pratique.

2.2. Notre méthode de travail

EDF R&D travaille depuis plusieurs années sur des projets de terminologie et d'ontologie dans le domaine de la production électrique d'origine nucléaire et hydraulique [Dougrnon-Hanoune, 2006]. Les travaux avec le CIH confirment la démarche ontologique déjà engagée par EDF R&D.

En effet, dans un domaine technique tel que celui de la production d'électricité et plus particulièrement celui de l'ingénierie hydraulique, l'opérationnalisation des terminologies à des fins de gestion de contenus et de capitalisation des connaissances est nécessaire afin de d'éviter des erreurs d'interprétations et leurs conséquences. Nous entendons ici par opérationnalisation, la conceptualisation d'objets ou de réalités au travers d'une terminologie normée et permettant leur désignation univoque [Million-Rousseau *et al.*, 2007].

Dans ce cadre, notre objectif est de détailler les matériels maintenus et exploités en vue de déterminer des libellés homogènes et partagés pour la terminologie des turbines hydrauliques.

3. La notion de turbine hydraulique

3.1. « Turbine hydraulique » ou « roue hydraulique » ?

Tout d'abord, il est intéressant de donner une définition du terme « turbine hydraulique ».

En consultant, dans un premier temps, un dictionnaire de langue générale comme le *Trésor de la langue française informatisé (TLFi)*, voici les premiers éléments d'informations que nous avons recueillis pour le terme « turbine hydraulique ».

« Dispositif rotatif destiné à utiliser la force d'un fluide et à transmettre le mouvement au moyen d'un arbre ». (TLFi consulté en date du 02-04-08).

En consultant ensuite une ressource terminologique spécialisée comme le *Grand dictionnaire terminologique (GDT)*, on commence à relever une certaine

variation de sens concernant le terme « turbine hydraulique » et il apparaît déjà clairement polysémique.

Nous trouvons ainsi comme définition « Partie tournante d'une machine recevant un fluide sous une certaine pression et transformant l'énergie de ce fluide en énergie mécanique » [...] et « Machine qui convertit l'énergie de l'eau courante ou d'un autre fluide en travail mécanique. Note : [...] Dans les usines hydroélectriques on trouve trois sortes de turbines hydrauliques : les roues de Pelton, les turbines Francis et Kaplan » (GDT consulté en date du 02-04-08).

Ces définitions s'accordent pour définir une « turbine hydraulique » comme un « dispositif » ou comme une « machine ». Ces définisseurs introduisent bien la notion « d'ensemble ». Mais lorsque l'on s'intéresse de plus près à la note fournie dans l'article « turbine hydraulique » du *Grand dictionnaire terminologique*, on relève que la turbine est soit une « roue », soit une « turbine ».

On trouve enfin dans un troisième article du *Grand dictionnaire terminologique* que le terme « turbine hydraulique » est synonyme de « roue hydraulique ».

La consultation de ces ressources lexicales met immédiatement en avant « qu'une unité lexicale donne naturellement lieu à la rencontre d'une vaste gamme de sens et emplois généraux et spécialisés, d'emplois concrets et abstraits » [Calberg-Challot *et al.*, 2007].

3.2. Le rôle des experts

Il nous semble pertinent de compléter les définitions précédemment citées par les connaissances des experts. En effet, « l'interrogation des spécialistes du domaine peut remplacer l'introspection du lexicographe » [Thoiron *et al.*, 1996]. Les experts décrivent la turbine hydraulique (Fig. 2) comme étant composée d'une amenée, d'un organe d'admission, d'une roue hydraulique, d'une évacuation, d'un palier, d'un arbre. Il est possible de la représenter de la façon suivante :

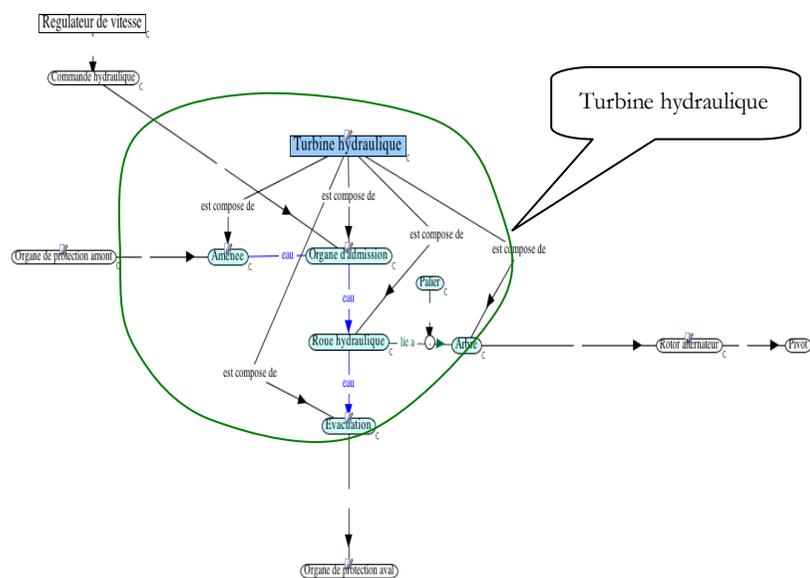


Fig. 8 : réseau conceptuel¹ d'une « turbine hydraulique »

De plus, quelques témoignages d'experts que nous avons recueillis viennent compléter nos connaissances :

« Un groupe pompe utilise en général une (roue) Francis, parfois une (roue) Pelton. En toute rigueur, n'importe quelle turbine (roue) peut être utilisée pour un groupe pompe. L'alternateur est utilisé comme moteur » (expert 1, 04-04-08).

« On parle aussi de roue pompe ? » « C'est un mode d'utilisation de la roue ». « Mais la roue pompe est-elle différente d'une (roue) Francis ? » « Pas vraiment, la pente des aubes est différente. » « (pas de pente pour une utilisation en (mode) pompe) » (expert 1, 04-04-08).

Pourquoi cette « confusion » entre la « turbine hydraulique » et la « roue hydraulique » alors que les experts savent parfaitement de quoi il s'agit ?

¹ Réalisé à l'aide de SNCW (Semantic network craft Workbench), éditeur de schémas (Ontologos corp.).

Comment définir alors les différentes turbines hydrauliques en exploitation sur le parc français ?

On ne peut pas se passer des connaissances et du savoir des experts. Pour se faire, le passage par la langue d'usage est obligatoire pour partager les connaissances et le savoir des experts.

Mais il semble important de sortir de la langue pour tenter de trouver un consensus entre les experts, ces variations d'usage et de pratiques entre les ingénieurs, les exploitants et les techniciens étant inhérentes à la langue et « il nous semble nécessaire de distinguer un niveau conceptuel et un niveau sémantique » [Thoiron *et al.*, 1996].

3.3. L'intérêt d'un recours aux schémas

Les ingénieurs ont transmis leurs connaissances à travers des schémas. Au-delà des mots, ils s'accordent sur le concept grâce au schéma qui l'explique. Comme l'a écrit Pierre Lerat à juste titre, « 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]. Sur le schéma suivant (Fig. 3), la distinction est bien faite, d'une part, entre une « roue Pelton » et un « groupe Pelton » et, d'autre part, entre un « groupe Pelton » et une « turbine » (qui ne comprend pas l'alternateur).

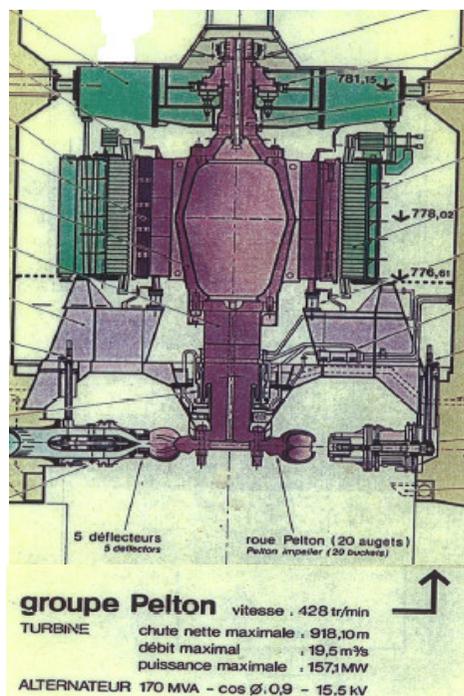


Fig. 3 : schéma d'un groupe Pelton (Source : EDF R&D)

Aujourd'hui, les cadres réglementaires ont changé et impliquent une production documentaire plus importante et plus pointue que par le passé. Ces documents font plus appel à des travaux d'experts qu'à des schémas de principe et sont donc plus fréquemment sujets à ambiguïtés. Ceci nous est confirmé par le témoignage suivant :

« Avant on avait des schémas et on se comprenait plus aisément. Maintenant on écrit et on a parfois du mal à se comprendre » (expert 2, 04-04-08).

Ce témoignage est explicite et il est donc essentiel de sortir de la langue naturelle pour se comprendre.

« 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, 2007] et « se référer à la conceptualisation

du domaine peut être une autre manière d'apporter des éléments de réponse » [Roche, 2007].

4. De la langue d'usage au langage normé

4.1. Termes normés²

« Ces termes normés, s'ils n'ont pas à être imposés, sont indispensables à la désignation du système notionnel. Ils participent également à l'identification et à la définition des termes d'usage » [Roche, 2007].

Comme nous l'avons écrit plus haut (Fig. 2), les hydrauliciens emploient le mot « turbine » à la fois pour désigner le « groupe hydraulique », la « turbine hydraulique » et la « roue hydraulique » qui la compose. Nous proposons alors de définir de façon formelle la turbine hydraulique comme étant composée d'une amenée, d'un organe d'admission, d'une roue hydraulique, d'une évacuation, d'un palier, d'un arbre tandis que le groupe hydraulique (Fig. 4) sera composé de la turbine hydraulique et de l'alternateur et de le représenter de la façon suivante :

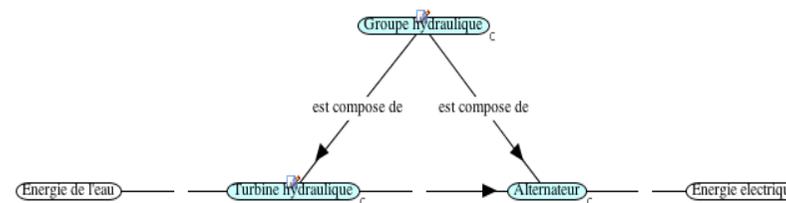


Fig. 4 : réseau conceptuel d'un « groupe hydraulique »

Le terme « turbine hydraulique » est fortement polysémique puisqu'il désigne aussi bien la « roue hydraulique » que le « groupe hydraulique ». Il faut en conserver l'usage, tout en maintenant les concepts auxquels il réfère. Ainsi, nous préconisons l'élaboration d'une terminologie normée extralinguistique, à laquelle se réfère la terminologie d'usage.

² Dans cette dernière partie, les expressions soulignées sont des expressions normées.

4.2. Définition formelle des concepts

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, p.226].

Les groupes hydrauliques sont caractérisés par la roue hydraulique qui les compose, cette dernière possédant de façon exclusive des augets, des aubes ou des pales. La figure 5 illustre une roue à augets.



Fig. 5 : roue à augets tournant avec le poids de l'eau (II^{ème} s. av. JC) (Source : www.musee-hydrelec.fr)

C'est donc la roue hydraulique, plus que le groupe hydraulique, que l'on peut identifier de cette façon.

En interrogeant les experts, nous savons que seules les turbines hydrauliques Pelton ont une roue à augets, que la roue des turbines hydraulique Francis est à aubes alors qu'elle a des pales dans le cas des turbines hydraulique Kaplan ou des turbines hydraulique à hélices. La distinction entre ces deux dernières s'opère sur le caractère fixe ou mobile des pales. Ces différences sont tellement évidentes qu'elles restent implicites pour les experts qui notent dans la documentation d'exploitation : « les trois principaux types de turbines sont les turbines Pelton, les turbines Francis, les turbines Kaplan ou turbines à hélice », « les turbines Kaplan sont des turbines à hélice » ou encore « les turbines Kaplan ressemblent à des turbines à hélice ».

Les trois premiers caractères distinctifs sont exclusifs :

$$\text{roue à augets} \oplus^3 \text{roue à aubes} \oplus \text{roue à pales}$$

³ \oplus signifie "ou exclusif".

De plus,

les pales d'une roue sont fixes \oplus les pales d'une roue sont mobiles

	roue à augets	roue à aubes	roue à pales	roue à pales mobiles	roue à pales fixes
Roue Pelton	X				
Roue Francis		X			
Roue Kaplan			X	X	
Roue à hélice			X		X

Tab. 1 : caractères distinctifs des roues hydrauliques

Cette différenciation exclusive sous-entend bien évidemment que si un critère est présent, les autres sont absents (Tab.1). L'ensemble des connaissances implicites se représente dans l'arbre suivant où l'on comprend qu'une roue hydraulique Pelton est à augets, sans pales et sans aubes (Fig. 6).

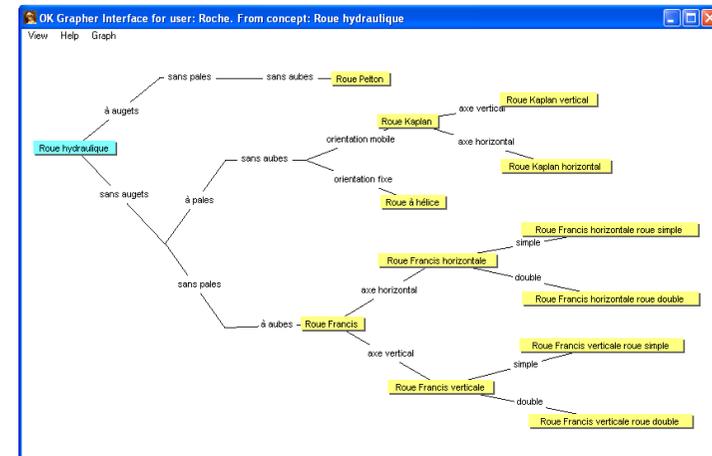


Fig. 6 : représentation conceptuelle⁴ d'une « roue hydraulique »

⁴ Construite à l'aide d'OCW (Ontology craft workbench), éditeur d'ontologies par différenciation spécifique (Ontologos corp.).

Cet arbre peut être vu comme l'arbre des possibilités déduites des différences exclusives. Dans la mesure où l'on s'entend sur la définition d'une roue hydraulique, il permet de définir les différentes roues de façon formelle. Les différences exclusives expriment des prédicats au sens de la logique. Une roue hydraulique Kaplan à axe vertical est, définie par l'ensemble de ses différences :

roue hydraulique Kaplan à axe vertical
 $\equiv \{ \text{axe vertical, orientation mobile, à pales, roue hydraulique} \}$

Les ambiguïtés de la langue sont levées et les connaissances implicites (la roue hydraulique Kaplan est sans augets et sans aubes) sont déduites de la présence de la différence exclusive « à pales ».

Nous pouvons ainsi donner une définition normée de la turbine hydraulique en disant que la turbine hydraulique Kaplan est une turbine hydraulique dont la roue hydraulique est une roue hydraulique Kaplan (Fig. 7),

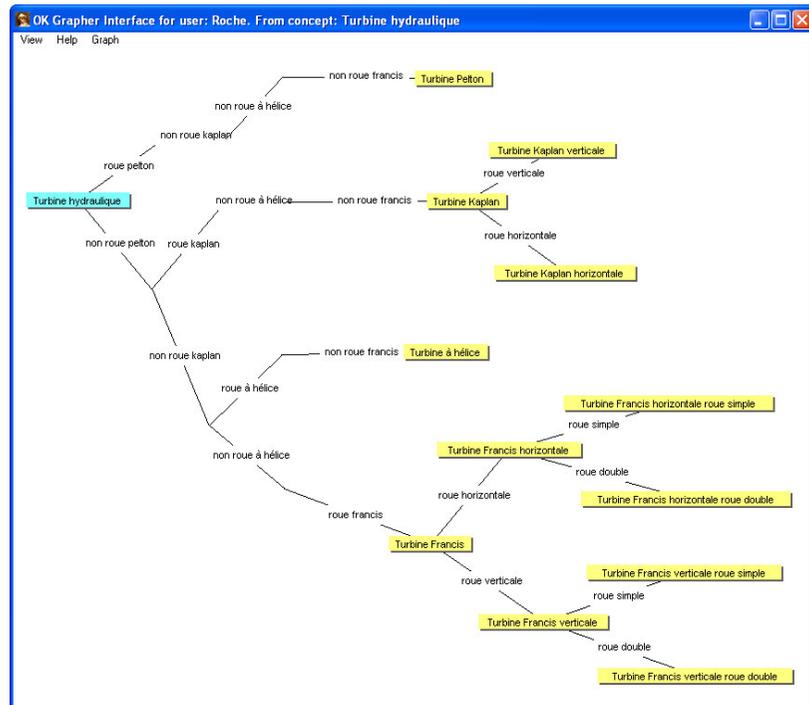


Fig. 7 : représentation conceptuelle d'une « turbine hydraulique »

et enfin déduire le groupe hydraulique en donnant pour définition normée, qu'un groupe hydraulique Kaplan est un groupe hydraulique dont la turbine hydraulique est une turbine hydraulique Kaplan (Fig. 8).

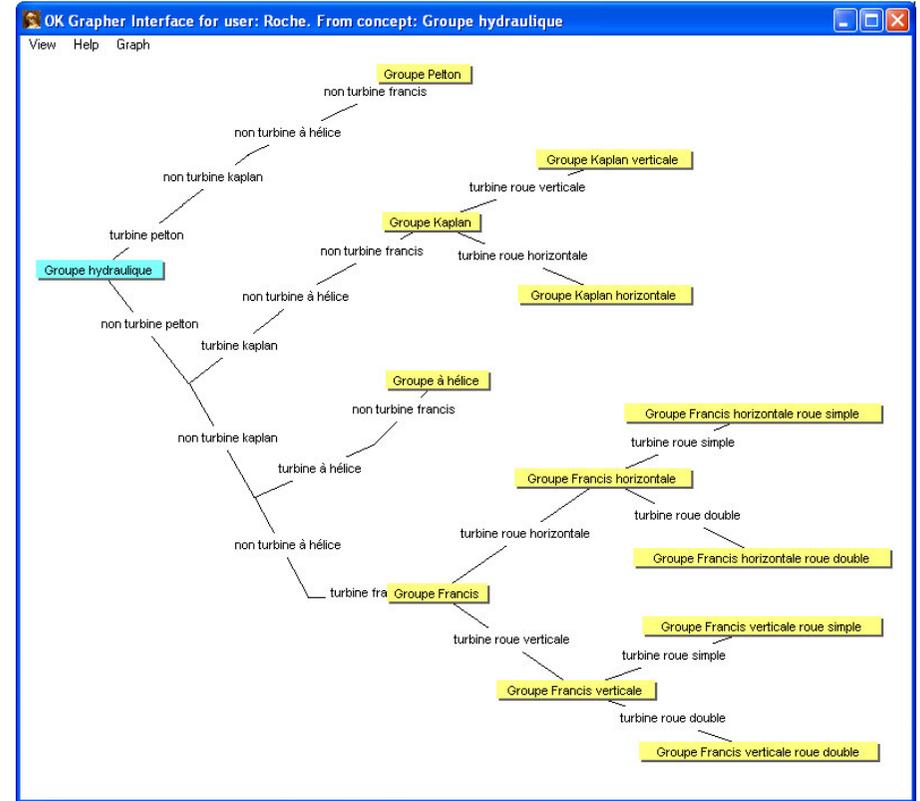


Fig. 8 : représentation conceptuelle d'un « groupe hydraulique »

5. Conclusion et perspectives

Nous avons tenté de montrer le besoin d'une terminologie formelle, base d'un système notionnel ou ontologique, et distincte de toute terminologie d'usage, afin de mettre en place la description d'un domaine, tel que celui des turbines hydrauliques, de façon pertinente et adaptée à l'ensemble des communautés de pratiques.

En effet, de notre point de vue, la démarche ontoterminologique que nous avons présentée au travers d'un exemple naît de la recherche d'une représentation commune d'une réalité face aux variations d'usage des différentes communautés de pratiques en interaction avec cette réalité.

Dans cette recherche d'une représentation commune, le concept correspondant à une réalité nous est apparu comme l'invariant pour le domaine de spécialité considéré sur lequel une description computationnelle peut être basée.

La constitution d'une terminologie formelle de ces concepts, distincte des terminologies d'usage apparaît alors comme la base pour la mise en place informatique d'un système ontologique relié et commun aux différents usages et par conséquent utilisables par tous.

6. Remerciements

Nous tenons à remercier les personnes consultées lors du projet ainsi que les experts qui ont participé à ce travail, et tout particulièrement Marc Delort, Patrick Laurier, Cédric Bernardi et Jean-Louis Ballester. Nos remerciements vont également à Alain Franco-Rondisson et à Philippe Miguel.

Bibliographie

Calberg-Challot M., Candel D. & Roche C. « 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*, ed. Christophe Roche, Annecy, Institut Porphyre, pp. 199-141, 2007

Dourgnon-Hanoune A., Salain P., Roche C. « *Ontology for long-term knowledge* », *LAE/AIE'06*, 2006

Dourgnon-Hanoune A., Mercier-Laurent E., Roche C. « *How to value and transmit nuclear industry long term Knowledge* », *ICEIS 06*, 2006

Kocourek R., *La langue française de la technique et de la science*, Brandstetter Verlag, 1991

Le Grand dictionnaire terminologique,
www.granddictionnaire.com/btml/fra/r_motclef/index1024_1.asp

Lerat P., *Les langues spécialisées*, PUF, 1995

Le Trésor de la langue française, version informatisée (TLFi), atilf.atilf.fr/tlfv3.htm

Million-Rousseau C., Damas L. & Pralon S., *Construction d'ontologies : retour d'expériences*", *Actes de la conférence Terminologie et ontologie : descriptions du réel*, Société française de terminologie, pp.11-125, 2007

Roche C. « *Le terme et le concept : fondements d'une ontoterminologie* », *Actes de la première conférence TOTh 2007, Terminologie & Ontologie : Théories et applications*, ed. Christophe Roche, Annecy, Institut Porphyre, pp. 1-22, 2007

Roche C. « *Dire n'est pas concevoir* », *IC 2007 : 18^e Journées Francophones d'Ingénierie des Connaissances*, Grenoble 2-6 juillet 2007

Thesaurus français-anglais, Département Systèmes d'information et de documentation, EDF R&D, 1997

Thoiron P., Arnaud P., Béjoint H. et Boisson C. P. « *Notion 'd'archi-concept' et dénomination* », *Meta*, vol. 41, N°4, pp. 512-524, 1996

Ontologie franco / anglaise du domaine informatique comme accès à un corpus de textes scientifiques

Gérald Kembellec

Equipe CITU-Paragraphe – Laboratoire Paragraphe

Université Paris 8

2, rue de la liberté

93260 Saint-Denis

gerald.kembellec@univ-paris8.fr

http://geka.ec-10.eu

Résumé :

Cet article présente une recherche visant à transformer un modèle d'organisation représentatif du domaine informatique en outil de recherche navigable intuitivement par un initié.

Mots-clés : Bibliothèque numérique, Ontologie de domaine, KBS.

1. Introduction

Dans le contexte universitaire français, il est courant d'observer des étudiants de 2^{ème} et 3^{ème} cycles éprouver de réelles difficultés à rassembler de la documentation sur leur domaine d'études ou de recherches. Dans notre optique, l'idée générale consiste à soutenir ces apprenants par l'outil, à suppléer leur perception par de la connaissance de domaine « stockée » dans l'ontologie. Il est à envisager, à espérer même, qu'à terme la maîtrise de l'outil le rende obsolète, car source intrinsèque de connaissance. Dans notre contexte, il s'agit de recherche bibliographique dans le domaine informatique pour des personnes « initiées » mais non expertes, qui de plus sont perdues dans un corpus majoritairement anglophone.

Nous avons entamé notre recherche par une réflexion sur la pertinence de projeter le concept d'ontologie de domaine sur les notions de portail et de moteur de recherche. Cette démarche tend à générer une interface

homme/machine (IHM) afin d'améliorer l'approche utilisateur (le chercheur) de manière transparente et intuitive. La question que soulève cette réflexion est l'influence de la représentation de l'accès à l'information sur la recherche de connaissance. Son impact sur les résultats obtenus par rapport aux recherches plus traditionnelles est-il significatif ?

L'objectif final de notre travail de recherche est de mettre au point un système incrémental, idéalement autonome, d'indexation de documents scientifiques. Cette méthode s'appuie sur l'extraction des mots clés et le positionnement d'articles dans une ontologie de domaine informatique. L'autonomie du système serait un facteur non négligeable de réduction de coût, mais surtout un gain de temps. En effet cela éviterait à un groupe d'experts du domaine une veille technologique sans fin. Nous échapperions à la fatalité que chaque soubresaut technique ou idéologique provoquerait des débats sur l'opportunité d'indexation du nouveau concept à un emplacement donné de l'ontologie. Notre travail passe par la collecte des articles scientifiques relatifs à l'informatique puis par une intégration à un corpus de documentation. Cette documentation sera représentée sous la forme d'une arborescence. Cela permettra l'émergence d'une visualisation globale du corpus de textes de la recherche informatique. Cette approche consentira également un accès facilité à l'information recherchée par moteur de recherche en langage naturel, par mots clés, contextualité, ou proximité sémantique. Ainsi, et c'est tout l'intérêt du concept, un utilisateur qui ne maîtrise pas encore l'ensemble du vocabulaire informatique (et la langue anglaise) pourrait trouver des articles pertinents en plusieurs langues, articles qu'il n'aurait pas su trouver seul par des méthodes de recherche traditionnelles.

Le présent article propose de mettre en œuvre la première partie de cette tâche, à savoir la construction de l'ontologie, le système de navigation pour la parcourir et un système de recherche dans un corpus scientifique.

2. Etat de l'art de la recherche d'informations par ontologies de domaine

Par ontologie de domaine nous entendons un ensemble de concepts hiérarchisés par un expert au sein d'une structure, et liés par des relations de proximité syntaxique ou sémantique.

La démarche classique d'utilisation des ontologies de domaine consiste à hiérarchiser les sous ensembles du domaine dans une optique de gestion. L'ontologie sert alors le plus souvent à hiérarchiser et classer les éléments composant le domaine ainsi qu'à décrire leurs relations. Une application courante est l'indexation de corpus spécialisé par ce biais.

Une utilisation plus novatrice de l'ontologie est d'inverser la démarche. Il est possible d'utiliser l'ontologie de domaine comme support de recherche dans un texte, un corpus, une bibliothèque numérique, ou même à l'extrême l'Internet.

Grâce à une combinaison de différentes technologies sémantiques, Stephan Bloehdorn a proposé une méthode intéressante de consultation de bibliothèques numériques [Bloehdorn *et al.* 2007]. Il a défini une approche par analyse de questions *structurées* en langage naturel avec une grammaire définie. Il s'agit pour le système de comprendre la question d'identifier les mots clés, les titres et les auteurs. Par exemple qui a écrit tel livre? Quel livre traite d'un sujet défini? Quel article fait partie de telle conférence et correspond à tels mots clés? Cette approche traduit le langage naturel en métadonnées, et reformule la question en langage SPARQL¹. Comme la réponse se trouve dans un fichier *Resource Description Framework* (RDF²), les mises à jour en temps réel sont supportées, ainsi que l'hétérogénéité des formats et la location des ressources. Cette méthode permet de s'abstraire de toute base de données au sens commun du terme.

3. Ontologie de domaine informatique, conception d'un modèle exploitable

Au départ, l'approche sera onomasiologique ou *top-down*, c'est-à-dire que le corpus sera classé à la volée dans une structure qui est donc un ensemble normalisé et fini. Ensuite nous ambitionnons d'enrichir éventuellement la structure si les corpus ajoutés en font émerger le besoin. L'ontologie de domaine est composée d'une arborescence de sujets allant d'une racine générique : le domaine (ici l'informatique), vers des feuilles de connaissance. Les arcs seront des relations de spécification / généralisation ou des liens de similarité. L'ontologie ne contient pas les articles, mais des mots clés dont l'héritage se fait de manière

¹ <http://www.w3.org/TR/rdf-sparql-query/>

² <http://www.w3.org/RDF/>

top-down (descendante), et qui permettent de générer une requête qui sera passée aux principales bibliothèques scientifiques en ligne.

Cette arborescence constitue le squelette externe ou *exo-squelette* du domaine, dont les premiers mots clés sont les mots constituant les intitulés des nœuds et des feuilles. Ces mots clés seront dits « natifs », par opposition aux autres mots clés ajoutés *a posteriori*, qui seront dits « ajoutés ».

3.1. Notion de pertinence utilisateur

L'informatique est un domaine très vaste, comprenant une multitude de sous-disciplines, et un outil puissant usité dans de nombreux domaines scientifiques. Il faudra donc autant que possible s'imprégner de *points de vues* pour la recherche, saisir le contexte d'étude de l'utilisateur.

Exemple : le terme de « stockage de données » n'aura pas le même sens pour un technicien en assemblage, un ingénieur système ou un documentaliste. Pour le technicien la représentation qui s'impose du stockage de données est le disque dur ou la clé USB. Le professionnel système et réseau aura lui une vision plus large de « stockage de données ». Il verra les concepts de périphériques mais aussi les méthodes de stockages tels les NAS, les redondances de données (niveau de RAID), la façon dont les informations sont partagées (Netbios, NFS, SMB³...) mais aussi les droits sur les données (lecture, écriture et exécution). Enfin, le documentaliste verra en ce terme principalement un progiciel de SIGB (Système intégré de gestion de bibliothèque) qui gère les prêts, les réservations, suivi des commandes ou encore l'état des livres. Ces trois professionnels, pointus en leur domaine font un usage différent du terme « stockage de données », cependant on ne peut pas parler ici de polysémie, mais plutôt de point de vue.

La question de la pertinence utilisateur se pose dans ce cas précis de la SRI. Cette observation a grandement influencé l'outil, en l'axant sur l'utilisateur et pas uniquement sur les données. Ce projet doit être une entité à l'utilisation souple, se mettant à la portée de l'utilisateur pour l'aider à maîtriser son domaine de connaissance.

³ Divers protocoles de partage de données en réseau

3.2. Perspectives d'évolution de l'ontologie

Par la suite, lors de la phase d'indexation de corpus, si un article semble « inclassable », nous proposons momentanément de le classer au plus proche dans une des branches temporaires de l'ontologie, comme *miscellaneous* ou *general*. Puis une fois une taille suffisante atteinte il conviendra de les classer définitivement en créant une ou des nouvelle(s) branche(s) à l'ontologie là où la proximité sémantique est la plus forte. Il s'agit d'un des vecteurs de l'évolution de l'ontologie, qui n'est pas statique mais évolue avec le corpus et le travail des usagers et experts.

Les extensions qui pourront être ajoutées à l'ontologie doivent être anticipées lors de son élaboration. Il doit être possible d'ajouter de nouveaux concepts sans avoir à toucher aux fondations de l'ontologie. Par exemple dans la branche ayant le plus de mots clés en commun ferait une racine convenable. Peut être même suffirait il de nommer la branche en résumant les concepts de l'article en un label en langage naturel.

4. Méthodes de recherches proposées et présomptions de modèles exploitables

Mots clés, Langage naturel, parcours de graphe de domaine, proximité sémantique. Nous sommes partis sur la définition du domaine de recherche avec un *exo-squelette* ontologique minimal. Cette partie du travail nécessite de trouver des approches taxonomiques représentant le plus finement et le plus exhaustivement possible le vaste domaine de l'informatique. Ensuite, pour conceptualiser ce domaine il va être nécessaire de segmenter les intitulés de chaque branche. Cette phase de spécification passe par une étape de construction des « grappes » de mots clés relatifs à chaque branche, grâce aux lemmes des mots extraits des intitulés.

D'un point de vue technique, pour une plus grande facilité de manipulation, nous intégrerons l'ontologie et ses mots clés dans une base de données, ce qui permettra de traduire de manière complète l'ontologie en *Extensible Markup Language* (XML⁴) en tenant compte de ses évolutions en temps réel.

Pour notre phase de test, le corpus de recherche sera composé des intitulés d'articles parus depuis 1945 et référencés dans la *DataBase systems and Logic*

⁴ <http://www.w3.org/XML/>

Programming (DBLP) par Michael Ley⁵ de l'Université allemande de Trier. Il s'agit à l'origine d'un document XML d'environ un million d'entrées au format BibTEX⁶ (format de description bibliographique de LaTeX⁷). Notons que les papiers sont rédigés dans diverses langues. Nous proposerons également des méta-requêtes vers les bibliothèques Computer Science Bibliography (CSBIB)⁸ et Association for Computing Machinery (ACM).

5. Construction du modèle

5.1. Choix d'une référence de classification informatique

D'un point de vue technique, pour une plus grande facilité de manipulation, nous intégrerons l'ontologie et ses mots clés dans une base de données, ce qui permettra de traduire de manière complète l'ontologie en XML en tenant compte de ses évolutions en temps réel. Nous allons dans un premier temps chercher un organisme spécialiste des questions informatiques proposant un système de représentation du domaine que nous ambitionnons de modéliser. Pour des raisons de simplicité, l'encyclopédie en ligne Wikipédia se détache dans un premier temps. En effet le domaine informatique y est classé selon une hiérarchie interne et un corpus abondamment pourvu est immédiatement disponible en XML et RDF. Cependant, à l'heure actuelle la caution scientifique de Wikipédia n'est pas démontrable. Nous allons donc nous concentrer sur la Computing Classification System⁹, dont la légitimité n'est plus à prouver. De plus, pour ne rien gâcher l'ACM possède sa propre bibliothèque numérique d'articles scientifiques, eux-mêmes indexés selon le modèle Computing Classification System (CCS).

Dans le contexte, le CCS n'est pas exploitable en l'état. Le CCS semble *a priori* plus être plus une taxonomie qu'une ontologie. Dans une taxonomie, le vocabulaire est organisé sous une forme hiérarchique. Cette hiérarchisation

correspond souvent à une spécification. Une taxonomie est une forme d'ontologie dont la grammaire n'a pas été formalisée. Dans le CCS cette grammaire a été définie par des rapports clairs de descendance et d'ascendance mais aussi des liens transversaux de proximité sémantique. Ainsi on peut un lien de ce type entre B.8 *Performance and reliability* et C.4 *Performance of systems* (cf. Figure 1).

Cependant, d'après Gruber, un des aspects importants d'une ontologie (en sus de la clarté, de la cohérence, d'un engagement minimal, et de la déformation) est l'extensibilité [Gruber 1993]. Il convient donc d'effectuer un traitement pour permettre d'anticiper les évolutions de l'ontologie. En effet le système d'identifiant du CCS ne s'applique qu'aux nœuds et pas aux feuilles. Cela empêche de conserver l'esprit de référencement si pratique proposé par l'ACM en cas de spécialisation d'une feuille. On ne peut pas imaginer une relation de spécification étendant un élément non référencé. C'est pourquoi nous choisissons de donner de manière arbitraire un identifiant aux feuilles pour les transformer en nœuds potentiels. Par respect pour le travail initial et pour distinguer nos évolutions du travail initial, nous avons choisi de donner comme identifiant des feuilles l'identifiant du père auquel s'ajoutera une lettre de l'alphabet. Notons que la notation CCS utilisant déjà le « m » pour *divers/miscellaneous* et le « g » pour *général/general*, nous avons ôté ces deux lettres de notre processus d'identification des nœuds et feuilles.

⁵ <http://dblp.uni-trier.de/>

⁶ <http://www.bibtex.org/>

⁷ <http://www.latex-project.org/>

⁸ <http://liinwww.ira.uka.de/bibliography/index.html>

⁹ CCS de l'ACM <http://www.acm.org/class/1998/>



Top Two Levels of The ACM Computing Classification System (1998)

- [A. General Literature](#)
 - [A.0 GENERAL](#)
 - A.1 INTRODUCTORY AND SURVEY
 - A.2 REFERENCE (e.g., dictionaries, encyclopedias, glossaries)
 - A.m MISCELLANEOUS
- [B. Hardware](#)
 - B.0 GENERAL
 - [B.1 CONTROL STRUCTURES AND MICROPROGRAMMING \(D.3.2\)](#)
 - [B.2 ARITHMETIC AND LOGIC STRUCTURES](#)
 - [B.3 MEMORY STRUCTURES](#)
 - [B.4 INPUT/OUTPUT AND DATA COMMUNICATIONS](#)
 - [B.5 REGISTER-TRANSFER-LEVEL IMPLEMENTATION](#)
 - [B.6 LOGIC DESIGN](#)
 - [B.7 INTEGRATED CIRCUITS](#)
 - [B.8 PERFORMANCE AND RELIABILITY](#) NEW! [\(C.4\)](#)
 - [B.m MISCELLANEOUS](#)
- [C. Computer Systems Organization](#)
 - [C.0 GENERAL](#)
 - [C.1 PROCESSOR ARCHITECTURES](#)
 - [C.2 COMPUTER-COMMUNICATION NETWORKS](#)
 - [C.3 SPECIAL-PURPOSE AND APPLICATION-BASED SYSTEMS \(J.7\)](#)
 - [C.4 PERFORMANCE OF SYSTEMS](#)
 - [C.5 COMPUTER SYSTEM IMPLEMENTATION](#)
 - C.m MISCELLANEOUS
- [D. Software](#)
 - D.0 GENERAL
 - [D.1 PROGRAMMING TECHNIQUES \(E\)](#)
 - [D.2 SOFTWARE ENGINEERING \(K.6.3\)](#)
 - [D.3 PROGRAMMING LANGUAGES](#)
 - [D.4 OPERATING SYSTEMS \(C\)](#)
 - [D.m MISCELLANEOUS](#)
- [E. Data](#)
 - E.0 GENERAL

Figure 9 : CSS d'ACM

5.2. Travail sur l'ontologie en français

Notons que l'ontologie devra être entièrement traduite en français, mais que les mots anglais ne devront pas être lemmatisés. Il est donc nécessaire de faire le distinguo de la langue pour les intitulés. Par exemple NFS, l'acronyme de « Network File System » n'a pas d'équivalent français, le concept sera traduit dans la version française par « Système de fichiers en réseau » pour donner un aperçu du concept. Cependant, l'emploi de l'acronyme sera à n'en pas douter utilisé dans les articles. L'informatique étant une science majoritairement anglophone, les articles français comporteront de toute façon par défaut des mots français et anglais. Il serait d'ailleurs judicieux de proposer des relations de synonymie entre

des mots étant indifféremment employés en anglais ou en français par les professionnels, les enseignants et les chercheurs spécialisés dans le domaine. Cependant en toute objectivité, ce travail entraîne nécessairement une spécification de la conceptualisation du domaine. Il y aura inévitablement des inexactitudes, qui seront corrigées *a posteriori*.

5.3. Traduction de l'ontologie en français

D'après la lettre ouverte à l'Agence d'évaluation de la recherche et de l'enseignement supérieur (AERES) par quelques milliers de chercheurs français, il est largement admis que la *lingua franca* de la recherche scientifique est aujourd'hui l'anglais. Pourquoi traduire les intitulés des branches de l'ontologie en français alors que le corpus est majoritairement anglais, la langue scientifique? Nous attirons l'attention sur le fait que si l'utilisateur final maîtrise peut être la lecture de textes techniques et scientifiques, il peut se sentir plus à l'aise en français pour effectuer sa recherche, quitte à lire les articles en anglais avec un bon dictionnaire sous la main.

Le choix le plus simple et le plus économique pour automatiser une traduction anglo-française est l'utilisation d'un outil de traduction en ligne. Les outils qui ont attiré notre attention ont été BabelFish de Yahoo et Google translate de la suite Google. Nous avons conçu et utilisé une Interface de Programmation Applicative (API) de *wrapping*¹⁰ pour générer une version française de l'ontologie basée sur un de ces outils. Notons au passage que ce type d'application en ligne gagnerait à posséder sa propre API.

Une fois cette étape terminée, nous avons rapidement compris que rien ne remplace une traduction manuelle, c'est pourquoi nous intégrons une notion de *folksonomy* par flux RDF Site Summary¹¹ (RSS) dans l'outil, cela permettra à l'utilisateur final de signaler une erreur de traduction, ou une imprécision, au comité de gestion. Ce groupe sera formé des chercheurs des laboratoires de l'unité de recherche et de formation et validera ou non la proposition. Selon Thomas Vander Wal, la valeur du marquage extérieur de la *folksonomy* vient des usagers en utilisant leurs propres mots ce qui ajoute une dimension explicite, qui va être une inférence de l'objet [Vander Wal 2006]. Le système de nommage

¹⁰ Technique qui consiste à créer un programme informatique permettant à deux autres programmes de communiquer.

¹¹ <http://web.resource.org/rss/1.0/>

français des nœuds de l'ontologie automatisé dans un premier temps, poursuivi et développé par les utilisateurs anglophones sera validé par des experts au besoin sans avoir eu recours à un traducteur professionnel, ni mobilisé un expert à plein temps. Cette procédure permet un évident gain de temps pour les chercheurs du groupe et une économie financière non négligeable.

L'aspect technique de cette démarche devra être simplifié au maximum pour l'utilisateur afin de ne pas le décourager de faire une proposition. L'opération ne doit également pas lui prendre plus de quelques secondes. Une fois la proposition faite, un flux RSS est généré et restera actif jusqu'à vérification par au moins deux membres du comité. Nous ambitionnons ainsi de corriger la partie française de l'ontologie sur une période de temps encore indéterminée.

Le procédé permet aussi de tenir compte des mutations terminologiques inhérentes aux évolutions du domaine *Information Technology* (IT). L'utilisateur final bénéficie grâce à son interaction avec le système d'un enrichissement de sa connaissance du pôle de connaissances tout en participant à son évolution.

5.4. La génération des mots clés et l'émergence de proximité sémantique

Considérons le corpus formé par les intitulés composant l'ontologie de domaine IT du point de vue de l'infométrie statistique. Selon Le Coadic [Le Coadic 2006], si l'on considère un ensemble d'articles scientifiques, il faut s'intéresser aux mots significatifs et à leur cooccurrence pour dégager des proximités sémantiques significatives. Ainsi lorsqu'un *n-uplet* de mots associés apparaît simultanément dans plusieurs intitulés de noeuds, il est probable que les sujets traités soient associés. Bien sûr, dans le cas précis nous n'utiliserons cette approche que sur des intitulés, mais gageons que les labels ACM sont suffisamment précis pour être représentatifs de l'ensemble des articles, tant du point de vue général, que particulier.

Ainsi, les mots les plus représentatifs du label seront ajoutés comme mots clés de l'article et de la branche, les autres notés dans la proximité sémantique. Ultérieurement, lors de la phase d'indexation d'une bibliothèque numérique, si un article semble pouvoir être indexé à deux endroits, il sera proposé de créer un lien de proximité entre deux branches de l'ontologie.

5.5. L'interface avec le corpus

Nous tenterons à terme une méthode compilation (ou *clustering*) de différentes bases d'articles en ligne comme CSBIB, DBLP, ACM entre autres... Le *clustering* passera par une phase de prétraitement. Chaque bibliothèque scientifique possédant sa propre interface d'interrogation, nous allons essayer de trouver le document RDF qui décrit chaque base. Notons que si chaque site de ce type fournissait des services de description de données comme RDFa¹², cette démarche serait grandement simplifiée.

Une base de données décrivant « la bibliothèque scientifique du domaine informatique » sera constituée, décrivant chaque article par son titre, son contexte de publication, son année et ses auteurs. Cette base de données, mise à jour automatiquement chaque semaine de manière incrémentale, permettrait idéalement de générer à la volée un document unique RDF décrivant le pseudo corpus. C'est sur ce document que s'appuieront les recherches internes à l'outil.

Le terme à la volée signifie qu'en théorie pour chaque requête, un cliché du corpus sera constitué en RDF par interrogation de la base et traité pour tenir compte des mises à jour hebdomadaires. Cette démarche, bien que souhaitable est techniquement utopique. Il est tout de même possible, voire souhaitable au vu de la masse de données (dans une optique d'économie de ressource système) de conserver un cliché en cache. Ce cliché de la base deviendrait fichier RDF et serait alors la représentation du pseudo corpus.

Le corpus d'articles scientifiques n'est pas hébergé localement sur la machine hôte de l'ontologie pour des raisons légales, mais également de capacité de stockage. C'est pour cette raison que nous préférons dans le contexte utiliser le terme de pseudo-corpus plutôt que corpus. En effet les labels, et éventuellement les résumés indexés dans les bibliothèques numériques ne constituent pas à proprement parler un corpus.

5.6. Choix d'un type de représentation

Pour rendre le corpus plus accessible, nous optons de faciliter la représentation de l'ontologie de domaine sous la forme d'une carte navigable. L'arborescence doit permettre un focus sur la branche contenant une formalisation du concept recherché.

¹² <http://www.w3.org/TR/xhtml-rdfa-primer/>

Il existe un certain nombre de manières de visualiser les ontologies, mais toutes ne sont pas propres à la navigation, en tous cas pas à une navigation intuitive. Dans notre contexte, l'outil de représentation doit se conformer à un certain nombre de règles exposées par Christophe Tricot et Christophe Roche [Tricot *et al.* 2007] suite à un certain nombre d'observations. A minima, pour être efficace le système de visualisation doit respecter les règles suivantes :

- Offrir une vue globale de l'ontologie. Cela permettra à l'utilisateur d'identifier tous les concepts du domaine.

- Utiliser une approche "focus + contexte" pour permettre à l'utilisateur de se concentrer sur certains éléments tout en ayant accès aux autres;

- Utiliser la géométrie plane, pour éviter de déranger la perception naturelle de manipulation dans le plan. Ce point en particulier, ne sera pas respecté, car au vu de la masse de données à afficher et de la volonté de respecter les points précédents, il est complexe, voire impossible de combiner un affichage en arborescence et la géométrie euclidienne.

Du retour d'expérience de C. Tricot, nous noterons également qu'il émerge deux types d'utilisateurs: "novices" et "experts". Les novices comprennent le domaine et ses concepts sans pour autant saisir la finesse de l'organisation et les interactions. Les experts quant à eux saisissent parfaitement la globalité du domaine tant du point de vue des concepts que des rapports qui les lient. Dans notre contexte d'utilisation, les utilisateurs ont un profil qui peut être hybride pour un étudiant en MASTER ou un docteur qui se renseigne sur un sujet transversal à ses travaux, jusqu'à un profil « expert » pour un spécialiste de domaine. Nous essayerons donc de trouver un compromis de représentation du domaine offrant des accès directs au contexte sur l'élément en focus. Dans l'article de C. Tricot, il semble que modèle de représentation par *radial tree* soit le plus approprié pour des experts et l'*eye tree* pour des novices.

La visualisation en *eye tree* permet une vision globale du domaine ainsi que la possibilité d'un grand angle focalisé (*fisheye polar*) sur un point de détail autour duquel s'articule le domaine. Son principal défaut, dans le contexte, est de se borner au plan ce qui empêche la mise en perspective autorisée par les *cones trees*. Le *radial tree* est assez similaire à l'*eye tree* combinant la vision globale du domaine et le *fisheye polar*. Cependant une plus grande place est faite au contexte et au focus au sein même du graphe. Il semble que ce qui fait l'intérêt du *radial tree* (le focus + contexte) cause également une perte de contact avec l'objectif premier qui est de

conserver la vue globale. De plus, un *radial tree* décrivant l'ACM serait parfaitement illisible du fait même de la taille de l'ontologie.

Au vu de ladite taille, une visualisation par «grappe d'informations» émerge grâce à la combinaison d'ontologie et de la technologie dite *Topic Mapper*, grâce à l'applet open source hypergraph¹³. Bien que n'étant pas spécialement préconisé pour représenter efficacement une ontologie, le *Topic Mapper* est une représentation de type *hyperbolic tree* qui consiste à cartographier l'ontologie et d'y naviguer à volonté. Nous allons l'adapter pour faire émerger des points de vues, mais aussi des focus avec leurs contextes. Il s'agira ainsi d'une approche hybride entre l'*eye tree* et l'*hyperbolic tree*.

5.7. Génération d'une méta-requête naviguée

La première étape de génération de requête est le filtrage du «bruit» sur le label grâce à des *stop-lists* (une dans chaque langue) qui va éliminer les mots vides comme les articles, les pronoms, ainsi que les substantifs trop communs pour avoir un sens significatif lors du positionnement de l'utilisateur dans le navigateur de l'ontologie. Une étape similaire préalable est effectuée lors de l'utilisateur du moteur de recherche en langue naturelle.

La deuxième étape consiste à une lemmatisation des mots français, puis un calcul de proximité statistique de l'ensemble des mots dégagés avec la grappe de mots clés d'une des branches de l'ontologie. Peut être est-il judicieux de donner une valeur aux mots clés dans le contexte (arc valué)? Ce point sera une perspective à approfondir.

¹³ <http://sourceforge.net/projects/hypergraph/>

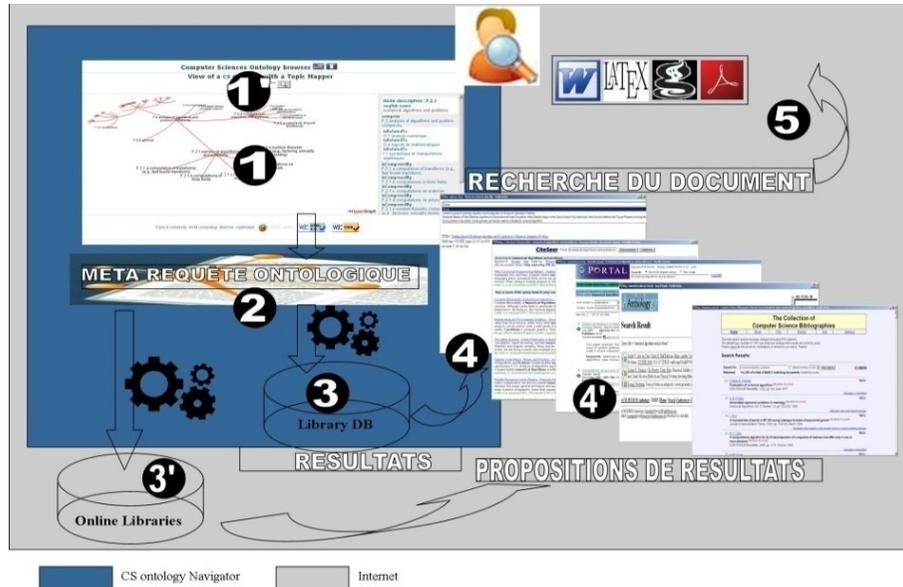


Figure 10 : modélisation conceptuelle

Description récapitulative:

- 1 Et 1' : Possibilité de se positionner dans l'ontologie par navigation ou par une requête en langage naturel.
- 2 Le positionnement permet de cerner un point de vue utilisateur et des centres d'intérêts,
- 3 Et 3.' : ce qui va dégager des méta-données et constituer des requêtes vers le RDF interne ou les bibliothèques numériques en ligne.
- 4 Et 4.' : Les intitulés des articles correspondants à la requête et trouvés dans le RDF ou les bases de connaissances scientifiques sont proposés.
- 5 Les articles sont cherchés sur le net si via Google scholar si l'Uniform Resource Identifier¹⁴ (URI) est absente de la base, ou directement proposés sur les

¹⁴ <http://www.w3.org/2004/11/uri-iri-pressrelease.html.en>

bibliothèques numériques. Si l'on utilise les bibliothèques numériques, l'accès aux documents est direct.

6. Essais de recherche naviguée

6.1. Navigation dans l'ontologie

La première étape de la recherche par navigation consiste à descendre dans l'arborescence jusqu'au nœud le plus représentatif du concept recherché.

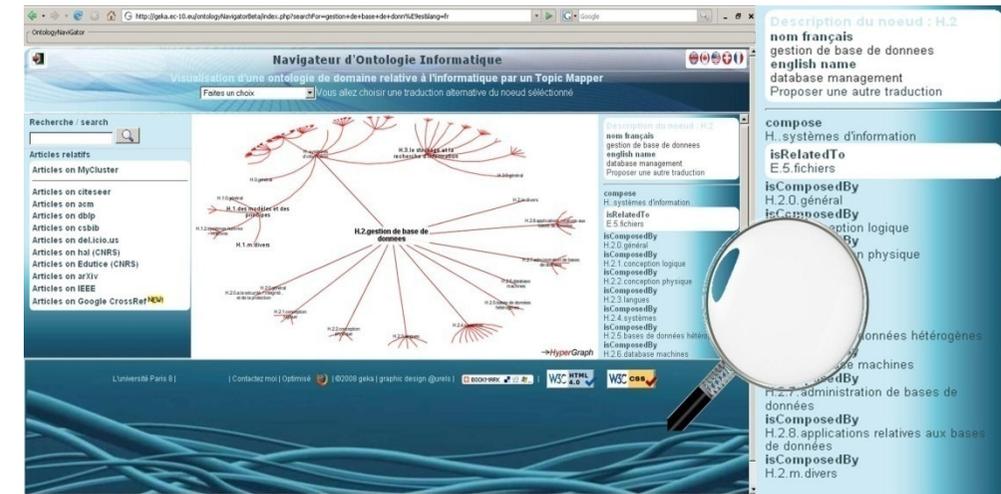


Figure 11: Recherche d'articles scientifiques par navigation de l'ontologie et zoom sur le contexte en focus.

Ici, la démarche de navigation pour atteindre le nœud «Gestion de base de données» a été de passer par la racine «Informatique» puis par « Systèmes d'information » et enfin de sélectionner nœud « Gestion de base de données ».

6.2. Exemple de recherche contextuelle d'articles



Figure 12: Articles scientifiques proposés par le système

Le bloc rouge du contexte (cf. figure 3) propose un accès direct aux articles des bibliothèques numériques en lignes comme CSBIB, DBLP, ou ACM en

générant des requêtes contextuelles vers ces sites (cf. figure 2). Mais l'outil propose également d'interroger la base interne d'intitulés d'articles. Dans l'exemple (cf. figure 4) une recherche sur « gestion de base de données » est générée et propose quelques dizaines de résultats. Nous choisissons « *Managing document taxonomies in relational databases* », la base nous donne l'auteur principal. L'outil vérifie la présence d'une URI relative à l'article dans la base, et en l'absence de celle-ci une requête est générée vers Google Scholar qui nous offre un accès direct à l'article.

Des tests ont été effectués vers les bases classiques, mais les résultats portant exactement sur le contexte de recherche sont encore trop rares. Le mécanisme de génération de requêtes est encore au stade d'heuristique, mais de bonnes perspectives sont ouvertes.

7. Limites et perspectives

Comme l'essai de l'outil l'a montré, la justesse actuelle des méta-requêtes générées est critiquable et de fait les résultats sont parfois approximatifs sur les bases externes. Il devrait cependant apparaître que, plus l'ontologie s'étoffera d'articles, plus la recherche et l'indexation seront précises. A cette fin, nous tenterons d'indexer un corpus préexistant de taille respectable. A cette fin, nous nous fixons comme ambition de créer un script d'extraction incrémental de contenu sur la bibliothèque en ligne DBLP. Ce travail en cours d'automatisation devrait affiner la pertinence d'indexation et de recherche par ontologie.

Une autre limite est l'accès physique aux articles qui est souvent soumis à un abonnement payant, quand ce n'est pas un paiement à l'unité. C'est pourquoi cette solution trouvera plus facilement une place dans les locaux d'un laboratoire universitaire ou une bibliothèque. Cependant, l'utilisation de proxy devrait permettre d'étendre l'accès aux bibliothèques numériques à tout un campus. L'outil va être mis à disposition des étudiants de second cycle du département informatique de l'Université Paris 8, au centre de calcul. Nous proposerons un formulaire en ligne pour enregistrer les retours en suivant les parcours utilisateurs.

Dans un avenir proche nous envisageons d'étendre l'application avec une ontologie basée sur les personnes au format *friend of a friend* (FOaF)¹⁵ pour mieux cerner les groupes de travail, équipe, et laboratoires ainsi que les liens de

¹⁵ <http://xmlns.com/foaf/spec/>

transversalité disciplinaire. Un autre objectif que nous ambitionnons est de rendre le système le plus autonome possible. Eventuellement le système de navigation de type *hyperbolic tree / eye tree* sera délaissé si un autre type de visualisation plus navigable ou ergonomique émerge.

8. Conclusion

Au cours de ce travail nous avons tenté de réaliser un outil de recherche pour les chercheurs dont les travaux sont liés à l'informatique. Cette interface consultable en ligne permet de lier un contexte de recherche ontologique à des bibliothèques scientifiques en ligne. Cette ontologie basée sur la CCS de l'ACM a été traduite en français de manière automatique pour proposer aux chercheurs francophones une alternative en langue maternelle. Notre solution propose aux chercheurs de trouver des articles relatifs à un contexte d'étude basé gravitant autour d'un nœud du domaine ontologique IT. Cette requête est générée par navigation graphique du domaine ou langage naturel. Une fois le contexte de recherche dégagé, un travail automatisé permet de trouver des articles en relation dans la base de données interne ou de proposer des méta-requêtes vers les bibliothèques numériques scientifiques en ligne. Le système de génération de méta-recherche étant uniquement basé sur les intitulés des nœuds de l'ontologie, ce projet en l'état a rapidement montré ses limites. Cependant les résultats sont encourageants, des perspectives d'amélioration ont donc été envisagées dans un futur proche ou sont déjà à l'étude.

Bibliographie

Bloehdorn S., Cimiano P., Duke A., Haase P., Heizmann J., Thurlow I., Völker J., *Ontology-based Question Answering for Digital Libraries. ECDL, Lecture Notes in Computer Science, Vol. 4675, pp. 14-25. 2007.*

Dragan G., Marek H., *Searching Web Resources Using Ontology Mappings. ntegrating Ontologies, CEUR Workshop Proceedings, Vol. 156, CEUR-WS.org, 2005.*

Gruber T. R., *Toward Principles for the Design of Ontologies Used for Knowledge Sharing, 1993.*

Le Coadic, Y.F, *Mathématique et statistique en science de l'information et en science de la communication. IBICT, 2006.*

Tricot C., Roche C., *Visualisation of Ontology: a focus and context approach, InSciT2006, 2006.*

Vander Wal T., *Understanding Folksonomy (Tagging that Works), dConstruct, 2006.*

Gestion des connaissances en médecine des assurances : modèle de représentation des connaissances et application technique

Juerg P. Bleuer¹, Kurt Bösch², Vincent Lampérière³,
Christian A. Ludwig²

¹Healthvidence GmbH
Jupiterstrasse 53/521
CP 6551
CH-3015 Berne
Suisse
bleuer@healthvidence.ch
http://www.healthvidence.ch

²Suva
Fluhmattstrasse 1
CH-6002 Lucerne
Suisse
kurt.boesch@suva.ch; christian.ludwig@suva.ch
http://www.suva.ch

³X8X Process Solutions AG
Bremgartenstrasse 7
CH-8003 Zurich
Suisse
vincent.lamperiere@x8x.com
http://www.x8x.com

Résumé :

La médecine est une science appliquée ; la coexistence d'avis différents provoque par conséquent chez les patients comme chez les médecins un sentiment d'insécurité. L'une des causes de la coexistence de ces différents avis médicaux réside dans la diversité

d'accès aux informations pertinentes et, de ce fait, dans l'hétérogénéité de la base de connaissances. Le service de médecine des assurances de la Caisse nationale suisse d'assurance en cas d'accidents (Suva) s'est attelé à ce problème dans le cadre d'un projet complet de gestion des connaissances. Le logiciel de gestion des connaissances InWiM (Integrierte Wissensbasen der Medizin, bases de connaissances intégrées pour la médecine) est l'une des composantes de ce projet. Cette application se fonde sur un modèle qui considère que les connaissances sont avant tout la mise en relation d'unités d'information. Pour rechercher les informations, le logiciel dispose d'une implémentation complète du thésaurus hiérarchisé de la United States Library of Medicine (index MeSH). À la connaissance des auteurs, InWiM est la seule application au niveau mondial qui met à disposition l'index MeSH, avec toutes les fonctionnalités, pour des recherches en interne.

Mots-clés : gestion des connaissances, représentation des connaissances, recherche d'informations, index MeSH

1. Introduction

La Suva est une entreprise indépendante de droit public assurant près de 100 000 entreprises, soit 1,8 million d'actifs et de chômeurs, contre les conséquences des accidents et des maladies professionnelles. Sur mandat du gouvernement, la Suva assume également la gestion de l'assurance militaire. Les prestations de la Suva comprennent la prévention, l'assurance et la réadaptation : afin d'offrir une réinsertion optimale aux victimes d'accident et aux patients qui souffrent de maladies professionnelles, la Suva met à la disposition des médecins généralistes comme des spécialistes en soins ambulatoires ou en milieu hospitalier une gestion complète des cas, et propose un service de conseils fourni par des médecins de son propre service de médecine des assurances.

La médecine est une science appliquée ; la coexistence d'avis différents conduit par conséquent non seulement à des débats scientifiques controversés, mais entraîne aussi des conséquences pratiques : en fonction des écoles et des expériences personnelles, des faits identiques ne sont pas évalués de la même manière par tous les médecins ; le pronostic et la thérapie s'en trouvent affectés, ce qui provoque une insécurité aussi bien pour le médecin que pour le patient. Cet état de fait a entraîné l'émergence de la médecine factuelle (*evidence-based medicine*). Le point central de l'approche factuelle réside dans l'exigence de décisions explicites basées sur les meilleures informations scientifiques disponibles, ainsi que sur la nécessité de devoir toujours motiver les avis médicaux [Sackett et al., 1996].

La coexistence de doctrines ou écoles diverses, partiellement contradictoires, a plusieurs causes. L'une d'entre elles réside dans la diversité d'accès aux informations pertinentes, et de ce fait, dans l'hétérogénéité de la base de connaissances sur laquelle les décisions se fondent.

Avec son ambitieux projet de gestion des connaissances, le service de médecine des assurances de la Suva s'est attelé précisément au problème de la non-uniformité des bases de connaissances sur lesquelles les décisions médicales se fondent : le projet est baptisé « InWiM », acronyme de « Integrierte Wissensbasen der Medizin » que l'on peut traduire en français par « bases de connaissances intégrées pour la médecine ».

InWiM définit et documente tous les processus de la gestion des connaissances. Le développement du logiciel de gestion des connaissances InWiM a démarré il y a quatre ans. La présente publication en fait la description.

2. Terminologie

Dans le langage courant, la distinction entre les termes « données », « informations » et « connaissances » est imprécise ; ils sont même fréquemment utilisés comme synonymes. On en relève de très nombreuses définitions dans la littérature. Les explications qui suivent se basent sur les définitions de Rehäuser et Krcmar [Rehäuser et al., 1996]. Les auteurs de cette publication utilisent ces termes avec les acceptions suivantes : « les données sont des représentations symboliques de faits ; les informations sont des données mises en contexte ; les connaissances naissent de la réflexion, c.-à-d. de la mise en relation logique et fonctionnelle des informations.

Les données brutes d'une étude ne sont déjà plus uniquement des symboles ; des unités de mesure (par ex. nombre de cas par 100 000 habitants) permettent de rendre les chiffres expressifs, ceux-ci sont donc déjà mis en contexte. Si le nombre de cas par 100 000 habitants – pour rester dans notre exemple – est placé dans un contexte supplémentaire (par ex. cas de maladie malgré la vaccination), il s'agit alors clairement d'informations, et non plus de données.

La transition de « informations » à « connaissances » s'effectue également de manière fluide : considérons par exemple les résultats d'une étude scientifique en particulier, on pourra y appliquer le terme d'« informations » ; si plusieurs études produisent des résultats concordants quant à une problématique, on s'approche alors toujours plus de « connaissances ». Ce sont finalement des directives qui,

dans le meilleur des cas, permettent de définir un recueil complet de connaissances relatives à une problématique donnée.

InWiM a pour objet de gérer et de mettre à disposition des informations et des connaissances. La gestion des données, leur recoupement et leur interprétation ne relèvent pas de l'application InWiM. L'entreposage et l'exploration de données ne sont pas non plus traités dans la présente publication.

Dans le langage courant, on accorde également au terme « connaissances » le sens de « vérité », que l'on distingue ainsi de « croyance » et de « supposition ». Dans le sens que nous leur donnons ici, aussi bien des informations que des connaissances ou bien encore des données peuvent être pertinentes, ou ne pas l'être. Pour souligner qu'une découverte désignée comme « connaissances » est de bonne qualité, on parle volontiers de « connaissances certaines ». Qu'une connaissance spécifique soit peu ou prou fondée, et quel est son degré de certitude, tient entre autres au fait qu'elle soit étayée par diverses sources, et au nombre de ces sources¹ [Canadian Task Force on the Periodic Health Examination, 1979].

3. Modèle de représentation des connaissances

Si l'on considère que les « connaissances » résultent de la réflexion au sens d'une mise en relation logique et fonctionnelle d'informations, cela implique également que les « connaissances » sont toujours subjectives. Les connaissances naissent non seulement de la mise en relation d'informations actuelles, mais aussi de la mise en relation avec des informations recueillies précédemment, et autant que possible dans d'autres contextes. Ces informations acquises par le passé, et les connaissances développées sur cette base, peuvent aussi être désignées comme « l'expérience ». Chaque individu possède des expériences personnelles différentes. Ces expériences ne sont pas qu'un élément des nouvelles mises en relation : elles participent également à déterminer la manière dont les nouvelles mises en relation se constituent.

Si nous partons de l'idée que chaque collaborateur d'une institution dispose de ses informations et que ses connaissances sont constituées par la mise en

¹ En médecine factuelle, la certitude d'une connaissance donnée est quantifiée par l'indication du niveau de certitude (level of evidence) et/ou du degré de recommandation: dans les directives factuelles (evidence-based guidelines), le poids des recommandations dépend du degré de certitude des connaissances sur lesquelles elles se fondent.

relation de ces mêmes informations, alors les connaissances d'entreprise (*corporate knowledge*) ne sont rien d'autre que l'ensemble de toutes les mises en relation effectuées par les collaborateurs. La gestion des connaissances d'entreprise (*corporate knowledge management*) consiste ainsi à mettre à disposition de tous l'ensemble des informations disponibles, avec leurs mises en relation.

4. Affinage du modèle

Considérons le modèle de mise en relation logique et fonctionnelle de deux unités d'information. La relation entre elles n'existe pas « en soi », mais dépend d'un point de vue donné [Rehäuser et al., 1996]. Des relations différentes peuvent alors exister en fonction de points de vue différents. Un modèle qui permet la représentation des connaissances personnelles doit donc être à même de relier entre elles des unités d'information selon plusieurs dimensions.

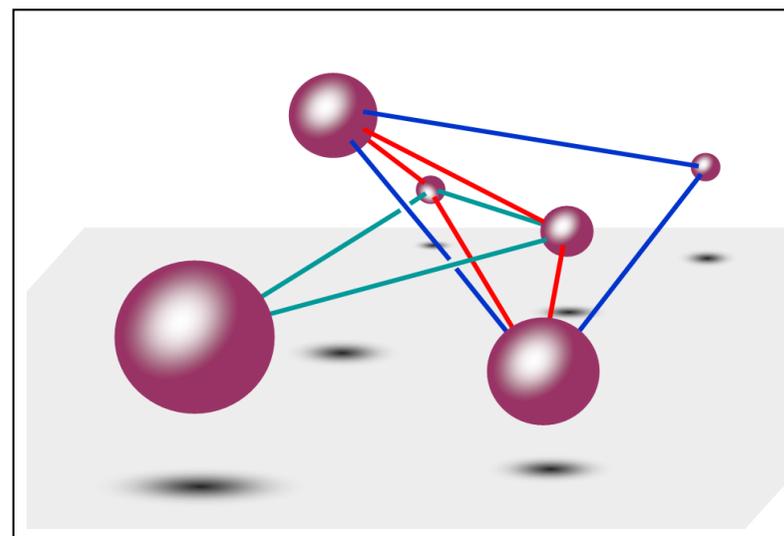


Fig. 1 : Mise en relation multidimensionnelle d'unités d'information représentées par des boules. Les couleurs des lignes représentent les diverses dimensions des mises en relation.

Chaque dimension de mise en relation correspond à un point de vue spécifique et se manifeste ainsi dans un contexte spécifique. Le modèle permet donc également de rendre compte du contexte de chaque mise en relation, et cela sous forme d'annotations.

Illustrons cela par un exemple : deux études rendent compte de l'efficacité et des effets secondaires des analgésiques. Alors que les deux études attestent de l'efficacité, elles sont contradictoires quant aux effets secondaires. Les conclusions correspondent à ces résultats : alors que l'une des études recommande la thérapie, l'autre exprime des réserves en raison des effets secondaires. Ces deux études fournissent donc des résultats identiques ou contradictoires, selon que le contexte est celui de l'« efficacité » ou celui des « effets secondaires »

Jusqu'ici, nous n'avons mentionné le contexte que dans le cadre des mises en relation logiques et fonctionnelles d'unités d'information. Toutefois, le contexte est aussi important pour chaque unité d'information en elle-même. En plus des informations usuelles habituellement désignées comme métadonnées² et qui indiquent la date de création, l'auteur, la version, etc., il est en outre possible d'ajouter des annotations à chacune des unités d'information. Le nombre des annotations n'est évidemment pas limité ; chaque annotation doit donc être considérée comme étant une méta-infomation dans un contexte donné.

5. Recherche d'informations

La représentation des connaissances avec des liens bidirectionnels entre les unités d'information et la possibilité d'ajouter des annotations aussi bien aux unités d'information qu'aux liens entre elles facilite la recherche rapide d'informations et de connaissances par « navigation » sur les liens correspondants. Il faut toutefois au préalable trouver le point d'accès adéquat.

Les homonymes et les synonymes sont fréquents en médecine, la recherche plein texte est rarement satisfaisante. L'index MeSH s'est donc imposé pour effectuer des recherches dans la littérature médicale. L'acronyme MeSH correspond à « Medical Subject Headings » et désigne le thésaurus de la United States National Library of Medicine (NLM). L'index MeSH trouve ses origines dans le répertoire des mots clés de l'Index Medicus publié depuis plus de 100 ans

² Le terme « métadonnée » est courant, mais dans le cadre de la terminologie utilisée ici, le terme de « méta-infomation » serait correct.

par la NLM, aujourd'hui disponible en ligne sous le nom de MEDLINE et librement accessible depuis PubMed.

L'index MeSH est un thésaurus hiérarchisé qui contient actuellement 24 767 descripteurs [United States National Library of Medicine, 2007]. Plus de 97 000 termes d'entrée (*entry terms*) permettent de trouver, en indiquant des désignations médicales usuelles, le terme MeSH adéquat. Ainsi par exemple, « *vitamin C* » est un *entry term* pour « *ascorbic acid* ».

6. Mise en application technique

Le logiciel de gestion des connaissances InWiM a été développé en tant que solution intranet sur la base du portail BEA (WLP) et du système de gestion de contenu d'entreprise (ECMS : *Enterprise Content Management System*) FileNet P8. En outre, une base de données Oracle contient l'index MeSH.

Le logiciel met à disposition toutes les fonctionnalités requises pour la collecte, la production, la gestion et la distribution d'informations et de connaissances. Conformément au modèle de représentation des données exposé dans les chapitres précédents, le système peut établir des liens croisés pour mettre des documents en relation bidirectionnelle. Les documents aussi bien que les liens croisés peuvent être munis d'annotations.

L'index MeSH a été totalement configuré pour la recherche d'informations, y compris en prenant en charge les sous-catégories (*subheadings*)³. À la connaissance des auteurs, InWiM est la seule application au monde en dehors de la NLM à offrir avec une implémentation MeSH toutes les fonctionnalités nécessaires à la recherche d'informations en interne. L'interface graphique utilisée pour la recherche en interne permet également l'accès à Pubmed ainsi que le téléchargement d'extraits de la littérature. Pour trouver la formulation de recherche adéquate, avec les termes MeSH corrects, un assistant a été créé par les programmeurs. Cet assistant permet également d'indexer les documents produits par l'utilisateur lui-même.

³ Les sous-catégories (*subheadings*) permettent de délimiter plus précisément le sens des termes recherchés. Ainsi, par exemple, la sous-catégorie « *adverse effects* » et le terme MeSH « *immunization* » permettent de trouver les effets secondaires de vaccins.

Bibliographie

Canadian Task Force on the Periodic Health Examination. *The periodic health examination. CMAJ* 1979;121:1193-1254.

Rehäuser J, Krcmar H, Schreyögg G. *Wissensmanagement in Unternehmen. Walter de Gruyter* 1996.

Sackett, DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. *Evidence based medicine: what it is and what it isn't. BMJ* 1996;312:71-72.

United States National Library of Medicine. *Facts sheet. PubMed: MEDLINE Retrieval on the World Wide Web.* <http://www.nlm.nih.gov/pubs/factsheets/pubmed.html>. 2007.

Remerciements

Les auteurs remercient tous les membres du groupe de projet pour leur collaboration et leur précieuse contribution à la réussite du projet :

Klaus Bathke, Erich Bär, Viktor Bydzovsky, Fiorenzo Caranzano, Massimo Ermanni, Bruno Ettlin, Pius Feierabend, Roland Frey, Franziska Gebel, Carlo Gianella, Raphael Good, Ulrike Hoffmann-Richter, Roland Jäger, Sönke Johannes, Bertrand Kiener, Hans Kunz, Jürg Ludwig, Wolfgang Meier, Bettina Rosenthal, Jan Saner, Rita Schaumann - von Stosch, Felix Schlauri, Holger Schmidt, Fred Speck, Klaus Stutz, Felix Tschui, Walter Vogt.

Les auteurs adressent un remerciement particulier au service linguistique de la Suva, pour la traduction comme pour la relecture du manuscrit.

Mise en évidence de la sémantique dans la conception d'un système d'aide au diagnostic des plaintes suite à des situations de pollution de l'air dans les logements

Zoulikha Heddadji*, **, Nicole Vincent*

Séverine Kirchner, Georges Stamon***

*Université René Descartes

45, rue des Saints Pères 75270 Paris CEDEX06

**CSTB

84, avenue Jean Jaurès Champs-sur-Marne

77421 Marne-la-Vallée CEDEX2

{*zoulikha.heddadji, severine.kirchner*}@cstb.fr

{*nicole.vincent, Georges.Stamon*}@math-info.univ-pris5.fr

Résumé :

Notre travail a pour objet de mettre en place un système à base de connaissance capable d'interpréter les circonstances à l'origine des phénomènes de pollution dans les logements exprimés par des particuliers dans des plaintes écrites. Le système s'appuie sur une base d'anciennes plaintes écrites, résolues suite à des audits de terrain. Par conséquent notre étude se place à la jonction de deux problématiques: celle de l'analyse (la compréhension) et la modélisation du raisonnement des experts lors des audits suite à des situations de pollution dans les logements, et celle de la gestion des problématiques classiques de la langue naturelle. Le problème majeur dû au besoin de la compréhension de la langue naturelle et auquel nous nous intéressons est la sémantique. Le module fonctionnel de notre application est établi à partir d'un système de recherche d'information. Dans ce papier nous montrons notre contribution pour l'amélioration du modèle de recherche fondé sur le principe de la proximité floue des termes des requêtes au sein des documents en l'adaptant au besoin de la gestion de la sémantique et à la recherche d'informations semi-structurées (XML). Nous montrons les résultats de l'analyse du lexique du corpus des plaintes qui nous permettent d'expliquer l'utilité du dictionnaire des synonymes DICTIONNAIRE dans la gestion de la sémantique et pour chaque système de recherche adaptée. À l'aide des courbes rappel/précision et pour chaque système de recherche nous comparons sa performance entre le cadre où il s'agit d'appariements directs et lorsqu'il s'agit d'appariements sémantiques. L'intérêt de la sémantique est aussi prouvé dans une évaluation objective de la qualité des partitions automatiques des plaintes.

Mots-clés : Système de recherche d'informations semi-structurées, logique floue, sémantique, dictionnaire des synonymes, plainte, air intérieur

1. Introduction

Les premières actions et enquêtes menées dans le domaine de l'air au sein des lieux de vie, l'habitat notamment, sont récentes et pour la plupart encore en cours. L'expertise est par conséquent nouvelle et non pluridisciplinaire. En effet, il existe des experts des milieux intérieurs spécialistes dans le domaine de la microbiologie, d'autres sont ingénieurs en ventilation ou ingénieurs chimistes. Entre-temps les plaintes liées aux malaises ressentis dans les environnements intérieurs où nous passons le plus clair de notre temps (plus de 80%) et qui ne cessent de parvenir aux organismes en charge de les traiter sont le plus souvent laissées sans réponses. L'objectif de notre travail est de développer un système d'aide au diagnostic, dédié à la compréhension des circonstances à l'origine des malaises en milieu intérieur. Grâce à ce système, nous souhaitons d'une part fournir une expertise centralisée et multidisciplinaire du domaine de la pollution intérieure et d'une autre part permettre un traitement automatique des plaintes des particuliers écrites en langue naturelle. Par traitement nous entendons une précision concernant la nature du problème de pollution domestique exprimé dans le texte de la plainte ainsi qu'une préconisation d'un ensemble d'options de gestion et d'actions correctives assurant l'amélioration du problème sanitaire cité.

2. Formalisme des plaintes

Les plaintes en notre possession sont issues de divers organismes en charge de les recueillir: 13 documents issus du département de microbiologie du laboratoire d'hygiène de la ville de Paris (LHVP), 14 dossiers provenant du département de pollution chimique du LHVP, 583 documents issus du Service des Analyses en Milieux Intérieurs de Liège en Belgique (Sami), 45 plaintes reçues au Centre Scientifique et Technique du Bâtiment (CSTB). Ces documents sont non-structurés et sont écrits en langue naturelle. Leur taille varie ; les dossiers issus du Sami de Liège et dont la partie problème a été retranscrite par l'expert en charge du diagnostic air intérieur sont très brefs, alors que les dossiers LHVP sont beaucoup plus détaillés et donc beaucoup plus longs. Concernant la structure des plaintes, malgré le fait qu'aucune régularité conversationnelle ne soit explicite sur l'ensemble des plaintes au départ, une structure récurrente apparaît à

travers les différents dossiers. En effet, le plaignant parle le plus souvent de ses symptômes, de son habitat et de ses activités principales au sein de son logement. Il décrit également l'environnement extérieur de son logement qu'il incrimine d'ailleurs très souvent. Nous avons retenu ces quatre rubriques conversationnelles en les formalisant à l'aide de balises XML sémantiquement pertinentes délimitant le contenu de chaque partie de la plainte en relation avec la rubrique concernée. Nous appelons cette partie la «structure de contrôle» (figure 1). L'interface usager, dédié à l'enregistrement des plaintes, se présente sous forme de quatre champs de saisie où l'utilisateur répond à quatre questions principales : parlez moi le plus précisément possible de vos problèmes de santé, de votre logement, de vos habitudes de vie, de votre environnement extérieur. Les autres informations périphériques comme le numéro de la plainte ou bien le nom de l'organisme en charge du dossier sont renseignées au niveau des champs correspondant aux balises de la partie «paramètres globaux».



Fig. 1 : Exemple d'une plainte air intérieur filtrée, lemmatisée et structurée au format XML

3. La méthodologie

Une certaine régularité des phénomènes de pollution domestique a été constatée à travers notre corpus de plaintes (648 documents). Il est certain que ces plaintes, si elles reflètent chacune un cas particulier, abordent des problèmes communs que les experts aimeraient identifier de manière objective. Selon nos analyses de corpus, nous avons constaté au niveau des parties solutions des dossiers des demandes d'intervention que les experts en charge d'enquêter sur site à la suite d'une plainte d'un particulier utilisaient des modèles de courrier récapitulant l'ensemble des actions à entreprendre en fonction du type de pollution. Ce constat nous a permis de définir notre approche de résolution des plaintes. La démarche que nous suivons pour atteindre cet objectif consiste à établir dans un premier temps un modèle de recherche. À l'aide de ce modèle et à l'aide d'une base archive d'anciennes plaintes résolues stockée en mémoire, la plainte courante est appariée à la plainte résolue la plus similaire. Ainsi le modèle de solution attribuée à la plainte la plus pertinente est assigné à la plainte nouvelle à traiter (figure 2). L'avantage supplémentaire de cette approche est que, une fois l'assignation de solution réalisée pour la plainte courante à traiter validée, cette dernière enrichit la base d'exemples et pourra ainsi contribuer aux futures assignations de solutions.

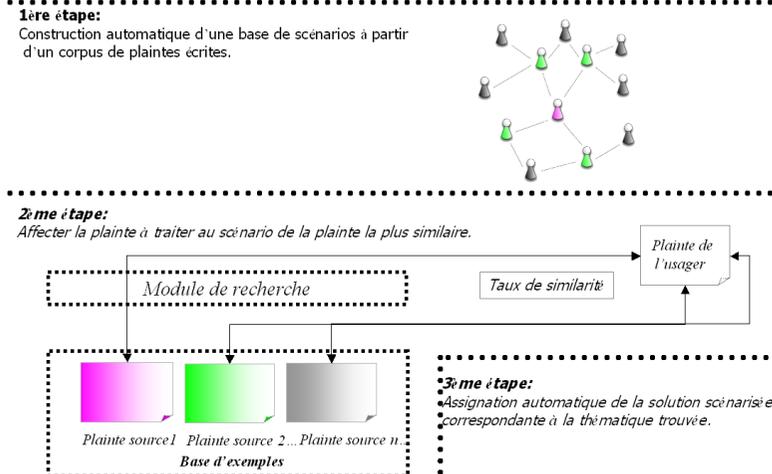


Fig. 2 : Schéma de la méthode générale de l'étude développée

3.1. Méthodes de recherche employées

Le module de recherche est le module fonctionnel de notre applicatif. Le système de recherche que nous employons doit être adapté à la nature et à l'origine du corpus utilisé et des documents que l'on doit traiter. Par l'origine nous entendons l'utilisation de la langue naturelle pour l'expression des plaintes, ce qui signifie qu'il devient nécessaire de gérer la sémantique des termes afin que deux plaintes exprimées différemment et ayant le même sens puissent être appariées par le modèle de recherche. Et par la nature nous distinguons la structure et la taille des documents. En effet, certains modèles de recherche sont plus efficaces pour l'appariement des textes longs alors que d'autres sont mieux adaptés et plus optimaux pour l'appariement des textes courts. Sachant que les modèles vectoriels connus sont efficaces pour la modélisation et l'appariement des textes longs, nous avons alors développé le modèle vectoriel étendu en vue bidimensionnelle. Aucune information sur la position des termes les uns par rapport aux autres n'est mentionnée dans les représentations vectorielles. Le modèle de recherche basé sur la proximité floue des termes des requêtes au sein des documents est fondé sur ce principe. Nous l'avons alors développé et adapté à la structure de notre corpus et de nos requêtes. Ces deux méthodes ont été employées suivant une approche lexicale où chaque descripteur de la requête est recherché en fonction du terme exact et une approche sémantique grâce à laquelle un terme sémantiquement proche d'un descripteur de la requête peut être rappelé par le modèle de similarité.

3.2. Approche lexicale

Le système de recherche de Salton étendu

Zargayouna [Zarga, 2004] a adapté le modèle vectoriel de Salton [Salton, 1998] aux documents semi-structurés XML. Elle définit ainsi le nouveau poids des termes TF-ITDF. Ce score attribué au lexique du corpus dépend de la fréquence d'apparition du terme dans le modèle de balise considérée, il dépend également de son apparition dans le reste des modèles de balises au sein du même document, et il dépend aussi de sa distribution au sein du même modèle de balise dans le reste des documents du corpus.

$$\begin{aligned}
 TF\text{-}ITDF(t,b,d) &= TF(t,b,d) \text{ ITF}(t,d) \text{ IDF}(t,b) \\
 IDF(t,b) &= \text{Log}(|B|_d / \text{TagF}(t,d)) \\
 ITF(t,d) &= \text{Log}(|D|_b / DF(t,b))
 \end{aligned}$$

$|D|_b$: est le nombre total des documents où la balise b apparaît

$|B|_d$: est le nombre total de balises dans le document d
 DF(t,b): (Document Frequency) est le nombre de documents qui contiennent la balise b et dans lesquels le terme t apparaît au moins une fois
 TagF(t,d): (Tag Frequency) est le nombre de balises dans le document d dans lesquelles le terme t apparaît au moins une fois
 TF(t,b,d): (Term frequency) est le nombre de fois que le terme t apparaît dans la balise b dans le document d.

Le système de recherche basé sur la proximité floue des termes

Le principe de cette approche est d'évaluer la densité des termes de la requête au niveau des textes pour évaluer leur pertinence. Mercier et Beigbeder [Mercier, 2005] font l'hypothèse que plus les termes de la requête sont proches dans un document alors plus ce document est pertinent. Fondés sur ce principe, ils calculent la pertinence relative μ des termes de la requête aux différentes positions x au sein d'un document source (document de la base):

$$\mu_i^d(x) = \text{Max}_{i \in d^{-1}} (\text{Max}(\frac{k - |x - i|}{k}, 0))$$

d^{-1} est l'ensemble des positions pouvant être prises au sein du document d. Le paramètre k^1 est une constante qui caractérise le degré d'influence d'une occurrence. À l'aide des opérateurs logiques «ET» et «OU» la pertinence d'une requête booléenne à une position x est calculée de la manière suivante:

$$\mu_{A \text{ ET } B}^d(x) = \text{Min}(\mu_A^d(x), \mu_B^d(x)) \quad , \quad \mu_{A \text{ OU } B}^d(x) = \text{Max}(\mu_A^d(x), \mu_B^d(x))$$

Le score de pertinence d'un document source d par rapport à une requête q est défini comme suit:

$$\text{Score}(q, d) = \frac{\sum_{x \in Z} \mu_q^d(x)}{|d^{-1}|}$$

¹ Une valeur aux alentours de 5 évalue la proximité dans le cadre d'une expression, une valeur de k égale à 100 traduit la proximité dans un contexte paragraphe et ainsi de suite.

3.3. Approche sémantique

Le système de Salton étendu et sémantique

Si l'on souhaite que les occurrences des termes n'aient plus besoin d'exister directement au sein d'un document pour avoir un poids non nul au niveau du vecteur des caractéristiques, nous devons utiliser le modèle vectoriel sémantique défini par Zargayouna [Zarga, 2004]. Le poids sémantique SemW du terme t au sein d'une balise b d'un document d au niveau du vecteur sémantique correspond à la somme de son poids TF_ITDF et les poids des termes qui lui sont proches sémantiquement.

$$\text{SemW}(t, b, d) = \text{TF} - \text{ITDF}(t, b, d) + \frac{(\sum_{i=1}^n \text{Sim}_{zs}(t, t_i) \text{TF} - \text{ITDF}(t_i, b, d))}{n}$$

Notre modèle de recherche

La mesure de Mercier et de Beigbeder rappelée précédemment est très intéressante, néanmoins elle ne tient pas compte de la sémantique. Nous présentons ici un modèle inspiré du modèle de proximité flou permettant de prendre en considération des liens de similarité latents entre documents. Notre contribution s'inscrit dans cette perspective, elle se propose alors d'ajouter la dimension sémantique au modèle de proximité flou [Heddadji *et al.*, 2007] et d'adapter également au formalisme XML. La pertinence locale et sémantique d'un terme à une position donnée dans le contenu d'une balise est modélisée par:

$$\mu_i^d(x) = \text{Max}_{i \in d^{-1}(\text{Synos}(t))} (\text{Max}(\frac{(k - |x - i|) \text{Sim}(t_i, t)}{k}, 0))$$

Un seuil de similarité est nécessaire afin de délimiter l'ensemble des termes sémantiquement pertinents par rapport à t (Synos(t)). Nous fixons cette limite à l'ensemble des termes dont le degré de similarité avec t est au-delà du score de similarité de ce dernier avec le terme auquel est rattachée la balise où son occurrence apparaît. Cette dernière spécification représente notre contribution dans l'adaptation du modèle de proximité flou au formalisme XML. Les modèles de similarité que nous avons cités nous permettent d'évaluer des appariements locaux (entre modèles de balises). La généralisation de ces mesures au niveau «document» est établie en effectuant une agrégation des similarités locales. La notion d'agrégation peut être définie selon les besoins de l'application. Par exemple, certaines applications accordent un plus grand intérêt à certaines rubriques plutôt qu'à d'autres, dans ce cas des poids différents sont accordés aux

différents taux d'accord locaux. Dans notre contexte précis, nous calculons la moyenne des similarités locales, avec des taux égaux, en ne considérant que les rubriques communes renseignées.

4. Gestion de la sémantique

Les plaintes ne sont pas formulées uniquement par des experts dans un langage technique et précis mais elles sont écrites par notamment des particuliers en langage ouvert et naturel. De nombreuses marques de produits sont utilisées, le vocabulaire est très vivant, d'où l'interrogation de la possibilité de construire une ontologie. Nous avons par conséquent étudié le vocabulaire des dossiers relatifs aux plaintes dont nous disposons.

4.1. Préparation des textes

Pour étiqueter les textes des plaintes nous avons utilisé l'outil de traitement morphologie flexionnelle Tree-tagger adapté au français. Après l'étiquetage morpho-syntaxique de chacune des plaintes nous récupérons en sortie uniquement les lemmes correspondant aux différentes flexions originelles de chaque document. Une stop-liste est utilisée afin d'éliminer automatiquement les mots vides de sens à partir de la forme lemmatisée des textes. La totalité des plaintes a été retranscrite en employant les lemmes de Tree-tagger. Une catégorie de mots n'a pu être reconnue directement par Tree-Tagger. Les sigles, abréviations et autres acronymes en relation avec le domaine de la pollution domestique décryptés naturellement et aussi bien par les experts que par les occupants ne sont pas connus par Tree-Tagger. Pour compenser cela, nous avons dédié un fichier récapitulant l'ensemble des mots du corpus inconnus par l'étiqueteur. L'objectif, à cette étape, est de substituer ces termes par des quasi-synonymes ou des termes plus génériques compréhensibles par l'étiqueteur dans le cas où il s'agit de marques de produits ménagers, ou une sorte très spécifique de champignons des milieux intérieurs par exemple. Ce fichier pourrait régulièrement être mis à jour par les experts spécialistes.

4.2. Lexique des plaintes

La compréhension de la langue naturelle est spécifiée formellement par la notion d'ontologie. L'ontologie organise sous forme logique les concepts d'un domaine donné dans une hiérarchie, ensuite chaque terme (instance d'un concept) est rattaché à un ou à plusieurs concepts. Dans des domaines précis qui

se présentent par des concepts clairs et des termes techniques, la communauté professionnelle s'accorde autour d'une «ontologie métier» mêlant la terminologie et la conceptualisation de la connaissance du domaine. Notre corpus contient des plaintes en rapport avec des problèmes de pollution intérieure uniquement. Nous montrons ici l'analyse de l'évolution du vocabulaire utilisé par les auteurs des parties «problème» des dossiers de pollution intérieure en notre possession afin de savoir si il est possible de contrôler le vocabulaire (figure 3). Le constat schématisé par l'allure de la courbe de la figure 3 est une preuve de l'insuffisance d'une éventuelle ontologie gérant de façon intégrée la sémantique et le vocabulaire terminologique issu du corpus des plaintes. En effet, la courbe ne cesse d'évoluer. On peut observer également que la richesse du vocabulaire utilisé dépend de l'origine du laboratoire recevant la plainte. Les points encerclés sur la courbe désignent des passages d'un corpus appartenant à un laboratoire d'analyse des milieux intérieurs à un autre. Afin de permettre l'usage de la langue naturelle libre dans la description des plaintes il devient nécessaire de construire un réseau conceptuel généralisé de façon à comprendre la langue naturelle. À cette date, il n'existe pas d'ontologie universelle correspondant au bon sens humain et dans laquelle on puisse retrouver de manière exhaustive la totalité des termes utilisés dans le langage naturel et qui puisse servir de base à un système de recherche général implémentant la sémantique. La base de données lexicales WordNet aurait pu servir de base pour définir le sens et les relations entre termes, mais le point essentiel sur lequel nous insistons est que notre corpus est rédigé intégralement en français et que le système d'aide à la décision sera dédié au traitement des plaintes écrites en langue française. Par conséquent, c'est une ressource en français et au moins aussi bien détaillée que WordNet qu'il nous faudrait dans le cadre de ce travail. Ces différentes constatations sont en quelque sorte une réminiscence de l'utilité des dictionnaires électroniques des synonymes pour le contrôle de la sémantique tout en assurant une couverture la plus exhaustive possible du lexique des plaintes que nous devons traiter.

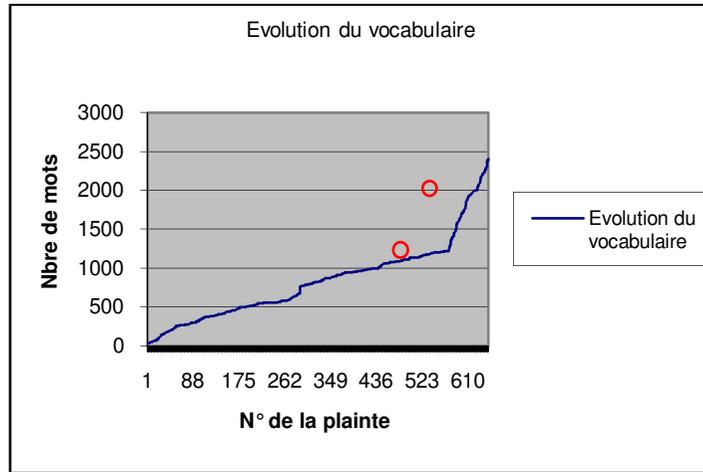


Fig. 3: Courbe de l'évolution du nombre des termes différents en fonction du nombre des plaintes analysées

4.3. Dictionnaire électronique des synonymes

Nous avons utilisé le dictionnaire électronique des synonymes du laboratoire CRISCO de l'université de Caen baptisé DICTIONNAIRE. Ce dictionnaire informatisé est accessible en ligne et regroupe les synonymes de 48 881 termes (vedettes) issus à l'origine de sept dictionnaires papiers classiques (Guizot, Lafaye, Bailly, Bénac, Du Chazaud, Grand Larousse et Grand Robert). En plus de la connaissance sémantique que nous offre DICTIONNAIRE, nous l'utilisons en tant que lexique de base pour les modèles vectoriels. Nous utilisons les vedettes de DICTIONNAIRE en tant que primitives vectorielles caractérisant les textes et permettant de les apparier à l'aide du module de recherche implémentant les formalismes vectoriels.

Relation de synonymie et calcul des taux de similitude

La proximité sémantique entre deux vedettes A et B du dictionnaire électronique repose sur le calcul de l'indice de communauté aussi appelé l'«indice de Jaccard » [Manguin, 2004]. L'indice de similitude S est calculé à partir des relations synonymiques entretenues entre A et B résumées dans le tableau de contingence Tab. 1.

		Synonymes de A	
		Oui	Non
Synony mes de B	Oui	a	b
	Non	c	d

Tab. 1 : Tableau de contingence des données synonymiques

La formule de calcul de l'indice de Jaccard est donnée par Legendre et Legendre [Legendre, 1998]:

$$S = \frac{a}{a+b+c}$$

Dans le tableau Tab.2 nous montrons les termes extraits de DICTIONNAIRE et avec lesquels le terme «pollution» partage un degré de similarité supérieur à 0,1. Dans Tab.2, excepté le terme «pollution» lui-même, nous ne constatons aucune autre forme fléchée de ce dernier, par exemple «polluer» et «polluant». Ce qui est normal, puisqu'ils sont de deux catégories grammaticales différentes. Néanmoins et si nous partons du principe que les formes fléchées d'un terme partagent le même sens, le dictionnaire des synonymes apparaît donc insuffisant et un système de codage des termes est nécessaire pour la gestion de la sémantique entre termes issus du même code.

Synonyme	Taux de similarité
contamination	0,111
corruption	0,134
dénaturation	0,15
déprédation	0,111
empoisonnement	0,107
infection	0,121
salissement	0,117
salissure	0,125

souillure	0,153
spermatorrhée	0,117
vandalisme	0,1
nuisance	0,263
viciation	0,352
pollution	1

Tab. 2 : Extrait de l'ensemble des termes sémantiquement proches de «pollution»

4.4. Heuristique d'Enguehard

L'aspect grammatical pris en compte dans Tree-Tagger n'est pas suffisant pour rapprocher certains mots relevant de la même famille. Nous avons choisi d'utiliser une heuristique, certes imparfaite, mais qui permet de rapprocher des termes ayant la même racine. Nous avons utilisé l'heuristique d'Enguehard [Enguehard, 1992] qui se résume comme suit: « le code d'un terme est la sous-chaîne de caractères rassemblant les premières lettres qui le composent jusqu'à l'obtention de deux voyelles non consécutives ». Nous avons choisi cette hypothèse parce qu'elle nous a semblé facile à mettre en œuvre et surtout parce qu'elle a fait ses preuves dans d'autres applications [Serradura *et al.*, 2002]. Afin d'évaluer le taux de similarité entre deux termes du même code, nous avons ré-utilisé DICTIONNAIRE. Entre chaque paire de termes A et B de dictionnaire et qui sont issus de la même famille selon l'hypothèse des codes on a effectué un échange de synonymes avec une influence de 1/2. Le taux de similarité entre les éléments de chacune de ces paires est calculé à l'aide de l'indice de Jaccard. Après différentes simplifications, le taux de similarité pour les termes de la même famille est de 1/2.

A : A, A', A'', 1/2 B, 1/2 B', 1/2 B''

B : B, B', B'', 1/2 A, 1/2 A', 1/2 A''

$$Sim(A, B) = \frac{\frac{1}{2}A + \frac{1}{2}A' + \frac{1}{2}A'' + \frac{1}{2}B + \frac{1}{2}B' + \frac{1}{2}B''}{A + A' + A'' + B + B' + B''} = \frac{1}{2}$$

5. Evaluation de la prise en compte de la sémantique

5.1. Intérêt de l'analyse du corpus et de la construction automatique des scénarii de pollution

Comme nous l'avons cité précédemment, suite à une analyse manuelle du corpus nous avons fait le constat qu'il existait une certaine régularité thématique concernant les phénomènes de pollution exprimés dans les plaintes ce qui ne nous a pas été rapporté initialement par les experts. Nous avons souhaité appuyer ce constat à l'aide d'une analyse automatique effectuée à partir d'un échantillon représentatif regroupant 100 documents d'entraînement. Pour les modèles vectoriels (direct et sémantique) nous avons utilisé la méthode de classification supervisée des k-moyennes. Pour ce qui est des modèles flous où nous ne disposons plus de représentation vectorielle, nous avons utilisé la méthode du meilleur représentant (de manière itérative le centroïde devient l'élément qui est le plus au centre du cluster considéré). Ce travail de segmentation a été réalisé parallèlement par 3 experts du CSTB. À partir du même échantillon nous leur avons demandé de les classer dans des groupes selon la nature des plaintes. Les experts se sont entendus sur l'existence de 3 thématiques traitant chacune d'un phénomène de pollution isolé. Le but de cette étude est d'une part, évaluer les performances des modèles de recherche documentaire par rapport à la capacité de raisonnement des experts, et d'une autre part de voir si il est possible de synthétiser tous les cas possibles à l'aide d'un nombre fini de scénarii (clusters). Le but de ce travail est de tenter une synthèse des différentes solutions possibles et de pouvoir ensuite assigner la bonne solution à la nouvelle plainte à traiter.

5.2. Evaluation de la qualité des partitions obtenues

Une bonne partition présente des classes bien séparées. L'objectif de la classification est de maximiser les écarts entre classes et de minimiser les écarts entre les éléments au sein d'une même partition. Pour évaluer la qualité de nos différentes classifications, nous avons calculé le rapport entre la distance inter-classes et la distance intra-classe. Les rapports d'inertie concernant les partitions établies à partir des modèles sémantiques vectoriels et flous à l'aide des méthodes de classification supervisée enregistrent leurs meilleurs scores au niveau du partitionnement à 3 classes. Des scores moins importants sont constatés pour des partitions ayant un nombre de classes différent. À partir du graphe de la figure 4, nous constatons que les modèles sémantiques enregistrent des scores de qualité de partition moins importants que les modèles directs concernant les partitions

dont le nombre de classes est différent de 3. Ceci est dû au fait que les modèles de recherche implémentant la sémantique mettent en évidence les liens entre les documents des trois clusters isolés et les éléments des clusters supplémentaires. Et par conséquent le rapport entre la distance intra-classes et la distance inter-classe diminue sous le contrôle des modèles sémantiques par rapport au score des partitions établies à l'aide des modèles d'appariement surfacique.

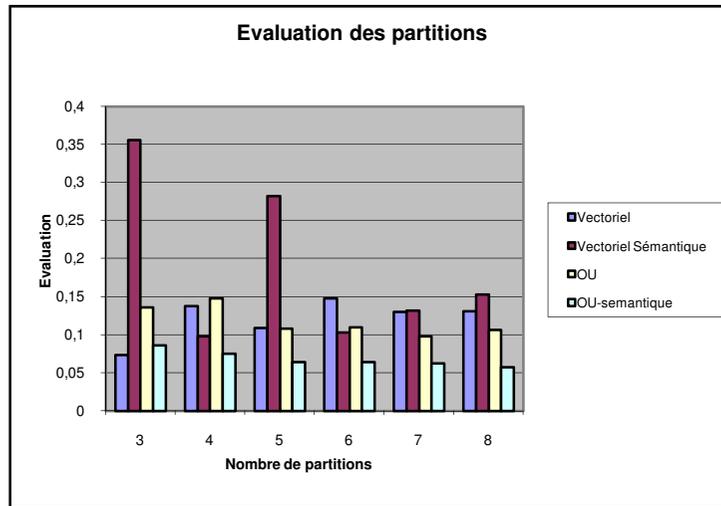


Fig. 4 : Evaluation objective des partitions réalisées à l'aide des systèmes de recherche employés

5.3. Evaluation des systèmes de recherche implémentés

Dans une première série d'expériences, nous avons voulu savoir quel système de recherche parmi ceux que nous avons développés propose le meilleur ordre de l'estimation de la pertinence d'un corpus d'entraînement de 100 plaintes en fonction d'une plainte de référence. Nous utilisons à cet effet les courbes «rappel/précision» relatives au classement des trois experts. Afin d'évaluer les différents systèmes nous utilisons la méthode de la moyenne des taux rappel/précision en utilisant 15 requêtes aléatoires différentes issues de l'échantillon des 100 plaintes. Dans le cadre de notre application, la précision est privilégiée par rapport au taux de rappel étant donné que l'assignation de solution à la plainte à traiter est effectuée en fonction de l'élément positionné en tête de liste dans le classement du modèle de recherche employé. Etant donné que nous nous intéressons dans cette partie à l'évaluation de la prise en compte de la

sémantique, alors, à l'aide des courbes rappel/précision des figures 5 et 6 nous comparons les systèmes de recherche implémentés dans le cadre d'un appariement lexical et sémantique. Nous constatons que le modèle vectoriel sémantique donne un meilleur taux de précision jusqu'au taux de rappel de 35% (figure 5). Le modèle de proximité flou augmenté sémantiquement et basé sur des requêtes disjonctives (OU) donne de meilleures valeurs de précision jusqu'à un niveau de rappel de presque 30% (figure 6). Cette expérience montre que l'exploitation de la sémantique améliore la performance des systèmes de recherche que nous avons implémentés dans le cadre de la recherche initié par des requêtes écrites en langue naturelle.

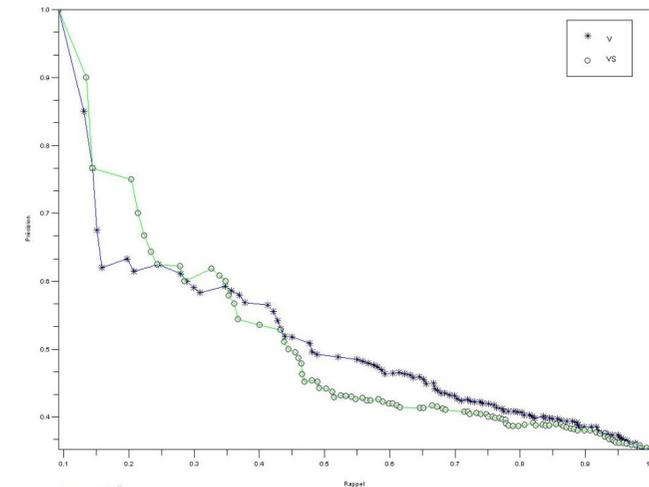


Fig. 5 : Courbes rappel/précision pour l'évaluation de la prise en compte de la sémantique par le système de recherche de Salton étendu

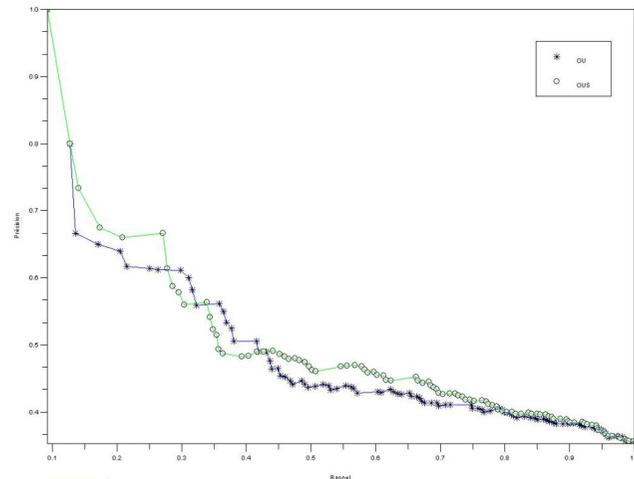


Fig. 6 : Courbes rappel/précision pour l'évaluation de la prise en compte de la sémantique par le système de recherche flou

6. Conclusion

Nous avons présenté dans cet article le contexte général de notre travail, les ressources et les méthodes que nous utilisons afin d'expliquer de manière automatique la nature d'une plainte à traiter et de guider l'utilisateur du système pour améliorer les conditions de vie dans son logement. Pour des raisons diverses, nous n'avons pu exposer les différentes expérimentations que nous avons menées. Concernant l'évaluation subjective des partitions par exemple, nous utilisons l'indice de Rand corrigé pour connaître le taux d'accord entre les partitions des experts et celles réalisées à l'aide des différents modèles étudiés. Au final, un seul système de recherche d'information est utilisé pour distinguer la plainte résolue la plus similaire à la plainte courante afin de réaliser l'assignation de solution. Pour savoir quel système utiliser et dans quel contexte nous souhaitons réaliser une analyse en composantes principales en fonction des résultats des tests d'assignation suivant différents critères de sélection. Pour le moment nous distinguons 2 critères candidats à la sélection: la taille du texte (de la requête et du document source) et le bruit (au sein de la requête et au sein du document source). Par le bruit nous entendons la quantité d'information issue du domaine mais non très pertinente lors du processus d'appariement.

Bibliographie

- Bellia, Z. *Modélisation d'un système informatique pour la gestion des demandes d'intervention dans le domaine des ambiances intérieures: Une approche basée sur le Raisonnement à Partir de Cas*, mémoire de stage DEA-IMTC, ParisVI, 2004.
- Enguehard, C. *ANA, Apprentissage Naturel Automatique d'un réseau sémantique*, Thèse de doctorat, 1992.
- Heddadji, Z. et N. Vincent et G. Stamon et S. Kirchner. *Extension sémantique du modèle de similarité basé sur la proximité floue des termes*, RNTI, numéro spécial Extraction et Gestion des Connaissances (EGC'2007), 2007.
- Legendre, P. et L. Legendre. *Numerical Ecology*, Elsevier, 1998.
- Manguin, J-L. *Regroupements de synonymes par indices de similitude : exemple avec l'adjectif ancien*, dans le Cahier de Lexicologie, n° 86, 2005.
- Mercier, A. et M. Beigbeder. *Application de la logique floue à un modèle de recherche d'information basé sur la proximité*, dans les Actes LFA 2004, 231–237, 2005.
- Salton, G. et C. Buckley. *Term-Weighting Approaches in Automatic Text Retrieval*, Journal of Information Processing & Management, 513–523, 1988.
- Serradura, L et M. Slimane et N. Vincent. *Classification semi-automatique de documents Web à l'aide des Chaînes de Markov Cachées*, dans les actes du colloque Inforsid, 215-228, 2002.
- Zargayouna, H et S. Salotti. *Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML*, dans Actes de la conférence IC'2004, 2004.

Un témoignage issu d'une expérience professionnelle à la Banque de France et deux suggestions

Alain Dequier

Banque de France/SGCB

115, rue Réaumur

75002 Paris

Alain.dequier@Banque-france.fr

<http://www.banque-france.fr>

Avertissement :

Cette contribution est un témoignage personnel, quelques réflexions qui n'engagent aucunement l'Institution qui m'emploie.

Résumé :

Au cours d'une carrière professionnelle les occasions ne manquent pas de constater l'importance d'une bonne terminologie pour la compréhension entre collègues ; celle-ci peut être obtenue en explicitant des relations sémantiques, ce texte en donne quelques exemples. Le recours à certains outils mis à disposition des internautes par des laboratoires du CNRS et utilisés dans un cadre professionnel m'a suggéré une extension qui pourrait être très profitable pour le bon usage du français : la mise en évidence de la proximité des mots et de leurs nuances selon les échelles sur lesquelles ils peuvent se placer. L'étude des gradations sémantiques et leur mise en œuvre dans un outil informatique est donc proposée. La seconde suggestion concerne une présentation des nécessaires traductions auxquelles l'internationalisation du travail nous contraint : une forme adaptée aux textes internationaux permet ce que j'appellerai « voir les idées en relief » et de préserver l'usage du français

Mots-clés : Proxisémie ; vocabulaire raisonné ; lexique structuré ; traduction de proximité.

1. Témoignages dans les domaines de

1.1. La conduite de projet

Au sein de la Banque de France au début des années 90, la méthode de conduite de projet retenue était la fameuse méthode française Merise largement enseignée avant qu'apparaissent les modélisations objet. Malgré cette unanimité sur le choix de la méthode, les projets avaient beaucoup du mal à démarrer. Une enquête a montré d'assez importantes divergences de définition dans les composantes d'une conduite de projet : les phases, les livrables et les acteurs. Ceci conduisait parfois à sacrifier les premiers mois de travail en commun, entre utilisateurs et informaticiens, pour en venir à travailler sur une compréhension bien partagée ou au pire se séparer sur un échec. Le constat était que l'enseignement de cette méthode variait notablement d'un établissement universitaire à un autre et qu'une harmonisation était nécessaire au sein de l'entreprise pour la rendre efficace. Une action a donc été lancée pour personnaliser la méthode au contexte de l'entreprise. Les responsabilités des acteurs ont été précisées, maîtrise d'ouvrage avec responsabilités budgétaires, de définition des besoins, de réception des livrables, d'une part, maîtrise d'œuvre, pour les choix techniques, la conception et la réalisation des systèmes, d'autre part. Les acteurs, leurs rôles, leurs relations ont été en quelque sorte positionnés sur un axe ressources humaines (RH). Les étapes, sur l'axe temporel (T), ont été bien ponctuées par des réunions de pilotage, des livrables, des points de décision. Les relations entre acteurs ont été précisées pour chaque phase, pour les tâches à accomplir. La méthode est toujours utilisée, plus de quinze ans après son adoption.

Sur le plan de la terminologie, les choix ont été très importants, les termes bien compréhensibles, bien définis et situés les uns par rapport aux autres ont permis d'entamer un projet sans hésitation sur le « qui fait quoi et quand ». Ainsi, même les plus haut responsables, les présidents de comité de pilotage, n'avaient qu'une soixantaine de pages à assimiler pour être à l'aise dans le projet et dans leurs responsabilités. Aux termes répandus dans les différentes variantes de la méthode Merise, il a fallu par exemple ajouter des notions nouvelles indispensables au bon déroulement d'un projet dans le contexte de l'entreprise. Ainsi ont été par exemple définis des « pôles de compétences », petites structures spécialisées dans une des technicités requises par un projet informatique et capitalisant sur les expériences acquises. Ce terme a par la suite fait flores.

Enfin il est intéressant de relever l'apport d'un bon nom de « baptême » pour l'acceptation d'une méthode. Un mémorable remue-méninges nous a conduits à l'acronyme suivant : MÉLODIC, pour **m**éthode pour **l'**organisation des **d**éveloppements **i**nformatiques **c**oncertés. Ce mot avenant donne à la fois l'objet de la méthode, l'organisation des développements informatiques, mais affiche aussi la façon privilégiée de faire : en concertation.

La méthode faisait aussi des choix sur la modélisation des données, mais cet aspect a présenté beaucoup moins de difficulté et a eu moins d'impact que la partie conduite de projet.

Cette expérience conduit à la recommandation générale et pressante suivante : avant de faire un travail en commun assurez-vous d'une compréhension partagée du domaine à traiter.

Voici quelques exemples d'outils disponibles pour faciliter cette compréhension commune :

D'abord un modèle de données selon les conventions Merise, où apparaissent les concepts utilisés (processus, risques...) et leurs relations (est porté par, est atténué par ...) caractérisées par des cardinalités (un processus peut porter de 0 à n risques, un risque doit être rattaché à un processus au moins...).

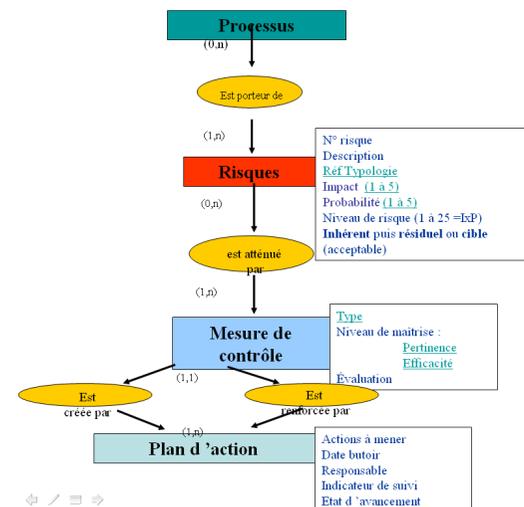


Fig. 1 : Une modélisation des données dans le domaine des risques :

Ensuite, une visualisation graphique, croisant axe chronologique et d'intensité, d'un choix raisonné de vocabulaire, élaboré avec des banquiers, pour décrire les composantes d'une période de crise. Le travail complet donnait les équivalents anglais des termes retenus, sur le graphique seuls les noms anglais des phases sont rappelés

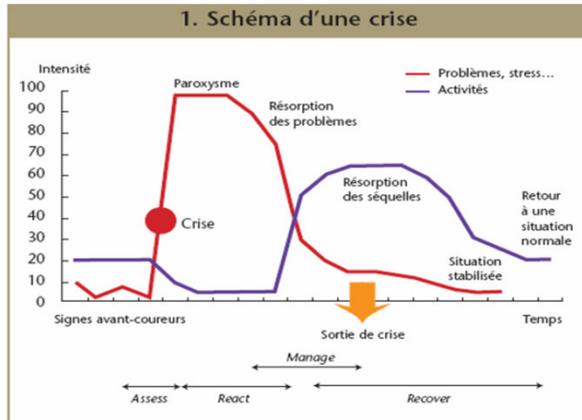


Fig. 2 : Visualisation d'un vocabulaire raisonné de crise.

Enfin, une présentation structurée partielle d'un lexique bilingue, pour faciliter la compréhension de l'emboîtement des concepts et proposer une traduction.

High level principles for business continuity
Principes directeurs en matière de continuité d'activité

Structured glossary **Lexique structuré**

Domain **Domaine**

Operational risk **Risque opérationnel**

Major operational disruption **Perturbation opérationnelle majeure**

Resilience **Résilience**

Business continuity **Continuité d'activité**

Business continuity management **Gestion de la continuité d'activité**

Business continuity plan **Plan de continuité d'activité**

Business impact analysis **Analyse d'impact sur l'activité**

Communication protocols **Protocoles de communication**

Recovery **Reprise**

Recovery objectives **Objectifs de reprise**

Recovery level **Niveau de reprise**

Recovery time **Délai de reprise**

1

Fig. 3 : Structuration d'un lexique relatif à la continuité d'activité

8

1.2. La sécurité de l'information.

Il s'agissait ici d'attirer l'attention des établissements financiers sur le caractère vital de leur informatique, la banque étant devenue à l'époque, en 1995, une usine à traiter l'information.

Un travail collectif a conduit à la rédaction d'un livre blanc mais aussi à l'introduction d'un nouveau facteur de sécurité pour caractériser une information.

Il préexistait trois facteurs de sécurité et plus généralement de qualité d'un système d'information : DIC, la disponibilité, la confidentialité et l'intégrité. Ceux-ci n'apparaissent toutefois pas suffisants pour caractériser la bonne connaissance de la genèse d'une information, les étapes de la constitution d'un résultat. Le choix s'est porté sur un facteur P pour « possibilité de preuve », de préférence aux mots « auditabilité », employé par les auditeurs informatiques, ou « traçabilité » qui est maintenant très répandu au-delà de la sécurité de l'information.

Par la suite les spécialistes ont introduit un facteur L de légalité, pour le respect des contraintes légales. Ces facteurs bien définis et bien distingués les uns des autres, facilitent par exemple la compréhension entre divers acteurs d'une entreprise : le responsable de la sécurité du système d'information (RSSI) partagera un souci de disponibilité avec les exploitants du centre de calcul, d'intégrité avec les gestionnaires des bases de données, de confidentialité avec le correspondant de la CNIL comme il en existe maintenant dans les grandes entreprises. Le facteur P sera examiné par les commissaires aux comptes et le responsable de la conformité se focalisera sur le facteur L.

Dans ce domaine des spécialistes indiquaient à l'époque que le facteur I suffisait. L'ajout du facteur P a permis de préciser le champ de ces deux facteurs de sécurité : l'intégrité est la cohérence entre les différentes occurrences d'une information dans une base de données ou un système d'information, la possibilité de preuve est la connaissance de l'historique de la création d'une information.

1.3. La continuité de l'activité.

Cette préoccupation plus récente, dont on a pris conscience de l'importante suite à différents événements extrêmes, dont l'attentat de septembre 2001, avait aussi besoin d'un vocabulaire bien défini. Une obligation réglementaire en la matière a été imposée en France en 2004 au secteur financier. Ce règlement contient par exemple les termes « chocs extrêmes » pour indiquer ce qu'il faut savoir surmonter et « mode dégradé » pour désigner un mode de fonctionnement nouveau que peut justifier une situation de crise.

Dans ce domaine une échelle de gravité est très utile pour bien situer ce sur quoi portent les investissements en continuité d'activité. Cette échelle peut aller de la non-qualité jusqu'au cataclysme, avec comme niveaux intermédiaires l'incident, l'accident, le sinistre, la catastrophe. Il faut se mettre d'accord sur des définitions objectives, par exemple la non-qualité, qui n'est qu'une situation perfectible du point de vue économique, sera en dehors du champ de la continuité, les incidents, événements qui provoquent une perturbation, peuvent être des signes avant-coureur (voir figure 2) ou être hors du champ, de même pour les accidents, qui pourraient être caractérisés par des séquelles définitives ou de longue durée. À l'autre bout de l'échelle, car on ne peut pas demander l'impossible, le cataclysme pourrait se caractériser par l'intervention des pouvoirs publics (état de guerre par exemple). Un des buts d'une bonne gestion de la continuité d'activité serait de résister à une certaine gamme de sinistres située

aussi à droite de l'échelle de gravité que possible, sans par exemple avoir à convoquer une cellule de crise, les décisions à prendre étant bien anticipées.

Pour se comprendre il y a donc à choisir les termes, les situer les uns par rapport aux autres avec si possible un élément objectif de différenciation. De tels attributs « palpables » ont fait en leur temps le succès de l'échelle de Beaufort, qui a caractérisé des mots en les associant avec des observations prises dans la nature.

En voici un exemple tiré de l'ouvrage « L'invention des nuages » par Richard Hamblyn, qui explique aussi pourquoi le choix, l'« invention », de mots nouveaux par Luke Howard, pour désigner les nuages a permis les progrès de la météorologie :

Un axe : la force du vent ;
des niveaux bien répartis, rattachés à des faits observables

0 Calme	la mer est comme un miroir ; la fumée s'élève verticalement
1 Très légère brise	il se forme des rides, mais il n'y a pas d'écume ; girouettes immobiles
2 Légère brise	vaguelettes courtes, leurs crêtes ne déferlent pas ; vent sensible sur le visage
3 Petite brise	très petites vagues, écume d'aspect vitreux ; petits drapeaux déployés
4 Jolie brise	petites vagues devenant plus longues, moutons nombreux ; petites branches en mouvement
5 Bonne brise	vagues modérées, allongées, moutons nombreux ; petits arbres en mouvement
6 Vent frais	des lames se forment, crêtes d'écume blanche plus étendues ; grandes branches en mouvement
7 Grand frais	la mer grossit, l'écume est soufflée en trainées, lames déferlantes ; arbres entiers en mouvement
8 Coup de vent	lames de hauteur moyenne, tourbillons d'embruns sur leurs crêtes ; rameaux cassés
9 Fort coup de vent	grosses lames ; leur crête s'écroule et déferle en rouleaux ; dommages aux bâtiments
10 Tempête	très grosses lames à longues crêtes en panache ; déferlement en rouleaux intense et brutal ; arbres déracinés
11 Violente tempête	lames exceptionnellement hautes ; mer recouverte de bancs d'écume blanche ; dommages
12 Ouragan	air plein d'écume et d'embruns ; mer entièrement blanche ; visibilité très réduite

15

Fig. 4 : Une échelle bien constituée

Un autre exemple sur le vocabulaire des événements fâcheux, qui consisterait à faire un choix pour pouvoir s'entendre sur les termes préjudices et dommages, souvent employés indifféremment mais à qui on peut ajouter une nuance, pour distinguer des effets matériels ou corporels :

En deux dimensions :

H Vulnérabilité (caractéristique d'une personne)



φ Défectuosité (caractéristique d'un objet)

Domage (corporel, psychologique...)

H



φ Préjudice (matériel, financier...)

t → Incident / Accident

Axe du temps et de la séquence : Cause → événement → conséquence



Fig. 5 : Un exemple de choix pour 4 termes selon 2 échelles

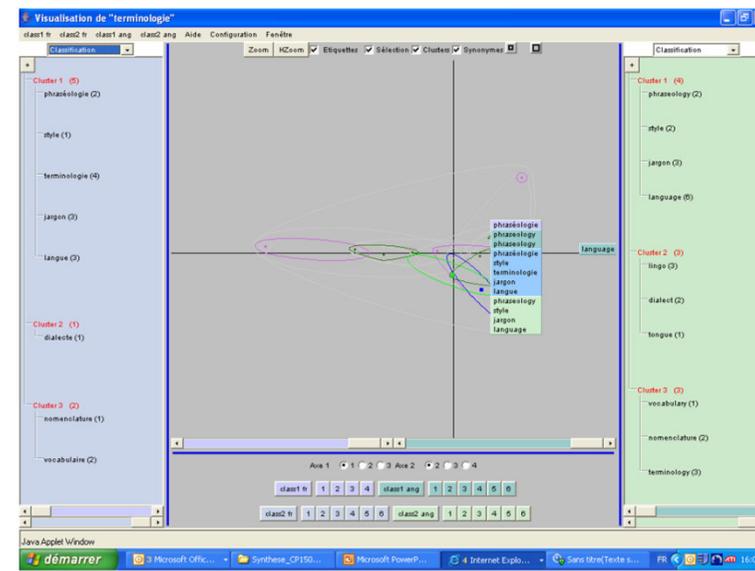


Fig. 6 : Les regroupements de mots du dictionnaire de l'ISC

1.4. Des outils informatiques d'accès aux ressources linguistiques

Deux outils issus de laboratoire du CNRS sont régulièrement utilisés avec grande satisfaction. Leur emploi suggère des prolongements, c'est ce que l'on tentera de formaliser dans la première suggestion.

Voyons d'abord un exemple du dictionnaire français-anglais de l'Institut des sciences cognitives de Lyon -ISC- (accès aux explications par : http://dico.isc.cnrs.fr/dico_html/fr/information2.html)

Ici sur le mot « terminologie » :

L'outil propose tous les mots de sens proche de celui retenu, il en calcule des proximités et crée des regroupements. Cette analyse de données donne une visualisation d'un espace multidimensionnel dont il est toujours difficile de caractériser les axes très composites. Par contre il devrait être possible de caractériser la nuance qui existe entre deux mots proches et de déterminer sur quelle échelle ils se distinguent.

Le second exemple est tiré du Trésor de la Langue Française informatisé du laboratoire ATILF du CNRS de Nancy 2

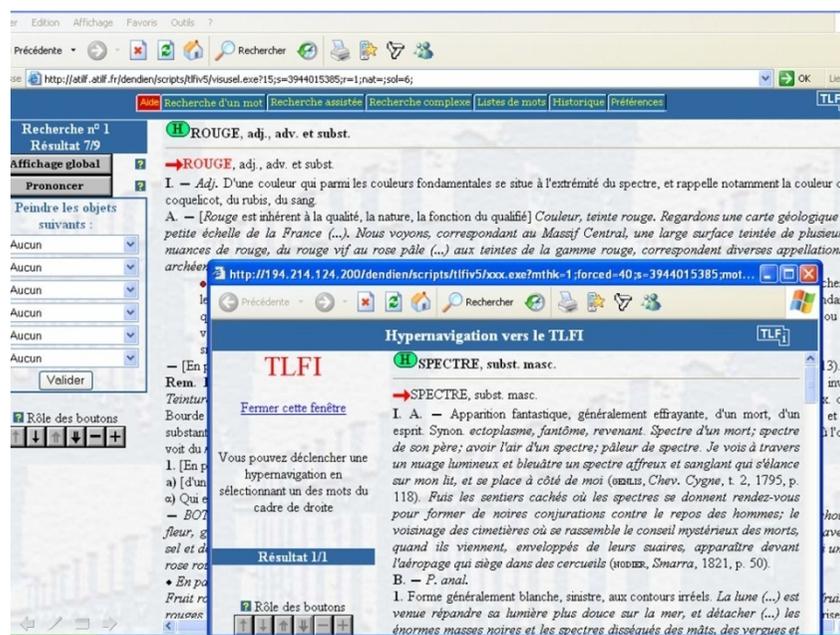


Fig. 7 : L'hyperlien du TLFI vers « spectre » dans la définition de « rouge ».

Ici dans la première phrase de la définition de « rouge » nous avons cliqué sur le mot « spectre », dont le sens implicite est celui du spectre chromatique, et la facilité d'hypernavigation nous conduit directement à sa définition. Toutefois il est donné la première entrée de la définition du spectre « apparition fantastique » (... qui ferait croire que les fantômes sont rouges alors qu'il est de notoriété publique qu'ils sont blancs livides !). Ici il conviendrait de pointer directement vers l'entrée II A « **II. A. — PHYSIQUE 1.** *Spectre (solaire, visible), spectre (de la lumière blanche).* .. » en tenant compte du contexte où le mot a été sélectionné.

2. Deux suggestions

2.1. Un dictionnaire « proxisémique » ?

Les considérations précédentes montrent l'intérêt de décrire les mots autant par de bonnes définitions que les écarts, différences ou nuances qui existent entre mots « voisins de sens », pour bien les distinguer et bien les comprendre.

La nuance entre deux mots voisins est porteuse de sens et peut souvent être le thème d'un axe différenciant. Cet axe tracé entre deux mots peut être prolongé et porter d'autres mots plus éloignés que les deux premiers. Nous arrivons ainsi à placer des mots sur une échelle dont le sens est connu, avec si possible un échelonnement régulier pour bien couvrir le domaine du thème de l'axe (échelonnement qui peut être en progression géométrique comme pour la gravité donnée dans le témoignage sur la continuité d'activité).

Pour illustrer ceci revenons au mot rouge, et à la consultation du TLFI à son propos. Rouge se situe indiscutablement sur un axe bien déterminé, celui des longueurs d'onde, ses voisins immédiats y sont l'orange et l'infrarouge. Rouge peut aussi être vu comme une couleur symbolique, une échelle qualitative peut être imaginée à son propos, ou plutôt qu'imaginée déduite des emplois du terme dans ce sens dans la littérature française, ses voisins les plus proches pourraient être le sang, le feu.

Il pourrait être intéressant de faire la même approche pour les mots propres, sur quels axes un personnage célèbre s'est-il exprimé, qui sont ses proches sur ces axes...

Un tel dictionnaire, qu'il est plus facile de concevoir informatisé qu'en édition papier, présenterait un mot avec des propositions de cheminement selon les axes pertinents pour ce mot, le déplacement donnerait les voisins successifs avec l'explication de la nuance qui les sépare, voire les points communs ou leur relation. L'axe complet, plus ou moins rectiligne devrait pouvoir être visualisé.

Un tel dictionnaire serait précieux pour la compréhension et la promotion de la langue française. Le choix du mot juste serait grandement facilité par rapport à l'existant des dictionnaires de synonymes, qui en fait propose des mots proches sans rappeler leur nuance ou axe de différenciation. Pour désigner ce dictionnaire, il faudrait un terme mélioratif, comme l'est trésor, il est proposé d'user d'un suffixe soit : « mélidico ».

Sa constitution représente certainement un investissement important, mais il semble se situer dans le prolongement des deux réalisations CNRS citées. Et il devrait pouvoir être partiellement généré automatiquement à partir de l'analyse de la littérature française.

2.2. Proximité visuelle pour les traductions

La seconde suggestion contrairement à la précédente, correspond à une formule simple de présentation des traductions déjà mise en œuvre¹, mais dont il est important d'explicitier l'intérêt. Cette fois la proximité en question n'est pas une proximité sémantique entre mots mais simplement de disposition entre paragraphes dans un ouvrage bilingue.

Le constat est le suivant : les textes internationaux produits par les institutions professionnelles tendent à prendre une forme universelle, donnant une succession de courts paragraphes numérotés de 1 à n contenant chacun le développement d'une idée. Ces textes peuvent contenir de termes difficilement traduisibles ou qui même n'ont pas d'équivalent dans la langue cible.

La formule classique de traduction est de faire un ouvrage séparé ou parfois en deux parties. La formule des pages face à face est plutôt réservée aux ouvrages littéraires. Dans ce cas on se demande quelle partie lire, l'originale qui fait foi, la traduction plus accessible mais dont les écarts naturels ne seront pas perçus.

Ce qui est proposé est simplement de promouvoir la succession des paragraphes en langue d'origine et en langue cible, la traduction venant immédiatement après le court paragraphe qui exprime généralement une seule idée. Ainsi on pourra rapidement en cas de difficulté avoir sous les yeux les deux expressions, et comprendre l'idée selon les deux cultures, ce qui pourrait se désigner par donner du relief aux idées.

Prenons un exemple réel, d'un document bilingue, et bicolore pour plus de lisibilité, qui en anglais présente des « *high-level principles* » ce qui donne en français des « principes directeurs », c'est volontairement que la notion de diriger est introduite. Une lecture séparée ne mettrait pas l'accent sur la traduction adéquate, mais non triviale, de « *board* » en « organe délibérant » et de « *senior management* » en « direction générale ».

¹ http://www.banque-france.fr/fr/supervi/supervi_banc/travinter/pca.htm

Cette présentation permet de mettre en évidence les faux amis, ces parents d'un même ancêtre qui se ressemblent toujours beaucoup (*hazard* et hasard) mais qui n'ont plus le même sens.

Ce document unique, tout en reconnaissant le « *global english* » comme langue de travail internationale, permet de défendre la langue française en n'obligeant pas à opter uniquement pour la lecture de la langue source, et en diffusant des documents accessibles aux anglophones qui trouveront comment régénérer leurs souvenirs du français.

Remerciements

Je tiens à remercier Jean-Yves Gresser qui a porté attention à ces suggestions de non spécialiste et m'a incité à les rédiger ainsi qu'à Christophe Roche qui m'a permis de les présenter en conférencier invité lors du premier séminaire Toth à Annecy en 2007.



Cette édition a été imprimée
en deux cent cinquante exemplaires
le treize mars deux mille neuf