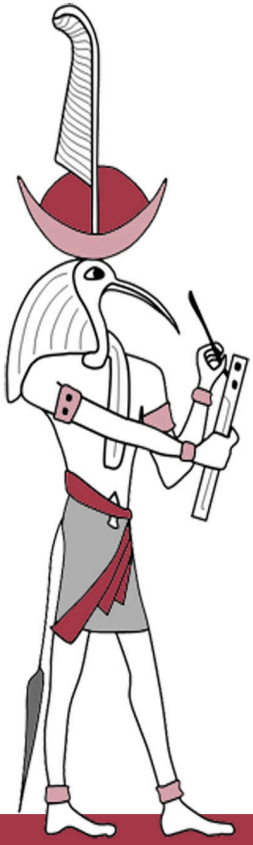
TOTh 2007 : Terminologie et Ontologie : Théories et Applications. Annecy 1er Juin 2007

Christophe Roche



TOTh

Terminologie & Ontologie : Théories et Applications

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

Institut Porphyre
Savoir et Connaissance

Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2007

Annecy - 1er juin 2007

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
- la Chambre de Commerce et d'Industrie de la Haute-Savoie
- l'association EGC (Extraction et Gestion des Connaissances)
- l'association AFIA (Association Française d'Intelligence Artificielle)
- la société Ontologos corp.

Institut Porphyre
Savoir et Connaissance

http://www.porphyre.org

Organisation

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 honoraire, Université de Lyon II*
Henri Zinglé	*Professeur, Université de Nice*

Comité de programme

Bruno de Bessé	*Professeur, Université de Genève*
Pierre Blanc	*EDF SEPTEN, représentant EDF à la commission terminologique de l'ingénierie nucléaire*
Marc van Campenhout	*Professeur, Institut Supérieur d'Interprétation et de Traduction de Bruxelles*
Stéphane Chaudiron	*Professeur, Université de Lille III*
Luc Damas	*MCF, Université de Savoie*
Sylvie Després	*MCF, Université René Descartes*
Anne Dourgnon-Hanoune	*EDF R&D*
François Gaudin	*Professeur, Université de Rouen*
John Humbley	*Professeur, Université Paris 7*
Hendrik Kockaert	*Professeur, Lessius Hogeschool (Anvers)*
Jean-Paul Haton	*Professeur, Université de Nancy 1*
Marie-Claude L'Homme	*Professeur, Université de Montréal*
Michel Léonard	*Professeur, Université de Genève*
Michel Simonet	*CNRS Grenoble*

Comité d'organisation : responsable : Luc Damas
Valérie Breasch
Joëlle Pellet

Avant propos

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 en 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.

La première édition des conférences TOTh a connu un franc succès. J'aimerais ici remercier, au nom du comité scientifique, tous ceux qui par leur présence et leur participation ont fait de TOTh une réussite.

C. Roche
Président du Comité Scientifique

Le mot du Président de la Société française de terminologie

La première édition des conférences TOTh, organisée en juin 2007, s'inscrit dans la continuité du colloque annuel organisé à l'École normale supérieure de la rue d'Ulm à l'initiative de la Société française de terminologie.

Ces conférences se sont révélées indispensables pour approfondir et diversifier les problématiques abordées lors de ce colloque annuel, en permettant aux spécialistes d'échanger et de collaborer sur les thèmes d'avenir. C'est ainsi que la question de la *terminologie et des sciences de l'information* (colloque de 2005) ou celle de la *terminologie et des ontologies* (colloque de 2006) se retrouve à différents moments de cette première conférence TOTh.

La correspondance entre ces deux types de réunion, en décembre et en juin de chaque année, permet une progression et un enrichissement mutuel. Abordées de façon souvent panoramique lors du colloque de Paris, les thématiques se retrouvent explicitées ou précisées lors des conférences TOTh d'Annecy.

L'un des enjeux qui apparaît fondamental pour les milieux de la recherche et de l'industrie est celui de s'efforcer d'accompagner de fondements théoriques fiables les applications, dans des domaines difficiles qui s'interpénètrent et qui tendent à créer de nouvelles voies de recherche. L'autre enjeu étant de permettre aux idées et applications nouvelles d'être débattues et de bénéficier rapidement à la communauté des chercheurs.

Nos remerciements vont aux organisateurs de ces conférences TOTh, en l'espèce l'Équipe Condillac de l'Université de Savoie travaillant sous la direction du Professeur Christophe Roche, et à tous les initiateurs de ces recherches, qui ont permis depuis 40 années qu'émerge en France un paysage de recherche riche dans les domaines de l'ingénierie de la connaissance.

Loïc Depecker
Professeur, directeur de recherches
Université de la Sorbonne nouvelle (Paris III)
Président de la Société française de terminologie

Table des matières

Le terme et le concept : fondements d'une ontoterminologie

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

Résumé :

La terminologie connaît depuis plusieurs années un tournant linguistique important. On s'intéresse aujourd'hui davantage aux mots et à leur utilisation en discours qu'à connaître les choses qu'ils peuvent dénoter. Si effectivement l'approche wüstérienne et l'approche normative sont difficilement applicables *stricto sensu* et que la terminologie a tout intérêt à s'approprier le *signifié*, il n'en demeure pas moins que tous les mots n'ont pas le même statut et que la terminologie ne se réduit pas à une lexicographie technoscientifique. La société numérique pose de nouveaux besoins, réclame une opérationnalisation des terminologies et réactualise le primat du concept pour de nombreuses applications – il suffit de penser aux problèmes que soulève l'ingénierie collaborative –. L'appellation *ontoterminologie* traduit ce besoin de replacer le concept et sa dénomination au centre de la terminologie, tout en préservant sa dimension sociolinguistique par la prise en compte des termes d'usage à travers la langue de spécialité. Si une conceptualisation se dit bien en langue naturelle, elle se définit dans un langage formel selon des principes épistémologiques où l'ontologie occupe une place prépondérante.

Plan

Le tournant linguistique
Les besoins d'opérationnalisation
La terminologie : un ensemble de pratiques
L'ontoterminologie
Conclusion

1. Le tournant linguistique

Les mutations technologiques et économiques de ces dernières années impactent profondément nos structures sociétales. La notion de *communauté* est devenue centrale, rendant encore plus cruciaux les besoins de communication et de partage de l'information et, par conséquence, les besoins en terminologie et en normalisation.

Si la terminologie, et de façon plus générale les langues de spécialité, connaissent un intérêt grandissant, force est de constater que la doctrine wüsterienne est difficilement applicable et que l'approche prescriptive soulève de nombreux problèmes tant au niveau de la définition des normes que de leur mise en œuvre. Les critiques des principes fondateurs de la terminologie semblent justifiées, et en particulier le premier d'entre eux, en citant le manuel de terminologie de Felber disciple de Wüster : *« Il convient de se rappeler que tout travail terminologique devrait être fondé sur des notions et non sur des termes »*. L'approche onomasiologique serait apparemment inadaptée à la réalité d'une pratique considérée avant tout comme langagière. Tout cela conduit à dénier à la terminologie un statut de discipline autonome et milite pour la « ramener » sous la coupe des sciences du langage.

La terminologie connait donc depuis plusieurs années un tournant linguistique indéniable, la réduisant parfois à une lexicographie technoscientifique. Il existe plusieurs raisons à cela. Avant tout parce que la terminologie est mobilisée au sein de discours liés à une pratique et relève donc de la langue, certes de spécialité. L'étude de la terminologie

se focalise alors sur les mots et leur utilisation en discours avec une attention toute particulière pour les textes – un mot isolé n'ayant pas de sens –. On s'intéresse plus aux expressions linguistiques qui dénotent les choses qu'à savoir ce que sont les choses. Aujourd'hui *être* c'est *être dit* et non plus *être pensé*. Une autre raison est la difficulté à cerner ce que peut être un concept et son rôle dans la détermination du sens du terme. La confusion entre *conceptualisation* et *classification* d'une part – il suffit de penser à l'approche prototypique –, et *sens* et *signification* d'autre part ; ont pour conséquence au pire le rejet du concept, au mieux sa réduction à un *signifié normé* ou à un réseau de mots liés par des relations linguistiques. Certains iront jusqu'à dire que Wordnet est une ontologie.

L'approche est séduisante. La langue est un système : les textes et les mots qu'ils contiennent constituent des données objectives sur lesquelles nous pouvons appliquer des méthodes scientifiques. La sémantique différentielle et la sémantique distributionnelle (étude des cooccurrences) en sont de beaux exemples et les résultats des plus intéressants. La terminologie relève *ipso facto* des sciences du langage.

Mais peut-on réduire la terminologie à une branche de la linguistique et oublier sa dimension conceptuelle ? Le fait qu'un terme puisse être mobilisé au sein de discours de façon similaire à un signe linguistique l'identifie-t-il pour autant à un tel signe réduisant du même coup le concept à un signifié ? Il est vrai qu'aujourd'hui la scène est davantage occupée par la dimension purement linguistique de la terminologie ; et lorsque l'on invite des experts d'un domaine c'est principalement pour qu'ils témoignent de cette dimension de leur activité et rarement de la conceptualisation et de la représentation des objets de leur domaine.

Cependant, on ne peut comprendre un discours (écrit ou oral) que dans la mesure où l'on partage une même culture. Ainsi, 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 un même *extralinguistique* qui par définition n'appartient pas à la langue. Cette culture commune, cet

extralinguistique, ne constituerait-il pas le cœur même de la terminologie ?

Que la linguistique puisse être mobilisée pour l'étude de la terminologie, c'est une évidence. Ce qui ne veut pas dire que la terminologie relève des sciences du langage. Toute pratique scientifique met en œuvre non pas une langue, mais plusieurs systèmes de signes – nous ne pouvons pas penser sans signe nous rappelle Frege –. Mais qui dirait que la chimie, la thermodynamique, la mécanique quantique, la conception de systèmes d'information relèvent de la linguistique alors que leur pratique, par la compréhension des objets du monde et par la recherche d'une langue la moins ambigüe possible, pour ne pas dire normalisée, les rattache *ipso facto* à la terminologie ?

On identifie aujourd'hui trop souvent la terminologie à sa manifestation langagière – *verbalisation* d'une pratique à travers une *langue de spécialité* dont l'étude relève bien de la linguistique – en oubliant que la *conceptualisation* et la *représentation* des objets du monde sont des activités centrales, si ce n'est premières, de la terminologie. Activités qui font de la terminologie une discipline scientifique à part entière. Et si la terminologie met en jeu différents systèmes sémiotiques – conceptualisation et représentation nécessitent leurs propres langages – elle n'est pas uniquement une science des signes, mais aussi une science des choses. Le fait que la structure lexicale ne se superpose pas à la structure conceptuelle du domaine en est une illustration.

Enfin, l'existence de disciplines définies comme autant de spécialisations de la terminologie : terminologie textuelle, terminologie conceptuelle, terminologie cognitive, socioterminologie, ethnoterminologie, etc. n'est pas le fait d'irréductibles partisans d'une autonomie de la terminologie, mais traduit bien le fait qu'elle ne peut être réduite à une branche d'une discipline donnée.

2. Les besoins d'opérationnalisation

Entendons-nous bien. Notre objectif n'est pas ici de nier les apports de la linguistique. Le *terme d'usage*, ce terme mobilisé par une *langue de spécialité*, donne bien lieu à interprétation et la terminologie a tout intérêt à s'approprier le *signifié*. L'analyse des discours scientifiques, la compréhension de documents techniques le réclament. Nous souhaitons, dans le cadre de cette présentation d'ouverture à notre conférence, insister sur le fait que la terminologie ne se réduit pas à une analyse du discours scientifique et technique, à la recherche du sens des termes ou à une lexicographie de spécialité. La terminologie est une discipline scientifique dont le principal objet est de comprendre le monde et de trouver les mots « justes » pour en parler. La terminologie est une discipline autonome qui requiert pour son étude de puiser à l'épistémologie, la logique et la linguistique.

Il existe en effet des domaines scientifiques et techniques qui nécessitent une conceptualisation du monde et la création de dénominations univoques de ses constituants. C'est-à-dire « *un moyen d'expression qui permette à la fois de prévenir les erreurs d'interprétation et d'empêcher les fautes de raisonnement* » pour citer à nouveau Frege. Ces domaines reposent d'une part sur une compréhension consensuelle des choses et d'autre part sur leur représentation à des fins de manipulation. Ils donnent lieu à la réalisation d'applications qui s'appuient sur la définition d'une théorie – comprise ici comme une conceptualisation permettant d'appréhender les objets du monde – qui permet une certaine *objectivité*[1] dans la description et la manipulation de faits –. Les aspects descriptif et raisonnement de ces domaines priment sur les discours auxquels ils peuvent donner lieu.

Prenons pour exemple les applications de l'ingénierie collaborative qui connaissent avec l'ère informatique un essor considérable. La

[1] La théorie est « objective » au sens où elle est acceptée et partagée par une communauté. Les descriptions et leurs manipulations le sont au regard de la théorie qui en contraint la forme et l'interprétation.

conception et la fabrication d'un produit, qu'il soit manufacturé ou non, repose de plus en plus sur la collaboration de communautés de pratique qui, bien que partageant une, ou partie, d'une même « réalité », peuvent différer tant au niveau de leur vision du monde que de la façon d'en parler. La solution ici ne réside dans une démarche de traduction des langues de spécialité, mais dans la définition d'un format d'échange, plus que d'une langue, reposant sur une conceptualisation et des dénominations consensuelles. Il est à ce propos important de souligner que les termes de cette interlangue seront d'autant plus acceptés que d'une part ils n'appartiennent à aucune des langues vernaculaires – il est plus facile de créer de nouveaux termes que d'imposer ceux d'une communauté –, et que d'autre part leur lecture permet de *comprendre* le système notionnel.

Pour être encore plus précis dans nos illustrations, la définition du modèle conceptuel d'un système d'information est un exemple typique d'une démarche terminologique « classique ». L'objectif est ici, avant tout, la définition d'une conceptualisation du domaine permettant de décrire les objets du monde qui puisse donner lieu à une représentation manipulable d'un point de vue computationnel (calculable par un ordinateur). C'est ensuite la recherche d'une dénomination univoque des termes dont la signification est le concept dénoté : l'approche est onomasiologique. Pour qui a assisté et participé à des réunions de conception de systèmes d'information, on peut être étonné – en fait la démarche est naturelle pour un scientifique – de ce souci constant de vouloir sortir de la langue naturelle et de ses ambiguïtés. Et systématiquement d'entendre : « Qu'est-ce que cela veut-dire ? Qu'est-ce que cela veut dire de façon précise ? Quel en est le sens exact ? ». Souci constant de s'extraire de tout discours pour se référer à un « socle » stable de connaissances défini dans un système formel dont la syntaxe et la sémantique sont clairement définies.

Il existe donc de nombreuses applications, en particulier les applications liées au traitement de l'information, qui réclament une opérationnalisation des terminologies. Des terminologies centrées sur la

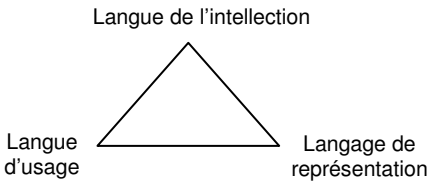
notion de concept qui soient *consensuelles*, *cohérentes*, *précises*, *partageables*, *réutilisables* et *calculables*. Autant de propriétés qui font que la terminologie ne peut relever du seul domaine de la linguistique.

Il nous semble important de rappeler que même lorsqu'on réduit la terminologie à une analyse du discours scientifique et technique, elle suppose une conceptualisation préalable du domaine [2]. Cette conceptualisation n'est pas du ressort de la linguistique. Sa définition relève d'une démarche épistémologique dans son appréhension des objets du monde et d'une démarche logique et computationnelle dans sa formalisation et sa représentation à des fins de manipulation. Elle met en œuvre un langage formel et un langage computationnel, dont les règles diffèrent de la langue naturelle. *Si une conceptualisation se dit bien en langue naturelle, elle se définit dans un langage formel selon des principes épistémologiques.*

3. La terminologie : un ensemble de pratiques

Le discours scientifique et technique « mélange » différents systèmes sémiotiques. La langue naturelle côtoie un langage symbolique et joue par rapport à ce dernier le rôle d'une métalangue – d'une glose – décrivant, expliquant, interprétant le langage symbolique. Ces systèmes sémiotiques ne répondent pas aux mêmes lois. Ils sont mobilisés par différentes pratiques qui ensemble constituent la terminologie proprement dite.

Ces différentes pratiques sont liées à la compréhension des objets du monde, à leur représentation à des fins de manipulation et aux discours auxquels ils



[2] Même dans le cadre d'une sémantique différentielle, la référence à cette conceptualisation est nécessaire à la compréhension des sèmes mais aussi à leur identification.

peuvent donner lieu. Même si ces pratiques sont liées, il est important de les distinguer et d'étudier les rapports qu'elles peuvent entretenir.

a. La langue de spécialité (langue d'usage)

Les discours[3] scientifiques et techniques relèvent de la *langue de spécialité*. Ils constituent, lorsqu'on étudie la terminologie d'un domaine sans en être un expert, la partie la plus visible et la plus directement accessible. Dans ce cadre, nous avons à faire à des *termes d'usage* qui donnent bien lieu à interprétation, à la recherche d'un sens qui se construit en discours. La notion de locuteur est centrale, et de façon plus générale la langue d'usage sous-entend la présence d'agents cognitifs tant au niveau de la production, et donc de l'intention, que de l'interprétation des discours. On s'intéresse ici aux rapports entre *signifiants* (termes d'usage) et *signifiés* en fonction d'un *contexte* donné. Cette pratique relève de la linguistique et de ses spécialités telles que la pragmatique.

L'analyse des discours, outre l'identification des termes d'usage, peut nous apporter une certaine connaissance du système notionnel[4]. Partant du fait que les documents scientifiques et techniques véhiculent des connaissances du domaine, il existe aujourd'hui de nombreux travaux qui portent sur l'extraction de connaissances, voire de terminologies, à partir de textes. L'existence de corpus numériques et l'utilisation de l'informatique permet d'obtenir des résultats intéressants (en particulier en sémantique distributionnelle). Cependant, il est indispensable de garder présent à l'esprit que l'incomplétude des textes est un des postulats de la linguistique textuelle. Ainsi, la compréhension des tropes suppose que l'auteur et le lecteur partagent un même *extralinguistique*. Mais comment prendre en compte l'intention de l'auteur à la base de toute interprétation, sachant qu'elle peut varier d'un texte à un autre au

[3] oraux ou écrits (textes).

[4] Par exemple la recherche des relations définitoires, des relations d'hyperonymie et de méronymie considérées comme des expressions linguistiques des relations de subsomption et de mérologie. C'est aussi l'étude des adjectifs substantivants *versus* qualifiants, etc.

sein d'un même corpus ? *In fine* il est important de souligner que les structures lexicales et conceptuelles que l'on peut extraire de textes ne se superposent pas avec la conceptualisation du monde : *dire n'est pas concevoir*. L'oublier c'est aboutir à des systèmes non réutilisables dépendants d'un corpus donné qui ne peuvent être qualifiés de systèmes notionnels ni de terminologies. *La variabilité du signifié ne permet pas de cerner la stabilité du concept*.

Se focaliser uniquement sur le discours scientifique et technique, c'est oublier que la terminologie résulte avant tout d'une activité scientifique. C'est-à-dire d'une activité qui consiste à comprendre, modéliser et représenter un réel et des modes de raisonnement dans un système formel afin de décrire, vérifier et prédire certains faits. Cette activité, propre à l'ingénieur, suppose d'une part la capacité à appréhender la réalité et d'autre part la capacité à l'exprimer dans une théorie donnée. Pour cela *il est nécessaire de redonner à l'ingénieur une place centrale au sein de la terminologie*.

b. La langue de l'intellection

L'appréhension des objets du monde repose, en terminologie, sur le *concept*. Défini comme une « *unité de connaissance créée par une combinaison unique de caractères* » (norme ISO 1087), il permet de regrouper sous une même appellation les objets qui partagent des propriétés communes.

Un des mérites de la terminologie classique est d'avoir insisté à la fois sur l'importance d'une expression extralinguistique des concepts comme un ensemble de *caractères* et sur leur organisation en tant que système : « *Toute notion occupe une place définie dans un système particulier de notions* » (Manuel de terminologie. Felber). La détermination d'une typologie de caractères – caractères distinctifs, essentiels – et de relations entre concepts – logiques, ontologiques, de combinaison – relève d'une préoccupation principalement épistémologique sur la nature des connaissances descriptives indépendamment de leur expression dans une langue donnée, qu'elle soit naturelle (« *La partie qui traite des notions s'applique à n'importe quelle langue* » *ibidem*) ou formelle.

La combinaison et la factorisation de caractères n'est pas la seule façon de définir un concept. La recherche d'attributs donnés par l'expérience, l'identification de propriétés essentielles issues de la raison, la définition de fonctions à valeur prédicative sont autant d'approches qui correspondent à des principes épistémologiques – et à des choix idéologiques : empirisme, métaphysique, positivisme logique – qui guident mais aussi conditionnent la construction du système notionnel. *La terminologie dépend directement de la théorie du concept qui la fonde.*

c. Les langages de représentation

La représentation du système notionnel à l'aide d'un langage formel répond à deux besoins. Le premier correspond à une démarche scientifique. L'utilisation de langages symboliques à la syntaxe et la sémantique clairement définies permet de nous affranchir des problèmes d'interprétation que pose la langue naturelle. Accepter les axiomes et les règles d'un système formel, c'est en accepter les constructions et donc le système notionnel. Le deuxième besoin est un souci d'opérationnalisation. Le système notionnel doit pouvoir donner lieu à un modèle calculable par ordinateur.

Les différents formalismes de représentation ne nous assurent pas tous des mêmes propriétés. Celles de cohérence et de consensus sont certainement parmi les plus importantes. Elles conditionnent l'acceptation de la terminologie et par conséquence sa réelle utilisation.

Bien que les « principes terminologiques » (*ibidem*) soient d'inspiration logique, ils n'en ont pas toutes les qualités – les opérateurs sur les notions et les manipulations des différents caractères ne sont pas formellement définis –. Il ne serait pas aisé d'en définir un formalisme et un modèle computationnel satisfaisant. C'est la raison pour laquelle, lorsque l'on souhaite opérationnaliser une terminologie, on se tourne généralement vers d'autres systèmes.

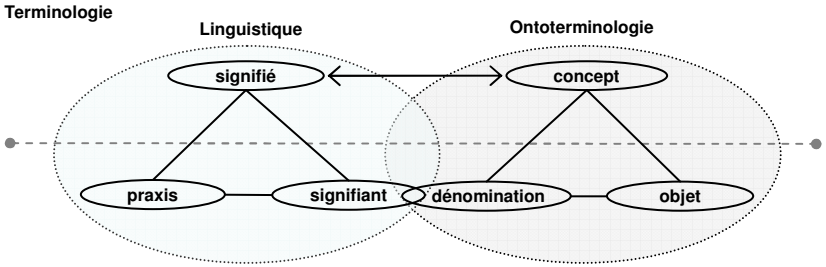
La logique joue un rôle important dans la formalisation d'une conceptualisation. Elle est l'archétype des systèmes formels dont la

syntaxe et la sémantique sont clairement définies. Le concept, fonction à valeur prédicative, est une *formule bien formée* et on dispose d'opérateurs et de mécanismes d'inférence pour la définition et l'exploitation des concepts. Enfin la logique est en elle-même un format d'échange. Autant de qualités qui nous garantissent certaines des propriétés recherchées dont la cohérence. A cela s'ajoute l'existence de logiques dédiées à la représentation des connaissances telles que les logiques des descriptions. *La logique est devenue aujourd'hui incontournable.*

Cependant les formalismes issus de l'intelligence artificielle demeurent les plus utilisés, principalement en raison de leur lisibilité : réseaux sémantiques, graphes conceptuels, systèmes à base de schémas. Le concept (ou classe) est défini par un ensemble d'attributs communs à ses instances. L'ensemble des concepts se structurent selon différentes relations : généralisation-spécialisation, partitive, etc.

d. Le triangle sémiotique

Distinguer les différentes pratiques – qu'elles relèvent du langage, de l'intellection ou de la représentation – qui toutes participent à la terminologie, c'est reconnaître à chacune son rôle et son importance sans vouloir en imposer une au détriment des autres. On peut ainsi reconnaître l'importance des termes d'usage (incluant variations terminologiques et figures de style) à côté des termes normés qu'il serait irréaliste de vouloir imposer. On peut également, par la séparation du concept et du signifié, permettre l'opérationnalisation des terminologies en garantissant un certain nombre de propriétés propres aux systèmes formels. En définitive, les notions mises en jeu – *signifiant*, *signifié*, *référent* versus *dénomination*, *concept*, *objet* – n'ont pas à être opposées à travers des triangles sémiotiques un peu réducteurs, mais ont tout à gagner à être mis en regard en insistant sur l'importance du contexte tant pour la définition de la conceptualisation (objectif, point de vue) que de la détermination du signifié (intention, interprétation).



un *double* triangle sémiotique

4. L'ontoterminologie

Si l'utilisation d'un langage formel permet de s'abstraire des problèmes d'interprétation et d'ambiguïté que pose la langue d'usage et d'assurer certaines propriétés comme la cohérence et l'opérationnalisation des systèmes notionnels, elle ne permet pas de résoudre tous les problèmes et en particulier celui de la construction des systèmes notionnels. La logique en est un des exemples les plus significatifs. L'introduction de la *rigidité de prédicat*[5] (rigidité ontologique) illustre bien l'existence de connaissances de nature différente – différence entre propriété essentielle et propriété accidentelle – nécessaires à la compréhension d'une conceptualisation. Introduction, *a posteriori* et non *a priori*, d'une propriété qui conditionne les descriptions du monde mais ne guide en rien leur construction. S'accorder sur la syntaxe et la sémantique d'un langage formel n'est pas suffisant. Un système formel est avant tout un système de *réécriture* (de formules) et non une théorie de la connaissance ou une théorie linguistique[6]. Le problème central

[5] Un prédicat est dit rigide si $\forall x\ [P(x) \rightarrow \Box P(x)]$; c'est-à-dire si P est vrai dans un monde possible, il l'est dans tous les mondes possibles. La rigidité relève *stricto sensu* d'une logique du second ordre (connaissance sur un prédicat).

[6] Les noms sont arbitraires et correspondent à des étiquettes sur des concepts.

demeure celui de la construction du système notionnel et du choix des principes épistémologiques sur lesquels se reposer.

La terminologie classique insiste avec raison sur l'importance d'une définition extralinguistique du concept sous la forme d'une combinaison de caractères (propriétés ou qualités d'un objet) et sur l'importance des relations qui lient les concepts entre eux. Elle propose de plus un certain nombre de principes pour la construction du système notionnel basés sur la nature des connaissances en jeu et en particulier des caractères : intrinsèques - extrinsèques, restrictifs, etc. La *définition spécifique* d'une notion en est un exemple.
Cependant il est à regretter une certaine confusion entre ce qui relève de préoccupations épistémologiques (classification des caractères, notions de genre et d'espèce, définition spécifique) et des systèmes formels (interprétation ensembliste des notions, opérateurs ensemblistes). Il en résulte trop d'imprécisions pour que les « principes terminologiques » puissent être directement utilisés : confusion entre notion et caractère dans leurs manipulations ; quelle est la définition intensionnelle de la notion résultante d'une disjonction ? Qu'en est-il des caractères distinctifs « hérités » par la conjonction (qui engendre une nouvelle espèce) de notions coordonnées créées par définition spécifique ? Et de façon plus générale, comment les opérateurs prennent-ils en compte la nature des caractères ? Comme si, sous l'influence d'un positivisme logique[7], on avait voulu faire passer ce qui relève de l'épistémologie – et donc d'une certaine façon de la métaphysique – sous les fourches caudines de la logique[8].
A cela s'ajoute un vocabulaire qui peut prêter à confusion sur l'emploi des mots *logique* et *ontologique* : la relation *genre-espèce* est qualifiée de « rapport logique » au même titre que l'intersection alors qu'elle relève de

[7] Le Cercle de Vienne pour ne pas le nommer.

[8] Le système notionnel donne bien lieu à une interprétation ensembliste et à une interprétation logique à condition de les définir de manière formelle.

l'ontologie[9] ; les relations entre objets sont qualifiées de « rapports ontologiques » alors que leur mise en relation nécessite de les définir au préalable – pouvons-nous parler d'une chose sans la connaître ? –, c'est-à-dire de définir l'ontologie.

Les principes épistémologiques qui permettent d'appréhender les objets du monde et la construction du système notionnel constituent une problématique à part. Ils relèvent de l'ontologie proprement dite.

a. Définition

Nous introduisons le néologisme *ontoterminologie* pour désigner cette approche qui place l'ontologie au centre de la terminologie. Une approche où l'ontologie joue un rôle fondamental à double titre : pour la construction du système notionnel et pour l'opérationnalisation de la terminologie. L'*ontoterminologie* insiste d'une part sur l'importance des principes épistémologiques qui président à la conceptualisation du domaine – c'est l'ontologie dans sa définition première –, et d'autre part sur la nécessité d'une approche scientifique de la terminologie où l'ingénieur joue un rôle fondamental – c'est l'ontologie dans ses définitions plus récentes –. Ainsi, les représentations formelles de l'ontologie permettent de « sortir » de la langue naturelle et de garantir certaines propriétés comme la cohérence, le partage et parfois le consensus. Et ses représentations computationnelles autorisent une opérationnalisation des terminologies – les modèles calculables par ordinateur jouent pour la terminologie un rôle similaire à celui qu'a pu jouer et que joue la logique pour le langage en définissant un cadre de vérifiabilité des propositions théoriques –.

Regardons en quoi l'ontologie, dans ses différentes acceptions, constitue une aide précieuse pour la construction du système notionnel, et le cas échéant pour la création de mots « justes » pour en parler.

[9] L'ontologie et son interprétation logique (prédicat, syllogisme) sont deux choses différentes.

b. Ontologie

L'ontologie [10], entendue comme « science de ce qui existe », constitue aujourd'hui une des voies les plus prometteuses pour la construction et la représentation formelle du système notionnel. C'est en particulier le cas pour la notion d'ontologie venant de l'intelligence artificielle, et plus précisément de l'ingénierie des connaissances. Issue de problèmes d'ingénierie collaborative au début des années 1990, elle vise des objectifs similaires à ceux de la terminologie classique : permettre la communication et l'échange d'information entre différentes communautés de pratique. Pour cela elle s'appuie sur une conceptualisation partagée d'un domaine sur laquelle repose la signification des termes. Les deux définitions suivantes résument la plupart des définitions existantes. La première insiste sur la dimension conceptuelle de l'ontologie : « *une ontologie est une conceptualisation d'un domaine – c'est-à-dire une définition formelle des concepts et de leurs relations – décrivant une réalité partagée par une communauté de pratique* » ; alors que la deuxième met en avant sa dimension terminologique – et normative – comme moyen de communication : « *une ontologie est un vocabulaire de termes dont les définitions sont données de manière formelle* ».
L'ontologie en ingénierie des connaissances est principalement un objet informatique, un moyen de représenter la réalité : *en intelligence artificielle, existe ce qui peut être représenté.* On comprend dès lors tout l'intérêt des ontologies pour la représentation du système notionnel et l'opérationnalisation de la terminologie. Mais le problème de leur construction reste entier.

L'ontologie est avant tout une théorie de la connaissance qui donne lieu à différents courants de pensée – et à différents principes épistémologiques – selon que l'on s'attache prioritairement à comprendre le monde où à le décrire tel qu'on le perçoit.

[10] Bien que le mot lui-même soit de création récente (généralement attribué à Christian Wolff « Philosophia prima sive ontologica » 1729), l'ontologie est la « science de l'être ». Elle relève dans son acception première de la métaphysique et remonte aux origines de la philosophie.

La définition des objets comme une somme de qualités perçues est une démarche naturelle et la plus immédiate. L'objectif ici n'est pas de comprendre le monde, mais de le décrire tel qu'il nous est donné, tel qu'on le perçoit à travers l'expérience, que ce soit par l'intermédiaire de nos sens ou de leurs prolongements technico scientifiques (appareils de mesure). Ces perceptions, que chacun partage parce qu'issues d'une expérience commune et sur lesquelles nous pouvons nous accorder (en particulier lorsqu'elles correspondent à des données scientifiques), définissent les qualités sur lesquelles se bâtit le système notionnel. Les concepts se construisent alors par abstraction, c'est-à-dire factorisation de qualités (caractères) communes : un concept est une « *unité de connaissance créée par une combinaison unique de caractères* » (ISO 1087-1), condition nécessaire et suffisante d'appartenance d'un objet à un concept. L'application itérative de ce processus d'abstraction aux différents ensembles de caractères permet de créer une structure notionnelle correspondant à un treillis[11] de concepts. Mais tout ensemble de caractères, s'il définit formellement un concept, n'est pas nécessairement porteur de sens. Cette démarche ne permet pas de prendre en compte les connaissances qui président à la formation et à l'organisation des concepts[12]. La factorisation de caractères reste une opération qui relève des systèmes formels.

Un concept est plus qu'une factorisation de qualités. Les connaissances qui structurent les concepts en système relèvent de la raison et non de la perception. La démarche ici concerne l'ontologie dans son acception première de « *science de l'être en tant qu'être indépendamment de ses déterminations particulières* ». On s'attache à comprendre ce que les choses sont, indépendamment de la façon dont elles peuvent être perçues. C'est-à-

[11] Ensemble (de concepts) muni d'une relation d'ordre partielle (inclusion sur les ensembles de caractères).

[12] Par contre cette approche, comme l'approche prototypique, semble bien adaptée à l'identification de concepts émergents dans un domaine en construction.

dire à rechercher les caractères *essentiels*[13] qui décrivent la nature de l'objet. Contrairement aux qualités – soumises « au plus et au moins » – qui décrivent l'état des objets, les caractères essentiels *définissent* et *différencient* les concepts. Issus de la raison et non de l'observation, ils structurent le système notionnel.
Le but est d'atteindre une description *stable* du monde sur laquelle on puisse s'accorder. Cette connaissance porte sur la *structure profonde* de la réalité. Souvent tacite[14], elle correspond à une *cristallisation* à un moment donné d'un savoir commun et partagé. Expliciter cette conceptualisation commune mais implicite est un problème difficile qui ne peut être résolu sans l'aide des experts du domaine. C'est la recherche de propriétés *objectives*, non pas de l'objet « en soi » indépendamment de tout observateur, mais de l'objet « pour soi » au regard d'une communauté de pratique. *L'ontologie est une modélisation d'une intersubjectivité.*

Quelle que soit la démarche – empirisme, métaphysique, logique – l'ontologie reste dépendante d'une pratique et non d'une langue qui découperait la réalité à la Sapir-Whorf. Elle n'est objective que dans la mesure où elle est partagée et acceptée par les membres d'une même communauté.

c. Dénomination

Bien que la majorité des termes des domaines scientifiques et techniques soient motivés au sens où leur forme reflète dans une certaine mesure la structure du système notionnel (« relais de tension »[15] par exemple), il n'est pas toujours aisé de distinguer ceux qui relèvent du discours de ceux plus directement liés à la conceptualisation (« relais de

[13] Un caractère est dit *essentiel* pour un objet si lorsqu'il est retranché de l'objet celui-ci n'est plus ce qu'il *est*.

[14] Cette conceptualisation tacite transparait à travers certains termes d'usage dans l'emploi par exemple d'adjectifs substantivants – preuve supplémentaire s'il en était besoin de l'intérêt de l'analyse linguistique de corpus.

[15] Les unités lexicales sont notées entre guillemets et les concepts entre les symboles inférieur et supérieur.

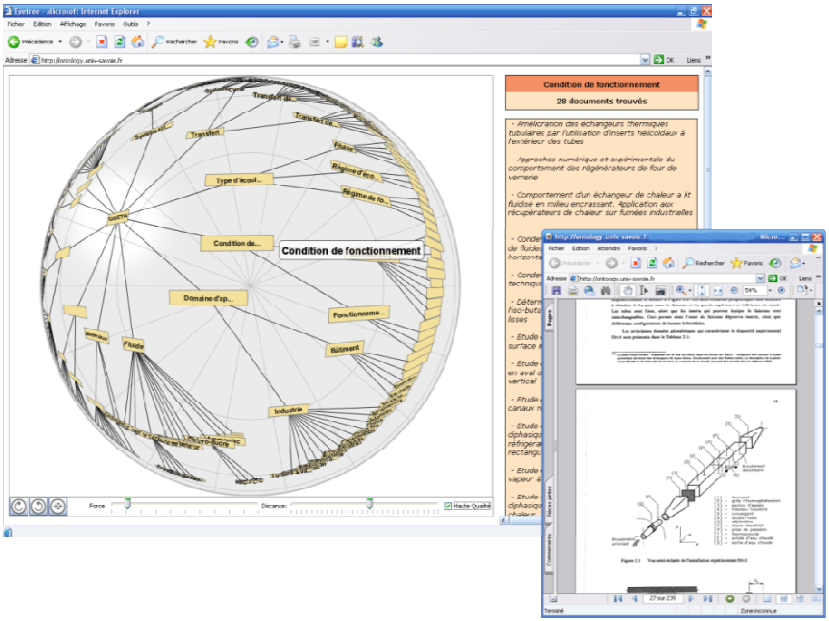
tension » désigne-t-il un concept de même nom ou est-il uniquement un terme d'usage ?). En distinguant les termes d'usage des termes normés, on redonne au processus de dénomination[16] tout son intérêt. Le nom d'un concept n'est pas arbitraire : sa forme traduit (devrait traduire) la place du concept dans le système notionnel (« relais à seuil de tension » désigne le concept <relais à seuil de tension> subsumé par le concept <relais à seuil>). Le choix de la théorie du concept pour la construction du système notionnel impacte donc également la dénomination des concepts. Une théorie prenant en compte des caractères distinctifs ou essentiels apporte une aide indéniable – la définition des concepts par différenciation spécifique est l'exemple type : le nom de l'espèce se construit à partir de celui du genre (régissant) et de la différence (modificateur) –. *A contrario* comment nommer les concepts dans une approche purement logique où il n'existe que des fonctions à valeur de vérité ? 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 (le syntagme « relais de tension » est une expression d'usage de l'expression normée « relais à seuil de tension » avec ellipse du premier modificateur. Il ne désigne pas un concept différent).

d. Exemple

La figure ci-dessous illustre une application de l'ontoterminologie à la gestion documentaire multilingue dans le domaine des échangeurs thermiques. Les documents sont indexés sur l'ontologie commune aux différentes communautés de pratique ; ce qui permet une recherche par concepts et non plus par mots clés. Ainsi, une recherche dans une langue donnée permet de retourner tous les documents, quelle que soit leur langue, associés aux concepts correspondant à la requête. Les termes d'usage et leurs relations linguistiques sont utilisés pour l'indexation automatique des documents et l'expansion (enrichissement) de requêtes.

[16] On préfèrera l'appellation *dénomination* pour insister sur une démarche onomasiologique à celle de *désignation* davantage liée à une utilisation du terme en langue.

L'exploitation de l'ontologie, également utilisée pour l'expansion de requêtes (extension aux concepts subsumés), permet d'obtenir des résultats pertinents sans perte d'information. Enfin, l'ontologie permet un accès interactif à la base documentaire multilingue par les concepts du domaine. On parle également de cartographie sémantique ou de navigation sémantique interactive.



Classification ontologique de documents multilingues

5. Conclusion

La terminologie ne peut se réduire à la seule étude des termes en langue. Ainsi, la détermination de leur sens requiert la connaissance préalable du système notionnel. De même, l'opérationnalisation des terminologies et la recherche de propriétés telle que la cohérence requièrent la mise en œuvre de systèmes formels détachés de tout discours. Le système notionnel, même s'il n'est pas toujours explicité, est au cœur de la démarche terminologique. Sa définition soulève de nombreux problèmes qui ne relèvent ni de la langue d'usage ni des formalismes de représentation.

La notion d'*ontoterminologie* met l'accent sur la dimension épistémologique de la terminologie dans son appréhension de la réalité. Elle permet de distinguer les pratiques – intellection, usage, représentation – et leurs fondements – terme d'usage *versus* dénomination, signifié *versus* concept –. Par la prise en compte de principes épistémologiques centrés sur la notion d'ontologie et de modèles computationnels respectant ces principes, l'ontoterminologie offre de nouvelles perspectives pour la construction de systèmes notionnels et leur représentation. Elle permet une construction du sens autour d'une sémantique référentielle et justifie l'intérêt de termes normés en regard des termes d'usage. En replaçant le concept et sa dénomination au centre de la terminologie, l'ontoterminologie redonne une place centrale à l'ingénieur dans son activité de conceptualisation. Enfin, elle propose des éléments de réponses aux enjeux de la société numérique et ouvre de nouveaux champs de recherche et d'applications.

Bibliographie

Manuel de terminologie. Helmut Felber, Paris, Unesco, 1984

Travail terminologique, NF ISO 704. AFNOR, ISSN 0335-3931, avril 2001.

Vocabulaire, NF ISO 1087-1. AFNOR, ISSN 0335-3931, février 2001.

La terminologie : nature et enjeux. Revue Langages n°157 mars 2005. Editions Larousse.

Le sens en terminologie. Henri Béjoint et Philippe Thoiron (dir.), ouvrage collectif. Presses universitaires de Lyon, 2000.

Les langues spécialisées, Pierre Lerat, Presses Universitaires de France, 1995.

Des fondements théoriques de la terminologie. Cahier du Centre Interlangue d'Etudes en Lexicologie 2004.

La terminologie noms et notions. Alain Rey, PUF, Que sais-je ? n°178, 2e édition corrigée, 1992.

Entre signe et concept : éléments de terminologie générale. Loïc Depecker, Presses Sorbonne Nouvelle, 2002.

Recent Trends in Computational Terminology. Special issue of Terminology 10:1. Edited by Béatrice Daille, Kyo Kageura, Hiroshi Nakagawa and Lee-Feng Chien, Benjamins publishing company, 2004.

Socioterminologie. Gaudin François (2003). De Boeck & Larcier s.a. 2003.

Sémantique et corpus. Sous la direction de Anne Condamines. Lavoisier 2005.

Handbook on Ontologies. Steffen Staab and Rudi Studer Editors, Springer, 2004.

A Translation Approach to Portable Ontology Specifications. Gruber T. Appeared in Knowledge Acquisition, 5(2):199-220, 1993.

Knowledge Representation. John F. Sowa, Brooks/Cole, 2000.

The Description Logic Handbook. Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, Peter Patel-Schneider. Cambridge University Press, 2003.

An Ontology of Meta-Level Categories. of Knowledge Representation and Reasoning. N. Guarino, M. Carrara, P. Giaretta. Proceedings of the Fourth International Conference (KR94), Morgan Kaufmann, San Mateo, CA, 1994.

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

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

Ontology Learning from Text: Methods, Evaluation and Applications. Frontiers in Artificial Intelligence and Applications, Vol. 123. Edited by Buitelaar P., Cimiano P., Magnini B., Ios Press Publication, July 2005.

Text analysis for ontology and terminology engineering. Aussenac-Gilles N., Soergel D. Applied Ontology. n°1. pp. 35-46, 2005.

Ontology for long-term knowledge. Dourgnon-Hanoune A., Salaün P., Roche C. XIX IEA/AIE, Annecy 27-30 June 2006.

Dire n'est pas Concevoir. Christophe Roche. 18èmes journées francophones d' « Ingénierie des Connaissances », Grenoble 4-6 juillet 2007.

La définition en terminologie : typologies et critères définitoires

Selja Seppälä
Laboratoire de terminologie/TIM/ETI/Université de Genève
Bd du Pont-d'Arve 40, 1211 Genève 4
selja.seppala@eti.unige.ch
http://www.unige.ch/eti/termino

Résumé :

La définition jouant un rôle important dans l'organisation et la transmission des connaissances (concepts) d'un domaine, nous nous proposons d'en étudier les caractéristiques en général et, plus particulièrement, en terminologie. Une brève étude des typologies de définitions nous amène à conclure à leur caractère inopérant et à nous tourner vers l'énumération des critères distinctifs des définitions, c'est-à-dire de leurs caractères. Nous proposons de caractériser les définitions utilisées en terminologie sur la base de ces critères. Nous concluons sur quelques questions soulevées par cet état de l'art sur la définition en terminologie.

1. Introduction

La définition joue un rôle important dans l'organisation et la transmission des connaissances d'un domaine en tant qu'elle est une des représentations possibles du concept (avec le terme, l'icone, etc.) (ISO 704, 2000). Elle figure dans les dictionnaires spécialisés et est parfois accompagnée d'autres modes de représentation. En terminologie, comprendre un concept suppose en connaître la définition. C'est elle qui permet de délimiter le concept grâce à une description de ses caractéristiques et par l'établissement de relations entre ses différents

éléments définitoires. C'est également elle qui permet de déterminer la place que le concept occupe par rapport aux autres concepts d'un domaine. Mais qu'est-ce qui caractérise plus précisément les définitions en terminologie ? Comment les définir dans toute leur complexité ? Qu'est-ce qui les distingue des définitions d'autres types de dictionnaires, par exemple ?

Pour tenter de répondre à ces questions, nous nous pencherons dans un premier temps sur les typologies de définitions en général. Ce bref panorama nous amènera à conclure à une certaine opacité voire un certain manque de cohérence des typologies, que nous proposerons de clarifier en dégageant les caractères autour desquels s'organisent les définitions, leurs critères définitoires. Ces critères, 17 au total, constituent autant de points de vue sur la définition et donnent souvent lieu à une dénomination propre (d'où l'existence de nombreux types de définitions). Dans chaque cas, nous nous attacherons à caractériser les définitions employées en terminologie suivant ces mêmes critères, afin de mieux comprendre leur nature. Nous conclurons sur quelques questions soulevées par cet état de l'art sur la définition en terminologie.

2. Remarques sur les typologies de définitions

La problématique de la définition en tant qu'objet d'étude est vaste. Elle englobe une multitude de types de définitions regroupés autour de différentes facettes, dont un certain nombre sont récurrentes (Blanchon, 1997). Blanchon relève, ainsi, trois axes majeurs repris par différents auteurs pour articuler leurs typologies des définitions : le premier rend compte de l'opposition « entre la lexicographie traditionnelle et la terminologie » et distingue principalement les *définitions lexicographique*, *terminologique* et *encyclopédique*, même si des distinctions plus fines sont parfois proposées ; le deuxième « concerne le contenu logique des définitions terminologiques » et oppose entre autres les *définitions en compréhension*, *en extension*, ou encore *générique* ou *partitive* ; le dernier axe relève « plus de la structure des définitions », (*définition synonymique*, *paraphrastique*, *métalinguistique*, *par analyse*, etc.). Malgré l'intérêt évident que

présente ce constat pour comprendre les caractéristiques des définitions, il n'en reste pas moins limité, car l'auteur n'explicite pas davantage ces dimensions.

On trouve également dans la littérature deux autres pôles d'articulation des typologies : la nature du défini (*définition de mot* vs *définition de chose*), ainsi que son rôle (*définition descriptive* vs *définition prescriptive* ou *créatrice de concept*). Si ces cinq dimensions sont les plus significatives, il en existe une sixième, plus rarement évoquée, qui est fondée sur les moyens utilisés pour définir et qui regroupe notamment les *définitions ostensive* et *par paraphrase*. On notera, curieusement, qu'aucune typologie ne semble s'articuler autour des fonctions de la définition, aspect pourtant essentiel pour comprendre la définition dans toute sa complexité, comme en témoigne la récurrence des réflexions à ce sujet dans la littérature. Sans compter qu'une analyse plus approfondie (cf. point 4) permet d'identifier de nombreuses autres dimension (nous en distinguons 17). Mais ce ne sont pas là les seules limites des typologies.

Une étude plus détaillée de différents auteurs montre, en effet, que certaines typologies sont fondées sur des critères de classification peu clairs ou non précisés, par exemple lorsque les caractères distinctifs ne sont pas énoncés de manière explicite, voire sont totalement absents. On trouve, en outre, des typologies qui opposent des types de définitions dont les traits distinctifs sont de nature différente, non exclusifs les uns des autres, donc souvent compatibles. Sager (1990) oppose, par exemple, la *definition by analysis*, qui se distingue par sa forme, à la *definition by demonstration* ou *ostensive definition*, qui se distingue par le type de moyens utilisés pour définir.[17] Pour terminer, il arrive que les définitions ou les explications concernant les *types de définitions* évoqués manquent de clarté ou soient totalement absentes, les auteurs se contentant de les nommer, en supposant peut-être que leur seule dénomination suffit à en saisir

[17] Pour d'autres exemples, se reporter à (Seppälä, 2004), où sont également développés la plupart des autres aspects.

toute la portée. Faut-il en conclure que les typologies ne sont pas le terrain privilégié pour aborder l'étude de la définition ?

Malgré la diversité des typologies proposées, que ce soit par des linguistes, des lexicographes ou des terminologues, il semble difficile de se fonder uniquement sur celles-ci pour établir une théorie de la définition complète et cohérente, qui rende compte de ses caractéristiques dans divers contextes, et encore moins pour comprendre les caractéristiques et le fonctionnement des définitions en terminologie.

Ce bref tour d'horizon des typologies de définitions met ainsi en lumière les limites des approches typologiques, qui ne sont pas systématiques et qui, étant fondées sur des définitions nommées, ne permettent pas d'aborder les aspects des définitions qui n'impliquent pas un *type de définition* donné, avec une dénomination particulière (cf. les discontinuités sémantiques des langues (Weinreich, 1970 [1962])). Malgré ce constat, nous pouvons néanmoins tenter de comprendre sur quoi se fondent les différentes typologies et en fonction de quels critères elles peuvent être créées. Cette approche descriptive permet ainsi d'aborder tous les aspects de la définition[18] sans que l'on se retrouve prisonnier de l'une ou l'autre typologie existante ni, surtout, des dénominations des définitions.

3. Méthodologie

C'est donc sur la base de ces typologies, mais aussi d'autres études plus descriptives que nous avons mené un travail de dépouillement systématique de la littérature relative aux définitions. Ce travail plus analytique, a consisté à faire l'inventaire des différentes informations disponibles sur les définitions, puis à les organiser, à les classifier, à faire des recoupements, pour identifier les différents types de caractères

[18] À l'exception des principes et conventions de rédaction des définitions, que nous écartons délibérément de cet exposé.

pouvant être utilisés pour définir les définitions en général et, plus particulièrement, en terminologie.

En nous penchant sur les différents caractères susceptibles d'être pris comme critères distinctifs des définitions et pouvant donner lieu à un type de définition particulier, nous écartons délibérément l'approche classificatrice pour adopter une méthode plus analytique. Celle-ci nous a permis d'établir une sorte de grille de lecture des définitions, à l'aide de laquelle il est possible de caractériser les définitions en terminologie.

Une définition peut en effet être définie de différentes manières, selon un ou plusieurs critères, par exemple suivant sa forme ou sa fonction, ou les deux à la fois. Selon notre approche, chaque *critère distinctif* correspond à l'un des 17 caractères de la définition (voir ci-dessous), où chaque *caractère* (figurant dans les sous-titres) peut prendre différentes *valeurs* (indiquées en gras), voire des *sous-valeurs* (indiquées en gras et italique). Ainsi, le caractère *fond*, qui se rapporte au type de sens[19] défini, peut prendre la valeur **chose-nommée** (c'est-à-dire la dénotation ou le sens référentiel), et la sous-valeur ***concept général***. La liste des 17 caractères proposés est la suivante :

Signalons, que bon nombre de caractères sont liés, l'un étant par

1. NATURE
2. SITUATION D'EMPLOI
3. MOYEN
4. MODALITE
5. FOND
6. TYPE DE DESIGNATION
7. PROPRIETE METALINGUISTIQUE
8. MODE
9. FORME
10. COMPOSANT
11. TYPE D'INCLUANT
12. TYPE DE SPECIFIQUE
13. PERTINENCE DES CARACTERES
14. FONCTION
15. ROLE
16. NIVEAU DE SPECIALISATION
17. DESTINATAIRE

[19] *Sens* est ici employé comme concept superordonné de *signifié* et de *concept*.

exemple la conséquence d'un autre. L'intérêt de les distinguer est que certains types de définitions ne s'articulent qu'autour de l'un ou l'autre des caractères et non des deux à la fois. Ils peuvent donc être considérés indépendamment les uns des autres. Nous choisissons néanmoins de mettre en avant ces relations en traitant les caractéristiques qui sont fortement liées dans une même partie.

Une définition d'un *type de définition* peut, par ailleurs, combiner plusieurs caractères, de même que certains caractères admettent plusieurs valeurs à la fois à l'intérieur d'une même définition. Ainsi, un type de définition peut être défini à la fois par sa *forme* et par plusieurs *fonctions*.

4. Critères définitoires des définitions

Dans cette section, nous passons en revue chacun des 17 caractères pouvant servir à définir un type de définition, en soulignant dans chaque cas les valeurs qui s'appliquent aux définitions en terminologie.

4.1. NATURE (1) et SITUATION D'EMPLOI (2) de la définition

Avant toute chose, il convient de préciser ce qu'est par essence une définition, sa *nature*. Elle correspond en fait à deux concepts : 1.1) une **opération** qui vise à produire une représentation d'un sens, et son résultat, c'est-à-dire 1.2) la **représentation** elle-même (d'après *Nouveau Petit Robert*, 2006). C'est le résultat qui nous intéresse en premier lieu et que nous souhaitons ici définir, puisque c'est lui qui apparaît dans les ouvrages dictionnairiques et, plus spécifiquement, dans les produits terminographiques.

Le besoin de définir peut, par ailleurs, survenir dans différents types de situation. La diversité des *situations d'emploi* implique ainsi différentes façons de définir, et partant, différents types de représentation. Nous en relevons principalement trois. Les activités humaines induisent, d'une part, des situations de 2.1) **communication**

générale variées et nécessitent des niveaux de connaissance variables qui appellent chacun un type de définition qui lui est adapté (communication parent-enfant, à l'école, etc.). D'autre part, si d'après Rey (1992), 2.2) toute **pratique** (le droit, la mathématique, la métaphysique, etc.) aurait également ses propres définitions, c'est à plus forte raison le cas des 2.3) **disciplines dictionnairiques** telles que 2.3.a) la ***lexicologie***, 2.3.b) la ***lexicographie spécialisée***, 2.3.c) la ***terminographie*** ou 2.3.d) l'***encyclopédie***.

En ***terminographie***, la **discipline dictionnairique** de la terminologie, les définitions sont donc des **représentations** de sens.

4.2. MOYENS (3) et MODALITES (4) définitoires

Deux catégories de *moyens* peuvent être employés pour représenter un sens, suivant différentes *modalités* ou manifestations concrètes de la représentation. On distingue ainsi 3.1) les **moyens non langagiers** et 3.2) les **moyens langagiers**.

Les premiers peuvent s'actualiser selon les modalités suivantes : 4.1.a) l'**ostension**, 4.1.b) l'**iconicité** (avec des degrés d'abstraction variables (ISO 704, 2000)), 4.1.c) la **sémiotique** (Depecker, 2000), ou 4.1.d) les **représentations formelles** (par exemple en traitement automatique des langues). Les moyens de représentation extralinguistiques ne sont généralement pas considérés comme des définitions, même s'ils peuvent parfois remplir la même fonction. Dans le contexte dictionnairique, et donc terminologique, ils sont plutôt considérés comme des compléments de la définition ; nous les écartons donc de notre analyse.

Par ailleurs, tout moyen langagier n'est pas forcément non plus considéré comme une définition. Si 4.2.a) les **modalités lexicales** – qui comprennent 4.2.a.i) l'***unité lexicale*** synonyme ou équivalente dans une autre langue, 4.2.a.ii) la ***locution***, voire 4.2.a.iii) le ***morphème*** – jouent parfois ce rôle dans les dictionnaires de langue ou les dictionnaires bilingues, s'agit-il pour autant de véritables définitions ? Ce qui est

certain, c'est que 4.2.b) le **syntagme libre**, comme celui des mots-croisés, destiné à « évoquer » ou à « faire deviner » (Rey, 1992), lui, ne l'est pas. En fait, seules deux modalités langagières le sont : 4.2.c) la **prédication définitionnelle** (Rey-Debove, 1971), c'est-à-dire une *proposition* (de Bessé, 1996) composée de deux membres – le *défini* ou *definiendum* (*thème*) et la *définition* ou *definiens* (*prédicat*) –, unis par une *copule* (exemple 1 ci-dessous), que l'on trouve dans les textes, et 4.2.d) la **définition** (ou *definiens*) à proprement parler (exemple 2), considérée comme 4.2.d.i) une ***paraphrase*** ou 4.2.d.ii) une ***périphrase***. (Pour une explication de cette distinction, voir Rey-Debove, 1971.)

Ex : 1. *(La)* baguette *(est un)* bâton mince et flexible. 2. Baguette *(signifie)* bâton mince et flexible. ⇒ défini + (copule) + définition (Rey-Debove, 1971)

Suivant ce découpage, la définition en terminologie est une représentation qui fait appel à des **moyens langagiers**. La modalité définitoire à laquelle les ouvrages terminologiques ont recours est celle de la **définition**, c'est-à-dire le dernier élément de la prédication définitionnelle qui suit le défini et la copule d'identité, laquelle est sous-entendue. En ce sens, les définitions en terminologie peuvent être considérées comme des ***périphrases***, soit des « Figure[s] qui consiste[nt] à exprimer une notion, qu'un seul mot pourrait désigner, par un groupe de plusieurs mots » (*Nouveau Petit Robert*, 2006), ayant valeur d'équivalence avec le défini.

4.3. FOND (5) de la définition, TYPE DE DESIGNATION (6) associé et PROPRIETE METALINGUISTIQUE (7)

Outre les moyens et les modalités, toute définition peut être caractérisée par son *fond*, c'est-à-dire le type de sens, et corrélativement, le *type de désignation*, d'unité signifiante, visés. Le fond sémantique ou conceptuel de la définition peut porter sur au moins deux types d'éléments, selon qu'il s'agit de décrire 5.1) le **signe-nommant**, c'est-à-dire la signification du signe linguistique, du mot lui-même (voir

l'exemple de *péjoratif*), soit 5.1.a) le ***sens autonymique***[20], ou 5.2) la **chose-nommée** (Martin, 1992, Rey-Debove, 1971), c'est-à-dire la dénotation ou le sens référentiel (Sager, 1990), qui recouvre notamment le *signifié* et le *concept* (Depecker, 2000, Rey, 1977).

Des distinctions plus fines encore peuvent être établies pour la chose-nommée. Le sens ainsi défini peut correspondre : au 5.2.a) ***sens des morphèmes*** d'une unité lexicale, à son 5.2.b) ***signifié***, au 5.2.c) ***concept général*** ou au 5.2.d) ***concept unique*** (ISO 704, 2000) qu'elle désigne, ou encore à la 5.2.e) ***connaissance encyclopédique*** du monde qu'elle évoque.

Les deux types de contenus peuvent en outre coexister pour un même signe (Rey-Debove, 1971), soit séparément, soit dans la même proposition, comme le montre l'exemple suivant :

Ex : *péjoratif, ive* adj.♦ Se dit d'un mot, d'une expression, d'un élément, d'une acception qui comporte une idée négative, déprécie la chose ou la personne désignée. (*Nouveau Petit Robert*, 2006) [Nous soulignons.]

Chaque type de fond renvoie à un *type de désignation*, d'unité signifiante. Ainsi le sens autonymique renvoie 6.1.a) au **signe** ; le sens des morphèmes et le signifié sont associés 6.2.a) au **mot** ; le signifié renvoie également 6.2.b) à la **locution** ; le concept général renvoie 6.2.c) au **terme** ; 6.2.d) l'**appellation** désigne, quant à elle, les concepts uniques (ISO 704, 2000) ; et les connaissances encyclopédiques sont principalement associées 6.2.e) au **nom propre**.

Le fond renvoie également à un autre caractère de la définition, à savoir sa *propriété métalinguistique*. La définition correspond en effet à une métalangue (Weinreich, 1970 [1962]) : dans certains cas, à une 7.1) **métalangue de signe**, lorsque le fond en est le signe, dans les autres, à

[20] « Qui se désigne lui-même comme signe dans le discours[…]. *Dans « très est un adverbe »,* très *est autonyme.* » (*Nouveau Petit Robert*, 2006)

une 7.2) **métalangue de contenu**, lorsqu'elle exprime « la substance du défini (le défini en soi) ». (Rey-Debove, 1971).

En terminologie, la définition occupe (avec la vedette et le domaine) une place capitale dans l'article associé à un **terme**. Ce type de désignation correspond à un type de fond particulier : la **chose-nommée** et plus précisément le ***concept général*** ou le ***concept unique*** qu'il désigne, car bien que le défini (le terme) soit d'apparence souvent indifférencié du mot, ce n'est pas sa forme signifiante qui est définie. La propriété métalinguistique des définitions figurant dans les dictionnaires ou bases de données terminologiques est donc celle de **métalangue de contenu**.

4.4. MODE (8) et FORME (9) de la définition

Parmi les principaux traits invoqués pour caractériser les définitions notons la forme logique, soit son *mode* définitoire, et la *forme* définitoire qu'elle implique, donc les structures logique et concrète de la définition. Les modes peuvent être répartis en quatre catégories :

8.1) Le **mode conceptuel**, lié aux concepts et aux systèmes auxquels ils participent. Ce mode correspond à 8.1.a) la ***compréhension*** et donne lieu à la forme définitoire 9.1) **en compréhension**. Il se caractérise par la présence d'un concept plus général, superordonné (l'*incluant* ou le *générique*), et d'au moins un concept spécifique ou différentiateur (appelé *caractère* ou *spécifique*), qui ramène le genre à une espèce et distingue le concept à définir des autres concepts appartenant au même système (de Bessé, 1996). Ce mode définitoire, qui prend la forme *générique* + *spécifique(s)*, est considéré comme le modèle classique de la définition et peut s'appliquer à tous les types de contenu définitoire, que l'on définisse le signe lui-même ou sa dénotation.

8.2) Le **mode référentiel**, plus ancré dans la réalité, correspond à 8.2.a) l'***extension***, dont découle la forme définitoire 9.2) **en extension**. L'extension « représente l'ensemble des objets auxquels s'applique [un] concept[,] » (Depecker, 2000) et se traduit par l'énumération de « toutes

les espèces situées au même niveau dans le système conceptuel, voire même de tous les objets individuels. » (de Bessé, 1996)

8.3) Le **mode langagier**, qui s'inscrit dans une perspective linguistique, caractérise les définitions qui s'attachent à décrire le sens d'un mot 8.3.a) ***par le contexte***, 8.3.b) ***par l'exemple***, ou 8.3.c) ***par renvoi*** à d'autres unités lexicales. Les deux premières s'actualisent sous la forme de 9.3.a) **contextes d'usage** et 9.3.b) d'**exemples**. Une définition qui opère par renvoi peut, elle, prendre différentes formes : 9.3.c) la **synonymie**, 9.3.d) l'**antonymie** ou 9.3.e) l'**équivalence** dans une autre langue.

Finalement, on relève l'existence de 8.4) **modes combinés** ou mixtes, qui caractérisent des définitions intégrant plusieurs modes et donc plusieurs formes dans une même phrase – forme définitoire 9.4) **mixte**.

Si, en terminologie, le fond des définitions reste toujours un concept, leur structure tant logique que linguistique ou définitoire peut varier. S'agissant du mode logique, le plus fréquent est le **mode conceptuel**, qui correspond à la forme logique de la ***compréhension*** et qui s'actualise dans la forme définitoire **en compréhension** : générique + spécifique(s). Il arrive que la définition suive le **mode référentiel** de l'***extension***, correspondant à la forme définitoire **en extension**, ou encore le **mode combiné**, aboutissant à des définitions de forme **mixte**. On trouve, en effet, des définitions qui sont à moitié en compréhension et en extension. En revanche, en terminologie, les modes langagiers (par renvoi ou par le contexte ou l'exemple) ne sont généralement pas considérés comme des définitions, le type d'information qu'ils recouvrent étant assigné à des rubriques propres (*terme*, *synonyme* ou *note*) de l'article terminologique.

4.5. COMPOSANTS (10) de la définition

Ces différents modes définitoires et les formes qui en découlent impliquent différents types de *composants*, les éléments définitoires ou

définissants (Rey-Debove, 1971). Parmi ces éléments, certains se réfèrent à la dénotation et constituent une sorte de description de traits conceptuels : l'indication de 10.1) **domaine**, qui dans certains cas participe de façon significative à l'explicitation du sens et peut de ce fait être considérée comme composant de la définition, 10.2) le **générique** (voir point 4.6.) et 10.3) le **spécifique** (voir point 4.7.) ; d'autres, 10.4) les **espèces isonymes**, renvoient directement aux référents, par exemple *Terre*, *Mars*, *Saturne*, etc. pour *planète*. Toutefois, on ne peut (efficacement) définir quelque chose en en énumérant les seules propriétés, les spécifiques, sans faire de rapprochement avec autre chose de (supposément) connu, sans mentionner qu'il s'agit d'un objet (Rey-Debove, 1971). En terminologie, une définition se doit donc de débuter par un **générique** (de Bessé, 1996) ou **incluant** (Rey-Debove, 1971), qui fournit une information catégorisante ; « C'est la réponse naturelle minimum à la question « Qu'est-ce qu'un X? » […] » (Rey-Debove, 1971).

Quel que soit le mode définitoire, les définitions en terminologie intègrent toujours le **domaine**, même s'il figure généralement dans un champ distinct, car sans domaine pas de terminologie. Celui-ci fait, en effet, partie intégrante du concept et est de ce fait au moins virtuellement présent dans la définition. Lorsque la définition est en extension, ses composants sont l'ensemble ou une partie des **espèces isonymes** que recouvre le concept défini, mais comme dans la plupart des cas les définitions sont en compréhension, les composants les plus fréquents sont le **générique** et le **spécifique**.

4.6. TYPES D'INCLUANTS (11) ou génériques

Rey-Debove (1971) distingue principalement deux *types d'incluants* : les vrais incluants et les faux incluants.

11.1) Les **vrais incluants** comprennent 11.1.a.i) l'***hyperonyme*** ou le ***genre***, ainsi que 11.1.a.ii) le ***genre prochain***, lesquels rattachent le défini à la classe sémantique ou conceptuelle à laquelle il appartient par essence. Il s'agit d'un concept superordonné (immédiatement au-dessus

ou plus éloigné) dont le sens est compris dans le sens du défini. Ces incluants comprennent également les éléments qui peuvent être considérés comme génériques, mais qui ne le sont que parce qu'ils 11.1.b) ***« jouent le rôle » d'incluant*** : les *incluants multiples* (ex. : *Reclus* = *enfermé et isolé*), *incluant de définitions d'autonymes*, c'est-à-dire lorsque le générique n'est pas un concept superordonné mais un quasi-synonyme dont le sens est précisé par d'autres éléments définitoires, et l'*incluant dans les définitions en métalangue de signe*, qui renseigne sur l'emploi du signe (*se dit de*, *sert à*, etc.) ou sur sa classe (*mot*, *onomatopée*, etc.).

11.2) Les **faux incluants** se retrouvent dans cinq types de situations : lorsque *la chose est définie par ses parties, par sa cause ou sa conséquence*, lorsqu'il y a *définition de la chose transformée*, lorsque *l'incluant marque le rapport de la chose à l'unité* (*ensemble*, *groupe,…* ; *partie*, *élément*, *membre,… ; tout ce qui*…, *chacun des*…, etc.), et lorsqu'il y a *faux incluant d'existence*, c'est-à-dire lorsque l'incluant exprime l'absence de quelque chose.

Une pratique rigoureuse de la terminologie voudrait qu'on ait autant que possible recours au **genre prochain**, ou pour le moins à un **vrai incluant**, même s'il est plus éloigné dans la hiérarchie (ISO 704, 2000). Or, si les éléments **« jouant le rôle » d'incluant** peuvent dans bon nombre de cas être évités, il n'est souvent pas possible de le faire pour les **faux incluants**. Les définitions en terminologie ont, en effet, souvent un générique dont la relation au concept défini est la *partie* ou le *tout*, voire la *cause* ou la *conséquence*, ou un générique qui rattache le concept à une classe conceptuelle qui n'est pas la véritable classe conceptuelle du défini. Par exemple, lorsque la *définition* est définie comme le *Résultat* d'une action (faux incluant : conséquence) plutôt que comme une *Représentation* (vrai incluant).

4.7. TYPES DE SPECIFIQUES (12)

Le spécifique d'une définition (également nommé *trait distinctif*, *caractère*, *caractère restrictif*, *qualificatif*, ou encore, dans une approche plus lexicologique, *sème*), est la partie de l'énoncé définitoire qui non

seulement renvoie à une particularité, à un aspect du sens, mais aussi distingue et/ou rapproche le sens défini d'autres sens. Chaque spécifique est en principe, comme le générique, *nécessaire* à la description de l'objet défini, mais pas toujours *suffisant* (pris isolément) pour le distinguer d'autres objets.

Le *type des spécifiques* peut parfois servir à caractériser les définitions selon deux axes : ils peuvent être rattachés à un 12.1) **type de caractère** (ISO 704, 2000) (*fonction*, *partie*, *cause*, etc.) en fonction de leur relation au générique ; ou être répartis en 12.2.a) **traits intrinsèques** et 12.2.b) **traits extrinsèques** en fonction de leur relation au « noyau conceptuel », à l'essence du concept. Ces deux perspectives sont compatibles et peuvent être appliquées simultanément à un même spécifique. La classification des spécifiques en traits intrinsèques ou extrinsèques est toutefois peu opérante pour l'étude des définitions en général et des spécifiques en particulier, ou pour la rédaction de définitions.

Parmi ces différentes façons de considérer les spécifiques, la plus répandue en terminologie est celle qui consiste à leur assigner un **type de caractère**, selon leur relation au défini, telle que la *fonction*, la *cause*, la *conséquence*, etc. Cette approche permet d'établir des modèles a priori simples à suivre par le terminologue. Elle soulève néanmoins quelques difficultés, du fait que le type et le nombre de relations n'est pas fini – la liste des relations possibles se termine toujours par un « etc. » très frustrant, et le type de relations proposées est variable –, ce qui peut poser problème lorsqu'on souhaite formaliser le fonctionnement des spécifiques à l'intérieur des définitions.

4.8. PERTINENCE DES CARACTERES (13)

La caractérisation des définitions peut également se faire selon la *pertinence des caractères*. L'ISO 704 (2000) en distingue trois niveaux, selon leur relation à l'objet défini ou au domaine : 13.1) les **caractères essentiels distinctifs**, 13.2) les **caractères essentiels communs** ou partagés (non distinctifs), et 13.3) les **caractères non essentiels**, stéréotypiques ou « superfétatoires ». On peut expliquer ces niveaux de

pertinence comme étant fonction du croisement de deux dimensions (voir ISO 704, 2000, Martin, 1992) : la nature plus ou moins *essentielle* ou *universelle* du trait, et sa *typicité*, c'est-à-dire le fait qu'il soit partagé par un concept associé ou, au contraire, qu'il l'en distingue. Ainsi, un *caractère essentiel* (ou *propriété universelle*) est satisfait par tous les objets dénommés. Il peut, par ailleurs, être *typique* ou *distinctif*, « c'est-à-dire satisfait seulement, à l'intérieur du genre prochain, par les objets en cause » (Martin, 1992), ou *non typique* ou *commun* lorsqu'il est aussi satisfait par d'autres objets de même genre (pour l'oiseau, le fait qu'il est ovipare, puisque d'autre animaux le sont aussi). Un *caractère non essentiel* ou *stéréotypique* correspond à une propriété généralement vérifiée, satisfaite par la plupart des objets dénommés (par ex. le fait qu'un oiseau vole) ; il peut donc être supprimé sans affecter le concept.

En terminologie, la pertinence des caractères suscite le débat. Celui-ci reste ainsi ouvert sur la possibilité de limiter la définition aux seuls **caractères essentiels distinctifs**, c'est-à-dire nécessaires et suffisants pour situer le concept à l'intérieur du système conceptuel considéré, ou de l'opportunité de l'étendre aux **caractères essentiels communs** *(typiques) non distinctifs* et aux *propriétés généralement vérifiées*. Dans les faits, les définitions en terminologie ne devraient donc pas comporter de traits qui ne soient pas pertinents dans le domaine concerné, mais rien n'oblige non plus à ce qu'elles se limitent au strict minimum et excluent le second type de caractères. La non-pertinence des traits *superfétatoires* (*propriétés généralement vérifiées*) devrait quant à elle faire l'unanimité, puisque ces informations à caractère plutôt « encyclopédique » sont recensées dans une *note*. Ce principe semble aller de soi, mais il pose en fait un problème majeur : où s'arrêtent les traits pertinents et où commencent les traits superfétatoires ?

4.9. FONCTIONS (14) de la définition

La définition peut également être caractérisée par sa, ou plutôt, ses *fonctions*. Selon Rahmstorf (1993), ces dernières peuvent être regroupées en trois ensembles : 14.1) les **fonctions orientées objet**, qui mettent l'accent sur le sens défini, 14.2) les **fonctions techniques** liées à la

communication, à l'organisation des connaissances, etc., donc à ses aspects appliqués, et 14.3) les **fonctions métascientifiques** liées à l'étude théorique des définitions.

La principale **fonction orientée objet** est de 14.1.a) ***décrire*** ou d'expliquer un sens. Ce faisant, la définition trace les limites de la compréhension d'un mot (Clas, 1985) ou d'un concept, elle lui assigne une fin (Rey, 1992). En 14.1.b) ***délimitant*** un sens, la définition le 14.1.c) ***distingue*** d'autres sens ou des concepts coordonnés. Ces trois fonctions lui permettent de répondre à la fois aux deux questions suivantes : *"Qu'est-ce que X ?"* et *"En quoi X se distingue-t-il de Y, Z, etc. ?"*. Elles lui confèrent en outre un pouvoir de 14.1.d) ***fixation*** ou de 14.1.e) ***création*** d'un sens (signifié ou concept), et parfois d'une réalité, à un moment donné dans le temps, voire dans l'espace (par exemple, l'espace textuel ou géographique, si l'on pense aux textes de loi).

En fixant un sens ou un usage reflétant la pensée d'une époque ou d'un milieu, la définition réalise un certain nombre de **fonctions techniques** : elle 14.2.a) ***facilite la communication*** et contribue à son efficacité ; elle permet ainsi la 14.2.b) ***transmission du savoir*** du passé et du présent. Si la définition sert souvent à transmettre un usage ou des connaissances à des fins d'apprentissage – fonction 14.2.b.i) ***didactique*** –, elle est aussi utilisée pour imposer un état de la langue ou de la connaissance à une communauté donnée – fonction 14.2.b.ii) ***normalisatrice***. La définition a par ailleurs une fonction 14.2.c) d'***attestation*** ou de 14.2.c) ***vérification***. Elle est garante de l'existence d'une signification ou d'un concept et vice versa (si un sens existe, il doit pouvoir être défini), ainsi que du référent. D'un point de vue linguistique, la définition peut servir à 14.2.e) ***établir la synonymie*** entre différentes unités lexicales de la même langue et 14.2.h) ***l'équivalence*** entre celles de langues différentes.

Dans une perspective **métascientifique**, la définition a donc une fonction de 14.3.a) ***lien*** ou de ***pivot entre unité(s) linguistique(s), concept et référent***. Relativement à un système sémantique ou conceptuel, la définition peut endosser la double fonction de 14.3.b)

structuration ou de 14.3.c) ***miroir du système***, en tant qu'elle opère à la fois des rapprochements et des distinctions entre des sens. En établissant ces relations, dont elle est en même temps le reflet, elle 14.3.a) ***situe*** le défini à l'intérieur d'un système, qu'elle structure simultanément à l'aide de caractères distinctifs (ISO 704, 2000).

En terminologie, la définition peut remplir **la totalité** des fonctions des différents niveaux identifiés par Rahmstorf (1993), même si elle ne les remplit pas forcément toutes à la fois. Non seulement elle ***décrit*** et ***délimite un concept*** de manière à ce qu'il puisse être reconnu, mais elle permet également, grâce aux traits conceptuels utilisés pour ce faire, de le ***distinguer des concepts coordonnés***. Par ailleurs, la définition d'un concept permet de révéler la structuration d'un domaine par l'intermédiaire d'un ou de plusieurs de ses spécifiques et, le plus souvent, par son générique. Elle permet, dès lors, sinon de construire, ***structurer le système*** dans lequel il s'insère, au moins de ***le refléter*** tel qu'il est perçu par les spécialistes, dans une perspective synchronique. Ce faisant, la définition ***indique également la place qu'occupe le concept dans le système auquel il participe***, et uniquement celui-là (Sager, 1990). Du moment qu'il existe une définition d'un concept, celle-ci aura forcément pour effet de le ***fixer*** (lui, ainsi que la forme linguistique du terme qui le désigne), sinon de le ***créer***. En fixant la relation qui unit un concept à son ou ses terme(s), la définition terminographique sert également de ***passerelle entre terme(s) et concept***, et ***donne accès au référent***, ce qui lui confère le pouvoir d'***attester*** l'existence d'un concept, mais aussi celui de le faire connaître, lui, et la réalité qu'il désigne. Elle a donc une ***fonction didactique***, qui va parfois de pair avec une ***fonction normalisatrice***, l'une comme l'autre visant à ***faciliter la communication et la transmission du savoir*** entre spécialistes, ou entre spécialistes et non-spécialistes. Au niveau linguistique, elle est la ***voie d'accès à la synonymie*** terminologique dans une même langue et ***à l'équivalence*** entre termes de langues différentes.

4.10. ROLES (15) de la définition

Quel que soit le fond de la définition, celui-ci peut dans tous les cas être abordé suivant plusieurs approches, qui correspondent en quelque sorte au *rôle* que joue la définition par rapport à son contenu : soit un 15.1) **rôle descriptif**, visant à consigner (a posteriori) l'ensemble des usages ou sens avérés pour un mot, ou décrire les concepts existants dans un domaine ou pour une population donnés ; soit un 15.2) **rôle prescriptif**, qui a pour but d'imposer un sens (même avec une portée limitée à un contrat, une théorie, un auteur, etc.) à travers son contenu informatif ; soit les deux à la fois – 15.3) **rôle mixte** –, ce qui est souvent le cas dans les dictionnaires.

Le rôle des définitions reste un critère fort débattu et souvent contesté en terminologie : les définitions ont-elles un rôle **prescriptif** ou **descriptif**, ou **les deux** à la fois ? Bien que dans la pratique terminographique les définitions soient davantage le résultat d'une démarche **descriptive** (dépouillement de corpus, etc.), il est cependant vrai que les produits terminologiques sent souvent d'une certaine légitimité qui confère de fait aux définitions un caractère **prescriptif**, pas toujours voulu. Les trois rôles sont donc possibles est dépendent fortement du contexte d'emploi de la définition.

4.11. NIVEAU DE SPECIALISATION (16) et DESTINATAIRES *(17)*

Le *niveau de spécialisation* de la langue définitoire, en corrélation très étroite avec le destinataire de la définition, peut aussi constituer un critère distinctif des définitions. Il reste cependant difficile de déterminer une échelle de niveaux de spécialisation, ce jugement ne pouvant être porté que si l'on connaît les définitions à comparer et donc à caractériser l'une par rapport à l'autre et/ou le public visé. En ce sens, le *destinataire* peut aussi être considéré comme un critère définitoire des définitions, mais il pose le même problème d'indétermination. Des typologies peuvent cependant être tentées pour des domaines d'application donnés, comme la lexicographie ou la terminologie.

Afin d'éviter au maximum toute imprécision linguistique, ainsi que par souci d'économie et de concision, la définition en terminologie comporte souvent – dans ses éléments générique et/ou spécifique(s) – des termes de la langue scientifique ou technique connus des spécialistes du domaine et/ou définis ailleurs dans le cadre d'un même projet. De ce fait, le niveau de spécialisation du vocabulaire définitoire en terminologie est souvent plus **spécialisé** (terminologisé) et donc plus « savant » que celui qui est utilisé ailleurs, le niveau étant également fonction du public-cible (des destinataires) des définitions. Rahmstorf (1993) en propose une liste non exhaustive s'articulant autour de huit catégories – à laquelle Blanchon (1997) en ajoute une neuvième – selon l'usage qu'ils font des définitions, leur centre d'intérêt principal (terme, concept, système, etc.) et la fonction terminologique remplie par la définition. Il est également possible de proposer une classification plus simpliste en fonction du bagage cognitif du destinataire : 17.1) **spécialistes du domaine**, 17.2) **spécialistes de domaines connexes** ou 17.3) **non-spécialistes**.

5. Conclusion

Dans cet article, nous avons cherché à caractériser les définitions employées dans les dictionnaires et les bases de données terminologiques sur la base d'une grille de lecture des définitions en général. Après quelques remarques sur les typologies de définitions, dont l'étude nous a conduit à souligner le caractère inopérant pour qui souhaite comprendre les définitions en terminologie, nous avons passé en revue un ensemble de 17 critères permettant de caractériser les définitions en général. Ceux-ci nous ont ensuite servi de base pour caractériser les définitions en terminologie.

Cet état de l'art sous forme d'inventaire montre que les points de vue sur la définition sont multiples et que chaque caractère peut donner lieu à un type de définition particulier. Chaque caractère exposé peut également servir d'axe d'opposition à différentes typologies de définitions. Certaines questions restent toutefois ouvertes, notamment

autour de la nature des traits spécifiques de la définition en compréhension, la plus fréquente en terminologie.

Si le mode logique et la structure générale – générique + spécifique(s) – qui en découle ne posent pas de problème particulier, il reste toujours des lacunes autour des types et de la pertinence des spécifiques, par exemple en termes de généralisation à tous les référents d'un concept (*conditions nécessaires et suffisantes* ou *stéréotypie*) ou en termes de relations conceptuelles (problème d'indétermination des types de caractères : combien de relations et lesquelles ?). Ainsi, la *fonction* semble être un type de relation privilégié en terminologie, mais pas toujours. Dans quels cas ne l'est-elle pas ? Pourquoi ? Tous les concepts sont-ils définis avec les mêmes types de spécifiques ? N'existe-t-il pas des types de concepts pouvant être justement distingués par la nature de leurs traits ? Par ailleurs, où placer la limite entre traits pertinents et superfétatoires ? Sur quels critères ? Autant d'interrogations qui rejoignent d'autres questions relatives à la relation entre définition et concept, et qui ouvrent un champ de recherche considérable pour la théorie de la définition en terminologie.

Bibliographie

É. Blanchon « *Point de vue sur la définition* » Meta 42:1, 168-173, 1997

A. Clas « *Guide de recherche en lexicographie et terminologie* » Paris, Agence de coopération culturelle et technique, 1985

B. de Bessé « *Chapitre 2.3.: La définition* » Notes de cours, 68-87, 1996

L. Depecker « *Le signe entre signifié et concept* » Le sens en terminologie, 86-126, 2000

ISO 704 « *Travail terminologique -- Principes et méthodes* » Genève, ISO, 2000

R. Martin « *Pour une logique du sens* » Paris, Presses Universitaires de France, 1992

Nouveau Petit Robert, Dictionnaires Le Robert, 2006

G. Rahmstorf « *Role and Representation of Terminological Definitions* » Actes de TKE'93, 39-49, 1993

J. Rey-Debove « *Étude linguistique et sémiotique des dictionnaires français contemporains* » The Hague, Paris, Mouton, 1971

A. Rey « *L'impossible définition* » Le lexique images et modèles: du dictionnaire à la lexicologie, 98-113, 1977

A. Rey « *La terminologie : noms et notions* » Paris, "Que sais-je ?" n° 1780, Presses Universitaires de France, 1992

J. Sager « *A practical course in terminology processing* » Amsterdam, Philadelphia, John Benjamins, 1990

S. Seppälä « *Composition et formalisation conceptuelles de la définition terminographique* » Université de Genève, École de traduction et d'interprétation, 2004

U. Weinreich « *La définition lexicographique dans la sémantique descriptive* » Langages 19, 69-86, 1970 [1962]

Un système logique pour les relations sémantiques entre concepts.

Christophe Jouis
Université Paris Sorbonne Nouvelle - Sorbonne Nouvelle
17, rue de la Sorbonne 75230 - Paris Cedex 05
cjouis@univ-paris3.fr
&
LIP6 (Laboratoire d'Informatique de Paris 6 - Université Pierre-et-Marie-Curie)
104, avenue du Président Kennedy- 75016 Paris
Christophe.Jouis@lip6.fr
http://www-poleia.lip6.fr/ACASA/

Résumé :

A main goal of recent studies in semantics is to integrate into conceptual structures the models of representation used in linguistics, logic, and/or artificial intelligence. A fundamental problem resides in the need to structure knowledge and then to check the validity of constructed representations. We propose associating logical properties with relationships by introducing the relationships into a typed and functional system of specifications. This makes it possible to compare conceptual representations against the relationships established between the concepts. The semantic system proposed is based on a structured set of semantic primitives – types, relations and properties- based on a global model of language processing, Applicative and Cognitive Grammar (ACG) (Desclés, 1990), and an extension model to terminology and ontology (Jouis, 1995, 1996, 1997, Jouis, 2002, 2004, 2006). The ACG postulates three level of representation of languages, including a cognitive level. At this level, the meanings of lexical predicates are represented by semantic cognitive schemes. From this perspective, we propose a set of semantic concepts, which defines an organized system of meanings. Relations are part of a specification network based on a general ontological scheme (i.e., a coherent system of meanings of relations). In such a system, a specific relation may be characterised as to its: (1) functional type (the semantic type of arguments of the relation); (2) algebraic

properties (reflexivity, symmetry, transitivity, etc.); and (3) combinatorial relations with other entities in the same context (for instance, the part of the text where a concept is defined)..

1. Introduction

La sémantique des relations sémantiques entre concepts (c'est-à-dire, pour chaque relation, le nombre et le type de ses arguments, ses propriétés algébriques, etc.) sont souvent trop vagues (par exemple dans les thesaurus, les structures conceptuelles, les ontologies ou les réseaux sémantiques). La sémantique des relations est vague parce que les principaux utilisateurs de ces relations sons des acteurs industriels. Toutefois, la consistance des ontologies construites doit toujours être garantie.

Par exemple, en terminologie, les relations sémantiques entre concepts sont souvent réduites à la distinction établie par les standards ISO 704 (1987) et ISO 1087 (1990) entre les relations sémantiques hiérarchiques (relations genre-espèces et relations partie/tout) et les relations non hiérarchiques (« temps, espace, relations de causalité », etc.). Une approche possible à ce problème consiste à organiser les relations sémantiques dans une typologie fondée sur des propriétés logiques. Par exemple, (Winston, Chaffin & Herrmann, 1987) ou (Pribbenow, 2002) distinguent plusieurs types de relations partie/tout. Cette typologie a inspiré le traitement des relations partie/tout dans WordNet (Miller, 1990). Des travaux récents appliquant les relations terminologiques à la recherché d'informations (information retrieval), en particulier pour la construction de thesaurus et d'ontologies, tentent de mieux spécifier les propriétés des liens entre concepts et de les étendre aux relations non hiérarchiques (Molholt, 1996); (Green, 1996, 1998), (Bean, 1996). D'autres travaux récents ont pour objectifs d'intégrer dans leur modèle terminologique des théories issues de la linguistique (la sémantique, par exemple) et de l'intelligence artificielle, en particulier la modélisation des connaissances pour la conception de systèmes à base de

connaissances et les ontologies, comme cela est défini, par exemple par (Sowa, 1984, 2000) ou (Hovy, 2002). Dans toutes ces disciplines, la nécessité de structurer la connaissance et ensuite de valider les représentations obtenues est fondamentales. En intelligence artificielle, des méthodes pour l'acquisition et la modélisation des connaissances, telles que KADSII, présentés par exemple dans (Wielingua, Schreiber & Breuker, 1992), ont été développées pour l'aide à la conception de systèmes à base de connaissances. Ces méthodes proposent de modéliser un domaine d'expertise sous la forme de concepts connectés par des relations sémantiques dans des langages orientés objets (appelé « niveau domaine » dans KADS). De notre point de vue, ces langages semblent très proches des structures des bases de données terminologiques. En terminologie, des logiciels ont été développés pour « naviguer » dans les réseaux de concepts structurant des micro-domaines, par exemple le système Termisti (Van Campenhoudt 1994, 2007; Lejeune & Van Campenhoudt, 1998), le système Code et le système Cogniterm (Meyer and Mchaffie, 1994), et le système Ikarus (Meyer and Skuce, 1998), qui incluent une gestion automatisée de base de connaissances terminologiques. Dans le but de mieux concevoir la structure des connaissances des concepts d'un domaine, et plus particulièrement, l'indexation et/ou la recherche d'information, nous proposons un ensemble structuré de relations sémantiques, fondé sur un modèle linguistique, la Grammaire Applicative et Cognitive (GAC) de (Descles, 1990). Ce modèle a été appliqué et étendu pour l'acquisition et la modélisation des connaissances par (Jouis, 1993, 1995), puis implanté sur ordinateur dans le système SEEK par (Jouis, 1995, 1998). (Mustafa and Jouis, 1996, 1997) et (Jouis 1998, 2004, 2006) ont reconsidéré ce modèle pour la construction de terminologies et d'ontologies. Les relations sémantiques de ce modèle font partie d'un réseau de relations. Nous proposons un schéma générique de relations (REL), qui est ensuite spécifié en fonction de propriétés algébriques dans des relations plus précises en fonction des propriétés qui leurs sont attribuées. Notre typologie est fondée principalement sur la distinction entre situation statique (« état de choses ») et situation dynamique (modification et changement dans le domaine). Il est à noter que notre typologie diffère de celle définie par (Felber, 1987), qui a établi une distinction entre «

relations logiques », « relations ontologiques » et « relations de cause à effet ». Avec notre approche, il est possible de vérifier la consistance des structures conceptuelles construites. Dans la suite, nous présenterons l'architecture sémantique et les quatre catégories de primitives de notre extension de la GAC.

2. Système sémantique et logique proposé

La GAC est une extension de la Grammaire Applicative Universelle (Shaumyan, 1987). Elle postule trois niveaux de représentation des langages :

- Le niveau phénotype décrit les caractéristiques superficielles des langages telles que l'ordre des mots, les cas morphologiques, etc. Chaque langage est appréhendé dans la diversité de ses expressions linguistiques, qui sont directement observables. Les expressions linguistiques de ce niveau sont vues comme des unités linguistiques concaténées.

- Le niveau génotype exprime des invariants grammaticaux et des structures qui sont à la base des phrases du niveau phénotypes. Le niveau génotype est structuré comme un langage formel appelé langage génotype. Il est décrit par une grammaire appelé Grammaire Applicative. (Biskri & Desclés, 1997). Les descriptions sont représentées sous la forme d'expressions applicatives formulées à l'aide d'opérateurs et d'opérandes de différents types.

- Dans le niveau cognitif, la signification des prédicats lexicaux sont représentés par des schèmes sémantiques et cognitifs. Ce niveau constitue la représentation des connaissances associé à un texte. Les représentations des niveaux génotype et cognitive sont des expressions de la logique combinatoire typée. (Curry & Feys, 1958). Avec le niveau cognitif, la GAC propose un ensemble de primitives sémantiques, qui définissent un système organisé de significations.

A l'intérieur du niveau cognitif, nous distinguons quatre catégories de primitives :

- Des types élémentaires des entités ;

- Des opérateurs de formation, qui permettent de créer des types plus complexes à partir des types élémentaires (listes, tableaux, types fonctionnels[21], etc.);

Des relations statiques fondamentales entre entités, où les relations statiques permettent la description d'états du domaine (situation statique) et où les situations statiques restent stables Durant un certain intervalle temporal où ni le début ni la fin ne sont envisagés (nous avons identifié plus de vingt relations statiques ; et

Des relations dynamiques fondamentales où les relations dynamiques permettent la description de processus ou d'événements dans un domaine : mouvement, changement d'état, conservation d'un mouvement, itérations, intensité, variations, contraintes, causes, etc.

Les relations sémantiques sont ainsi classifiées dans deux catégories principales disjointes : relations statiques et relations dynamiques. Dans cet article, nous allons décrire plus particulièrement les relations statiques, parce que ce sont celles que nous avons complètement formalisé dans notre système dans le but d'effectuer des tests de cohérence de modèles conceptuels. Ensuite, nous donnerons un bref aperçu des relations dynamiques.

3. Types sémantiques élémentaires

Nous distinguons un certain nombre de types élémentaires pour les entités. Par exemple :

[21] Au sens du lambda-calcul typé de Church ou de la logique combinatoire typée de Curry (Curry & Feys, 1958).

- Les Entités booléennes (notées H) sont des objets dont la valeur est soit vraie soit fausse.

- Les Entités individualisables sont des entités qui peuvent être désignées ou montrées par pointage. Elles peuvent être comptées individuellement ou regroupées dans des classes distributives. Par exemple, des entités telles que *Jean, chaise, fourniture, homme, enfant* sont distinguables. Les entités individualisables sont notes J. Par exemple : [J: table].

- Les entités massives telles *que eau, mer, vin, beurre, pain* ne sont pas des entités individualisables. Toutefois, nous pouvons noter qu'un certain nombre d'opérateurs (les classifieurs) permettent d'individualiser des notions massives : *un verre d'eau, un bras de mer, une bouteille de vin, une tranche de pain, un morceau de beurre*. Elles sont notées M. Par exemple [M: mer].

- Les classes distributives regroupent des entités individuelles qui ont une propriété identique. Elles sont notes D. Par exemple, [D: être-un-carré] représente une classe distributive d'entités individuelles ou un « concept ».

- Les classes collectives se distinguent des entités individualisables dans le sens qu'elles représentent des objets qui forment un « tout » à partir d'objets plus élémentaires. Elles sont notées C. Ainsi, [C: entité géographique], [C: armée], [C: molécule [22]], [C: corps humain] représentent des classes collectives.

[22] Une molécule est formée de différents types d'atomes, qui eux-mêmes sont formés de …

- Les lieux, vus comme un type sémantique (noté P), sont conceptualisés comme un ensemble de positions, chaque position étant assimilée à un point. Pour chaque entité (individualisable, collective ou massive), nous pouvons associer un ensemble de lieux. Par exemple, [P: Paris], [P: jardin], [P: maison] peuvent être vus d'une certaine façon comme une entité individualisable (*Paris est une ville)* et, d'une autre façon, chaque entité individualisable détermine un lieu spécifique (*Je suis dans Paris*).

4. Distinction entre classes distributives et classes collectives

Il est important de faire la différence entre les classes distributives et les classes collectives. Dans une classe collective, le « tout » est vu comme une « accumulation » des éléments que la constituent, disjoints ou non. Lesniewsky (1886-1939) a proposé une théorie générale des touts et des parties (méréologie), en réponse au problème de la théorie des ensembles (Cantor, 1982, 1962). Une analyse détaillée de la méréologie a été menée par (Miéville, 1984). Lesniewsky arrive à la conclusion que la notion de classe contient deux aspects : le distributive et le collectif. L'exemple qui suit, emprunté à (Grize, 1973) donne une idée de la différence : *« Une classe distributive est, pour être strictement correct, l'extension d'un concept. Si p est le concept planète, l'énoncé Jupiter est une planète correspond à poser soit p(Jupiter) ou Jupiter* $\in \{x / p(x)\}$*, et l'information transmise est la même dans les deux écritures. Ainsi,* $p = \{$*Mercure, Vénus, Terre, Mars, Jupiter, Saturne, Neptune, Uranus, Pluton*$\}$ *est une classe distributive. Elle contient neuf éléments et rien d'autre.*

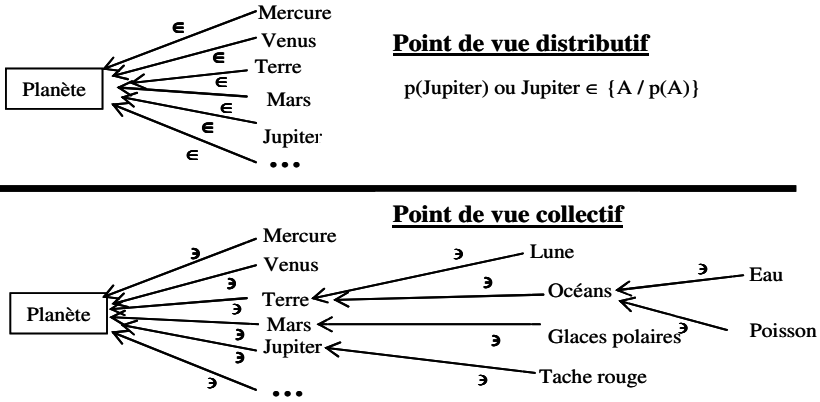


Fig.1 : Classes Distributives vs. Collectives : points de vue différents mais logiques

Les glaces polaires de Mars, la tache rouge de Jupiter, les anneaux de Saturne n'appartiennent pas à p. Tout ceci, et un millier d'autres choses ont un rapport avec le concept planète. La notion de classe collective doit pallier ce défaut. » (voir Fig. 1).

4.1. Types complexes

A partir de l'ensemble des types élémentaires S = {H, J, M, D, C, P …}, il est alors possible de définir un système de types plus complexes de façon récursive en partant des deux règles suivantes :

- Les éléments de S sont des types élémentaires.

Si x et y sont des types, alors Fxy est un type (fonctionnel).

Le symbole F est un formateur d'opérateur de types fonctionnels. Un opérateur E de type Fxy (noté [Fxy: E) est un opérateur unaire qui prend comme argument un objet de type x pour retourner un résultat de type y. Si nous considérons une entité A de type x, l'application de E sur A va construire une certaine entité B de type y :

([Fxy: E] [x: A]) → [y: B]

Par exemple, le type FJH est le type d'un opérateur qui, appliqué à une entité individualizable (J), retourne une valeur de vérité H (propriété unaire des individus, ensemble d'individus ou « concept », comme par exemple [FJH : « être-un-carré] ».

Une relation entre une entité individuelle et un lieu (localisation) aura le type FJFPH. Parce que la localisation est un opérateur binaire, l'application est effectuée en deux étapes. Par exemple, la localisation de Jean dans Paris est formalisée de la façon suivante. Nous avons les types suivants : [J: Jean], [P: Paris] et [FJFPH: localisation]. La localisation s'applique tout d'abord à Jean pour retourner un opérateur de type FP :

(FJFPH] localisation] [J: Jean]) → [FPH: localisation_Jean]

Le résultat est un opérateur de type FPH qui s'applique ensuite au lieu Paris pour retourner une valeur de vérité v de type H :

([FPH: localisation_Jean]] [P: Paris]) → [H : True].

Le connecteur logique « ET » est un opérateur binaire qui s'applique à deux entités booléennes :

[FHFHH: ET], etc.

4.2. Relations statiques entre entités

La relation statique générale notée REL est un modèle schématique : Une entité X est en relation avec une entité Y. Ce modèle schématique est ensuite spécifié en fonction de propriétés algébriques pour former des relations plus précises : identification, différentiation et ruption (voir Fig. 2).

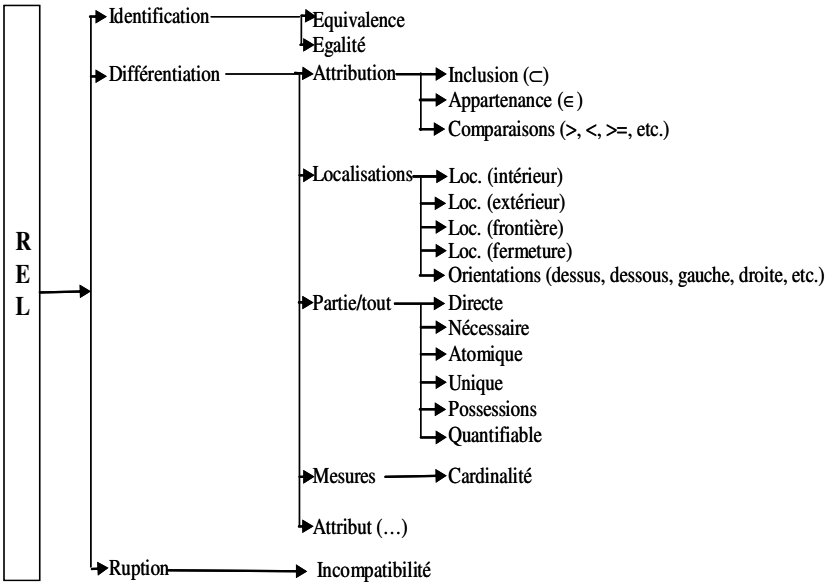


Fig. 2: Réseau de spécification des relations statiques

- La relation d'identification (qui peut être paraphrase en « X est identifiée à Y », c'est-à-dire l'entité Y est utilisé comme un identifiant de X) est une relation binaire, symétrique et réflexive. Elle est exprimée dans des énoncés tels que : *Paris est la capitale de la France* ou « rendez-vous » a la même signification en anglais que « appointement ». L'identification se spécifie en plusieurs relations telles que l'égalité extensionnelle et l'égalité intentionnelle. .

- La relation de différentiation (qui signifie « X est différent de Y » est non-symétrique.

- La relation de ruption (qui signifie qu'il n'y a pas de propriété commune entre X et Y) est une relation non-réflexive et symétrique. Elle s'exprime dans des énoncés tels que Protons et électrons sont

incompatibles. Très souvent, cette relation s'applique à des classes disjointes issues de la même classe par attribution : valeurs positives/négatives, propriétés quantitatives/qualitatives, etc. Ainsi, les relations sémantiques constituent un système de significations des relations entre entités (Culioli & Descles, 1982, Descles, 1987, Jouis 1993). Dans ce système, une relation peut être spécifiée en des relations plus précises en fonction de ses propriétés: (1) son type fonctionnel (les types sémantiques de ses arguments), (2) ses propriétés algébriques[23] (réflexivité, symétrie, transitivité, etc.), et (3) ses propriétés de combinaison avec d'autres relations et d'autres entités dans un même contexte (la partie du texte où un concept est défini par exemple). Par exemple, l'inclusion entre classes distributives est irréflexive, asymétrique et transitive. De plus, dans un même contexte, elle est incompatible avec d'autres relations, comme par exemple l'appartenance d'une entité individuelle à une classe distributive. Les relations statiques sont structurées et indépendantes d'un domaine particulier. Ce sont des relations binaires. Nous distinguons plus de vingt relations. Nous présentons dans la suite les propriétés des relations de différentiation. Parmi les relations issues de la différentiation, nous avons les attributions, qui sont caractérisées par l'asymétrie. Cette asymétrie se spécifie en :

[23] Rappelons en particulier les propriétés algébriques des relations binaires :

Etant donné trois entités X, Y et Z et une relation R :

A) Réflexivité:

R (complètement) réflexive = def $\forall$ □X (X R X), R non-réflexive = def $\exists$ □X ¬ (X R X),

R irréflexive = def $\forall$ □X ¬ (X R X).

B) Transitivité:

R transitive = def $\forall$□ X, Y, Z (R(X, Y) and R(Y, Z)) $\Rightarrow$ R(X, Z),

R non-transitive = def □$\exists$ X, Y, Z (R(X, Y) and R(Y, Z)) and NOT (R(X, Z)),

R JAMAIS transitive = def $\forall$ □X, Y, Z (R(X, Y) and R(Y, Z)) $\Rightarrow$ NOT (R(X, Z)).

- L'appartenance d'une entité individualizable (J) à une classe distributive (D). De type FJFDH, cette relation est non-réflexive, asymétrique et non-transitive. Elle est exprimée dans des énoncés tels *que PI est un nombre réel*[24] .

- L'inclusion entre classe distributives (*Les bactéries sont des micro-organismes* [25]), qui est de type FDFDH, et qui est non-réflexive, asymétrique et transitive .

- La comparaison, qui correspond à une relation d'ordre stricte (c'est-à-dire qu'elle est ni réflexive ni symétrique, mais est transitive), concernant les entités individuelles : son type est donc FJFJH. Elle se spécifie dans plusieurs relations : supérieur (>), et inférieur (<), etc.

Les relations de localisation spatiales sont exprimées dans les exemples suivants : *Paris est en France, Le jardin entoure la maison, le livre est sur la table,* etc. Les relations de localisation sont de type FxFPH où x est de type J ou de type P, en fonction du contexte de l'entité localisée. Chaque occurrence d'un objet, dans un environnement particulier, détermine un lieu (c'est-à-dire un « voisinage », dans la terminologie de la topologie). Des primitives de position peuvent être définies en faisant appel aux concepts élémentaires de la topologie générale. Un lieu est alors visualisé dans sont intérieur, son extérieur (excluant sont intérieur et sa frontière), sa frontière (excluant son intérieur et son extérieur) ou sa fermeture (son intérieur et sa frontière). Nous introduisons les

[24] Considérons, par ailleurs, l'attribution d'une propriété à une entité individuelle. Par exemple, l'énoncé Socrate est un humain signifie que l'entité individuelle Socrate appartient à la classe distributive des humains ou que le concept « être-un-humain » s'applique à Socrate.

[25] Considérons, par exemple, l'énoncé Les hommes sont mortels. Il est à noter que dans de nombreux thesaurus ou modèles de réseaux sémantiques, on utilise typiquement seulement la relation générale « est-un » sans distinguer l'appartenance de l'inclusion. Or, il il y a une différence fondamentale, puisque la première est JAMAIS transitive tandis que la seconde est transitive et permet l'héritage de propriétés.

opérateurs de détermination topologique d'un lieu x : in(x), ex(x), fr(x) and fe(x), déterminant l'intérieur, l'extérieur, la frontière et la fermeture de x, respectivement. Pour tout lieu x, nous avons, par exemple :

$in(x) \subset x \subset fe(x)$

$fr(x) \subset fe(x)$ (parce que $fe(x) = x \cup fr(x)$)

$x \cap ex(x) = \varnothing$

$fr(x) = co(in(x)) \cap co(ex(x))$

Les propriétés de ces quatre opérateurs nous permettent d'établir les propriétés des relations de localisation[26] . Les relations de localisations totplogiques sont alors les suivantes :

- Loc-in (« être-à-l'intérieur-de"): localisation à l'intérieur d'un lieu (L'oscillateur est placé dans la première zone). Cette relation est transitive, asymétrique et non-réflexive.

- Loc-ex (« être à l'extérieur de »): localisation à l'extérieur d'un lieu (Le Limiteur est extérieur à la troisième zone). Cette relation est transitive et JAMAIS réflexive.

- Loc-fr (« être-à-la-frontière-de »): localisation à la frontière d'un lieu (Alger est au bord de la mer). Cette relation est incompatible avec l'intériorité et l'extériorité ; elle est plus précise que la localisation à la fermeture d'un lieu.

- Loc-fe (« être-à-la-fermeture-de ») : localisation à la fermeture d'un lieu (Boulogne est située dans la banlieu de Paris). Cette relation est

[26] Kuratowski (1958) a montré qu'il y a exactement 14 opérateurs distincts en combinant les quatre opérateurs identité, in, fe, ex et le complémentaire co. A partir de ce résultats (Barbut, 1965) montra ensuite qu'il était possible de déduire entièrement les propriétés de combinaison des relations de localisation. Pour plus de détails, voir (Jouis, 1993).

incompatible avec l'intériorité ; elle est redondante avec la localisation à la frontière et à l'intérieur.

Nous pouvons distinguer des localisations orientées de la même façon en introduisant les primitives gauche(x), droite(x), devant(x), derrière(x), au-dessus(x) et en-dessous(x). Toutefois, ces primitives peuvent être définies seulement si l'objet de référence a une orientation intrinsèque : le devant d'une maison, l'avant d'un bateau, etc.

La relation partie/tout est un relation générale permettant la décomposition d'un objet en ses composants. En utilisant cette relation, chaque entité individuelle est vue comme une unité complexe organisée. La relation partie/tout admet deux arguments qui sont, respectivement l'objet tout et l'objet composant. Son type est alors FCFxH, où x est de type J ou de type C. Dans les relations partie/tout, nous distinguons les relations de composition (La fluorine entre dans la composition des os et des dents) et la possession (Jean a eu une voiture).

La relation de composition (notée □) est réflexive et asymétrique mais (généralement) non transitive, ce qui la différencie de l'inclusion. Elle est exprimée dans des énoncé tels que : La main forme une partie du bras. La composition est spécifiée dans plusieurs relations. En effet, il existe un grand nombre de propriétés décrivant les relations entre l'objet composant et l'objet tout. Par exemple :

- Composition nécessaire versus composition non nécessaire *(Le processeur est un des composant essentiel de l'ordinateur* versus *Un lecteur CD-ROM est un composant accessoire d'un ordinateur*). Les caractéristiques nécessaire et non nécessaire sont transitives pour la relation de composition.

- Composition directe versus composition non directe (*L'opium est un composant primaire de la Lamaline versus Une molécule est constituée de neutrons, protons et électrons qui sont des parties des atomes*). Un objet partie OP est un composant direct d'un objet tout OW, s'il n'y a pas d'objet OP1 (différent de OP) de telle sorte que l'objet partie OP soit un composant

de l'objet OP1 et l'objet OP1 soit un composant de l'objet OW. Sinon, OP est un composant non direct de l'objet OW. (voir Fig. 3). La composition non directe est transitive, tandis que le composition directe n'est pas transitive.

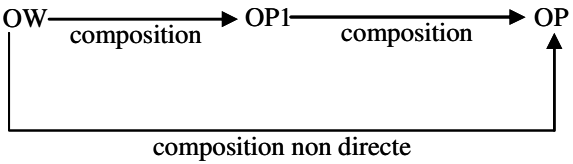


Fig. 3 : Composition directe vs. non directe

- Composition atomique versus composition non atomique (Le plus petit composant d'un programme est le bit versus Un livre se décompose en chapitre, qui eux-mêmes se décomposent en paragraphes). La composition atomique n'admet pas la transitivité, mais la composition non atomique l'autorise.

- Composition unique versus composition non unique (Une étoile jeune est composée uniquement d'atomes d'hydrogène versus L'atmosphère est un mélange de plusieurs gaz, dont les principaux sont l'oxygène et l'azote).

- Composition quantifiable versus composition non quantifiable (La main est compose de cinq doigts; Chaque cellule humaine contient 46 chromosomes versus L'eau est constituée d'atomes d'oxygène et d'atomes d'hydrogène).

La relation attribut permet d'ajouter des relations statiques spécifiques, qui sont en dehors du schéma général de relation, c'est-à-dire qui sont spécifiques à un domaine particulier. Par exemple, la relation « être-le-père-de » que l'on pourrait utiliser en généalogie ne peut pas être considérée comme une relation générale. En d'autres termes, l'ensemble structuré de relations que nous proposons doit être vu

comme un ensemble d'invariants sémantiques, indépendants d'un domaine de connaissances particulier, mais non exhaustif. Notre approche n'exclue pas la nécessité d'ajouter des relations spécifiques à un domaine.

4.3. Vers des relations dynamiques

Contrairement aux relations statiques, les relations dynamiques construisent (ou décrivent) des situations non statiques. Elles introduisent des modifications entre les objets du domaine. La modification est un processus qui fait passer d'une situation statique SIT1 vers une autre situation statique SIT2. Trois zones temporelles peuvent alors être distinguées : (1) avant la modification (SIT1), (2) pendant la modification et (3) après la modification. Si nous introduisons les nouveaux types élémentaires St et Dy pour désigner les situations statiques et les situations dynamiques, le schéma général d'une transition est alors :

[St: SIT1] → [FStFStDy]: DYNA] → [St: SIT2]

La relation dynamique générale DYNA décrit le type de transition. DYNA est ensuite spécifiée, par exemple, comme une relation de mouvement (MOVT), comme une relation de changement d'état (CHANG), comme une relation de conservation d'un mouvement (CONSV) comme une relation de causalité (CAUS), etc.

5. Vers un système de vérification de la cohérence des structures conceptuelles ?

Un des points sensible est la validation des structures conceptuelles durant leurs constructions. Cette validation ne peut être réalisée qu'avec la coopération de spécialistes (c'est-à-dire les spécialistes du domaine, les terminologues et/ou les ingénieurs de la connaissance).Avec des relations définies par des propriétés logico-sémantiques, il est possible de contrôler la consistance des représentations en vérifiant que toutes les propriétés des relations sont

bien appliquées (respectées). La consistance est une condition nécessaire à la validation de structures conceptuelles. Un module informatique, testant ces propriétés, est en cours de développement. Il intègre des procédures de contrôle pour chacune des propriétés. Ce module utilise GRAPHLET (Himsolt, 1994), un système informatique de gestion et d'affichage des graphes. Les procédures sont lancées au moment où l'utilisateur introduit une nouvelle relation entre deux concepts. Par exemple, la relation d'inclusion entre deux classes distributives est non réflexive, non symétrique mais est transitive. La Figure 4 illustre une des conditions nécessaires pour la maintenance de la cohérence.

Fig. 4: L'inclusion est irréflexive, asymétrique mais transitive.

Quand un utilisateur tente d'établir une relation d'inclusion entre une classe distributive D1 et une autre classe distributive Dk, il est nécessaire de vérifier les propriétés de non réflexivité et d'asymétrie, en effectuant la fermeture transitive (itinéraire de tous liens de Dk, pour vérifier qu'on n'arrive pas à D1). Notons que ce type de vérification peut s'avérer long et fatiguant pour l'ingénieur de la connaissance, qui peut être amené à gérer un grand nombre de concepts et de relations, tout particulièrement si cette tâche doit être menée manuellement.

Quand une relation partie/tout est établie entre deux classes collectives C1 et C2, alors, dans un même contexte, il y a incompatibilité avec la relation d'inclusion comme cela est représentée dans la figure 5.

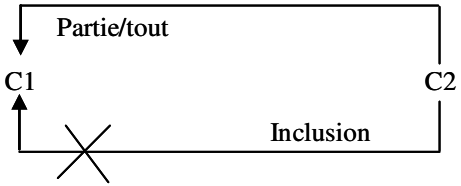


Fig. 5: L'inclusion et la relation partie/tout sont incompatibles dans le même contexte.

Notons qu'il est nécessaire de prendre en compte la transitivité pour ces deux relations ce qui mène à effectuer la fermeture transitive de ces deux relations autour de C1 et de C2.

Les tests de cohérence peuvent mettre en jeu plusieurs propriétés simultanément. Considérons par exemple la situation typique donnée dans la figure 6. Dans cet exemple, nous avons deux classes distributives en relation de disjonction (elles sont disjointes ; la disjonction étant une spécification de la ruption). En supposant que les hiérarchies d'inclusion issues respectivement de D1 et de D2 sont consistantes, alors, pour que le réseau reste cohérent, il n'est plus possible d'avoir :

- L'introduction de relations d'inclusion entre des sous-classes distributives issues de la hiérarchie de D1 et des sous-classes issues de la hiérarchie de D2 (et vice-versa) ; ou

- L'introduction de relations d'appartenance entre des entités individuelles de la hiérarchie de D1 vers des classes distributives issues de la hiérarchie de D2 (et vice-versa).

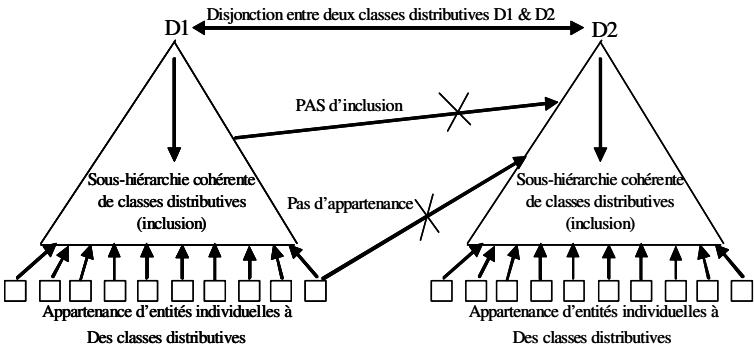


Fig. 6: Test de plusieurs propriétés de plusieurs relations

Ces trois exemples montrent que l'implémentation d'un système informatique qui vérifie la consistance de structures conceptuelles construites suivant notre modèle consiste en de simples procédures. Ces procédures sont simples, mais doivent parfois être combinées entre elles. Ces opérations, quand elles doivent être menées manuellement, exigent une grande rigueur et un long et fatiguant travail, qui notre système informatique propose de réaliser automatiquement.

6. Conclusions

La typologie des relations que nous proposons est fondée sur un ensemble organisé de primitives. Ces primitives sont organisées à partir de types sémantico-logiques dans un système logique de significations. Ce système est construit progressivement, en partant d'un schéma général de relations entre entités, pour obtenir peu à peu des relations sémantiques précises, par spécification progressives à l'aide de propriétés. Dans ce système, la sémantique de chaque relation correspond à des propriétés intrinsèques. En introduisant des relations sémantiques dans ce système, il est alors possible de développer un système automatique qui vérifie la consistance interne des structures conceptuelles (construites ou ajoutées en cours de construction), en

fonction des propriétés des relations. La consistance est une condition nécessaire (mais non suffisante) pour leurs validations. L'ensemble structuré des relations proposées est composé d'invariants sémantiques, indépendants d'un domaine de connaissance particulier ou d'un langage donné. Toutefois, Nous ne prétendons pas qu'il soit complet et exhaustif. Les relations devraient être validées par des expérimentations psycholinguistiques. De plus, notre approche n'exclue pas le besoin d'ajouter des relations spécifiques à un domaine. Par ailleurs, nous n'avons pas traité la représentation des entités atypiques. Enfin, nous avons décrit que les aspects statiques, auxquels doivent être ajoutés les relations dynamiques entre entités. Ces relations permettent de décrire des processus ou des événements.

Remerciements

Cette recherché est financée par le LIP6 (Laboratoire de PARIS 6 – Université Pierre et Marie Curie, France). Tout particulièrement, l'auteur remercie le professeur Jean-Gabriel Ganascia, responsable de l'équipe ACASA ("Acquisition des Connaissances et Apprentissage Symbolique Automatique"). Il a aidé l'auteur à mieux formaliser le modèle présenté dans cet article.

Bibliographie

Abraham, M. *Analyse Sémantico-Cognitive des Verbes de Mouvement et d'Activité : Contribution Méthodologique à la Constitution d'un Dictionnaire Informatique des Verbes*, Thèse de doctorat, Paris : Ecole des Hautes Etudes en Sciences Sociales, 1995

Barbut, M.. « *Topologie générale et algèbre de Kuratowski* », Mathématiques et Sciences Humaines, 12, 11-27, 1965

Bean, C. *« Analysis of non-hierarchical associative relationships among Medical Subject Headings (MeSH)",* Anatomical Terminology, Knowledge

Organization and Change: Proceedings of the Fourth International ISKO Conference, 80-86., 1996

Biskri, I. & Desclés, J.-P. *Applicative and Combinatorial Grammar: From Syntax to Functional Semantics*, Amsterdam and Philadelphia: John Benjamins, 1997

Cantor, G. *Gesammelte Abhandlungen* (E. Zemelo, Ed.) Hildesheim. Omls (original work published 1932), 1962

Culioli, A., & Descles, J.-P. *« Traitement formel des langues naturelles »*, Mathématiques et Sciences Humaines, 77, 3-125 ; 78 5-31, 1982

Curry, H., & Feys, R. *Combinatory Logic* (Volume I). Amsterdam: North-Holland, 1958

Desclés, J.-P. *Langages Applicatifs, Langues Naturelles et Cognition*, Paris : Hermes, 1990

Descles, J.-P. « *Réseaux sémantiques: La nature logique et linguistique des relateurs* », Langages, 87, 55-78, 1987

Felbert, H.. *Manuel de Terminologie*, Paris : UNESCO, 1987

Green, R. « *Attribution and relationality* » Structures and Relations in Knowledge Organization: Proceedings of the Fifth International ISKO Conference, 328-315, 1998

Green, R. *"Development of a relational thesaurus*", Proceedings of the Fourth International ISKO Conference, 72-79, 1996

Grize, J.-B. *Logique moderne* (Fascicule II), Paris : Mouton/Gauthier-Villars, 1973

Himsolt, M. "*GraphEd: A graphical platform for the implementation of graph algorithms*", Tamassia R. & Tollis, I.G. (Eds). Graph Drawing, Lecture Notes in Computer Science 894, n182-193,1994

Hovy, E "*Comparing Sets of Semantic Relations in Ontologies*", The Semantics of Relationships, 91- 110. Dordrecht: Kluwer Academic Publishers, 2002

International Organization for Standardization (ISO). *Terminology – Vocabulary = Terminologie – Vocabulaire*, Genève : Organisation internationale de normalisation (ISO 1087 – 1990.), 1990.

International Organization for Standardization (ISO). *Principes et Méthodes de la Terminologie,* Genève : Organisation internationale de normalisation (ISO 704 – 1987.), 1987

Jouis, C., "*Hierarchical Relationships "is-a": Distinguishing Belonging, Inclusion and Part/of Relationships",* LREC2006 (the 5th International Conference on Language Resources and Evaluation)., 571-574, Genoa : ELRA, 2006

Jouis, C. & Ferru, J.-M. "*Intranet Try To Find Project (ITTF): An approach for the searching of relevant information inside an organization*", LREC 2004: Language Resources and Technology Evaluation within Human Language Technologies, 1325-1329. ELRA – European Language Resources Association : Paris, 2004

Jouis, C., "*Logic of Relationships".* The Semantics of Relationships, 127-140. Dordrecht: Kluwer Academic Publishers, 2002

Jouis, C *"System of types + Inter-concept relations properties: Towards validations of constructed terminologies ?"*,Structure and Relations in Knowledge Organization. Proceedings of the Fifth International ISKO Conference, 39-47, 1998

Jouis & Mustafa, W. « *Vers un nouvel outil interactif d'aide à la conception de dictionnaires électroniques spécialisés »,* Lexicomatique et Dictionnairiques : Ivèmes Journées Scientifiques du Réseau Thématique « Lexicologie, Terminologie, Traduction ». 255-266. Beyrouth : AUPELF-UREF & F.M.A., 1996

Jouis, C. & Mustafa, W. *"Conceptual modelling of database sketch using linguistic knowledge: Application to terminological databases"*, Proceedings of the First Workshop on Applications of Natural Language to Data Bases, 103-118, 1995

Jouis,C. *Contributions à la Conceptualisation et à la Modélisation des Connaissances à partir d'une Analyse Linguistique de Textes : Réalisation d'un*

prototype : Le Système SEEK, Thèse de doctorat, Ecole des Hautes Etudes en Sciences Sociales : Paris, 1993

Lejeune, N. & Van Campenhoudt, M. *« Modèle de données et validité structurelle de fiches terminologiques; L'expérience des microglossaires de TERMISTI »*, La banque des Mots: Terminologie et Qualité, numéro spécial 8, 97-111, 1998

Meyer, I. & Skuce, « *Bases de connaissances et bases textuelles sur le web : Le système Ikarus. »,* Vèmes Journées Scientifiques : La mémoire des mots, 637-646. Tunis and Montréal : AUPELF & F.M.A., 1998

Meyer, I, & Mchaffie, C. B. *« De la focalisation à l'amplification: Nouvelles perspectives de représentation des données terminologiques »*, T.A.-T.A.O. : Recherches de pointe et Applications Immédiates : Troisièmes journées Scientifiques du Réseau Thématique de Recherche « Lexicologie, Terminologie et Traduction », 425-440, Montréal : AUPELF-UREF & F.M.A,. 1994

Miéville, D. *Un développement des Systèmes Logiques de Stanislaw Lesniewsky : Prothétique, Ontologie, Méréologie,* Bern : Frankfurt am Main, and New York : Peter Lang, 1984

Miller, G.A. *"Nouns in WordNet: A lexical inheritance system"*, International Journal of Lexicography, 3, 245-264, 1990

Molholt, P. *"Standardization of interconcept links and their usage"*., Knowledge Organization and Change: Proceedings of the Fourth International ISKO Conference, 65-71, 1996

Mustafa-Elhadi, W. & Jouis, C. "*Natural language processing-based techniques and their use in data modelling and information retrieval"*, Knowledge Organization for Information Retrieval : Proceedings of the Sixth International Study Conference on Classification Research, 157-161. The Hague: FID, 1997

Pribbenow, S. *"Meronymic Relationships: From Classical Mereology to Complex Part-Whole Relations",* The Semantics of Relationships, An Interdisciplinary Perspective. Chapter 3. 35-50. R. Green, C. A. Bean,

Sung Hyon Myaeng (Eds), Kluwer Academic Publishers: Dordrecht/Boston/London, 2002

Shaumyan, S. *A semiotic theory of language*. Bloomington: Indiana Univ. Press, 1987

Sowa, J. F. *Conceptual Structures: Information Processing In Mind And Machine*, Reading. Mass.: Addison-Wesley, 1984

Sowa, J. F. "*Knowledge Representation: Logical, Philosophical, and Computational Foundation",* Pacific Grove, CA: Brooks/Cole, 2000

Van Campenhoudt, M. *Abrégé de Terminologie Multilingue* [On-line]. Available : http://www.termisti.refer.org. [2007, March 16], 2007

Van Campenhoudt, M. *« Les relations notionnelles expérimentées dans les microglossaires de TERMISTI : Du foisonnement à la régularité »*, TA-TAO. : Recherches de Pointe et Applications Immédiates. Troisièmes Journées Scientifiques du Réseau Thématique de Recherche « Lexicologie, Terminologie et Traduction, 409-423. AUPELF-UREF & F.M.A.Montréal & Beyrouth, 1994)

Wielingua, B., Schreiber, A., Breuker, J. "*KADS:A modelling approach to knowledge engineering?"* Knowledge Acquisition, 4 (1), 5-53, 1992

Winston, M.E., Chaffin, R., Herrmann, D. *"A taxonomy of part-whole relation"*. Cognitive Science, 11, 417-444, 1987

Aide à la structuration d'ontologies à partir de l'analyse textuelle : travaux exploratoires

Henri Zinglé
CIRCPLES- EA 3159
Université de Nice – Sophia Antipolis

Résumé :

Nous présentons ici les premiers résultats d'une recherche destinée à évaluer l'intérêt de l'extraction de mots associés pour la constitution d'ontologies. Contrairement aux méthodes classiques qui se bornent à relever les coocurrents dans un certain cotexte, nous faisons appel à une grammaire qui sélectionne uniquement les associations vérifiées par des relations syntaxiques.

Dans la constitution d'ontologies l'exploitation de la documentation joue un rôle particulièrement important. Dans le logiciel ZText (Zinglé, 1998) nous proposions un ensemble d'outils pour effectuer l'extraction d'unités lexicales simples et complexes, la constitution de concordances et l'exploration des associations de mots. Dans nos travaux en cours nous entendons approfondir la réflexion en ce qui concerne les mots associés dans un certain cotexte. Parmi les problèmes rencontrés dans le passé avec ZText on peut citer d'une part la largeur du cotexte pris en compte pour l'analyse et d'autre part l'absence d'informations grammaticales concernant les unités lexicales associées par le logiciel à un vocable donné. En effet, la fonction "recherche de termes associés" du logiciel précité explore l'ensemble des mots associables à un vocable donné dans les limites de la phrase, à l'exception des mots outils (déterminants, prépositions, conjonctions et subjonctions); les unités lexicales associées sont listées avec l'indication

du nombre d'occurrences pour l'ensemble du corpus traité. On relève toutefois avec cette méthode un certain nombre d'associations intempestives, dues au fait que l'on se borne à relever les unités lexicales appartenant au même cotexte (la phrase) et qu'aucun analyseur syntaxique n'est utilisé. C'est ce problème que nous cherchons à corriger dans notre nouvel environnement de traitement linguistique ZTools.

ZTools est une refonte complète de la ZStation (Zinglé, 1998) qui tire parti d'une API linguistique programmée en Visual Prolog 6.3 orienté objet. Celle-ci repose sur un ensemble de classes, parmi lesquelles on peut citer

1. ZDBase : création et exploitation de bases de données relationnelles
2. ZDialogs : création de dialogues interactifs
3. ZDico : création et exploitation de ressources lexicales
4. ZGram : création et exploitation de ressources syntaxiques
5. ZList : gestion de listes
6. ZMorpho : création et exploitation de ressources morphologiques
7. ZSem : création et exploitation de ressources sémantiques
8. ZStats : fonctions et procédures d'analyse statistique
9. ZString : fonctions et procédures de traitement de chaînes de caractères
10. ZText : fonctions et procédures de traitement de données textuelles.

Le calcul des associations de mots appartient à la dernière classe, mais tire parti des autres classes mentionnées. L'idée principale vise à parcourir un document ou un ensemble de documents[27] et à rechercher pour chaque unité lexicale (UL) rencontrée qui n'est pas un mot outil[28]

[27] La classe ZText fournit également des outils pour l'organisation de corpus.

[28] Par mot outil nous entendons les déterminants, les conjonctions, les prépositions et les subjonctions.

l'ensemble des relations syntaxiques dont elle est soit le gouverneur soit le dépendant, puis à organiser les relations détectées.

Les relations sont notées r(UL,Info), où Info représente l'information associée à UL. Pour l'instant nous nous limitons aux cas suivants

adj(X)	X est un adjectif associable à UL qui est un substantif[29]
adv(X)	X est un adverbe associable à UL qui est un verbe ou un adjectif
sub(X)	X est un substantif associable à UL qui est un verbe, un substantif ou un adjectif
vb(X)	X est un verbe associable à UL qui est un substantif ou un adverbe

Les règles applicables à la détection des relations syntaxiques sont non déterministes, puisqu'il est évident d'une part qu'une relation r(A,sub(S)) implique logiquement la relation r(S,adj(A)) pour A représentant un adjectif et S un substantif et d'autre part qu'en un point de l'énoncé plusieurs règles peuvent s'appliquer aux mots précédents ou aux mots suivants.

Chaque règle fait appel à un ensemble de contraintes lexicales, morphologiques et syntaxiques qui peuvent être vérifiées par les fonctions appropriées de ZDico, ZMorpho et ZGram. Les paramètres morphosyntaxiques (genre, nombre, personne) sont extraits par

[29] Nous tenons compte ici non seulement de l'adjectif épithète mais également de l'adjectif attribut.

l'analyseur morphologique qui fait appel à un dictionnaire générique. Nous disposons actuellement d'un dictionnaire relativement complet pour le français (environ 45.000 entrées). Si d'aventure le dictionnaire se révèle insuffisant (dans le cas de textes spécialisés), il est possible d'effectuer grâce à une procédure appropriée de la classe ZText une analyse lexicale du/des document(s) et de déterminer les mots absents du dictionnaire. Grâce à la nouvelle approche choisie pour la codification mophosyntaxique[30] l'introduction de nouvelles unités lexicales dans le dictionnaire ne présente aucune difficulté. Les accords en genre et en nombre pour les associations adjectif/substantif et en nombre et personne pour les relations substantif/verbe lorsque le substantif est sujet sont vérifiées grâce à la fonction d'accord de la classe ZGram. Comme l'analyse morphologique d'un mot peut conduire à plusieurs analyses différentes pour une même catégorie morphosyntaxique[31], la difficulté principale a résidé dans l'introduction du déterminisme dans le cadre d'une règle non déterministe. Ce problème a été traité par l'introduction d'un coupe-choix dynamique limité à la vérification des contraintes morphosyntaxiques et syntaxiques :

[30] L'analyse et la génération morphologique adoptées dans ZTools est radicalement différente de l'approche utilisée pour la ZStation. En effet, nous avons renoncé à la programmation par classe morphologique pour adopter une articulation entre une grammaire statique (caractérisant les phénomènes morphologiques fondamentaux d'une langue) et une grammaire dynamique (décrivant les phénomènes qui ne sont pas en accord avec la grammaire statique). Cette nouvelle approche a le mérite de simplifier énormément la maintenance des dictionnaires.

[31] Par exemple, la forme verbale *traite* est analysée comme le verbe *traiter* à l'indicatif à la première ou troisième personne du singulier, comme subjonctif présent à la première ou troisième personne du singulier ou comme l'impératif de la seconde personne du singulier.

```
rel(LN,Dic,Mots)=Info:-
        Op=getBackTrack() and
        [vérification des contraintes morphosyntaxiques par
syntaxiques]
        [calcul des solutions] and
        cutBacktrack(Op) and
        Info=select(Sols).
```

De ce fait les règles restent globalement non déterministes tout en éliminant en interne les choix inutiles.

Les relations détectées sont mémorisées au fur et à mesure de la progression de l'analyse. Les unités lexicales impliquées dans une relation sont systématiquement lemmatisées pour faciliter le regroupement des informations au cours de l'étape suivante et pouvoir donner des indications sur la distribution des relations au sein du corpus.

A l'issue de cette première étape, on établit la liste des unités lexicales pour lesquelles des relations ont été détectées. Cette liste est triée et les doublets éliminés grâce aux fonctions appropriées de la classe ZList. Pour chaque unité lexicale on rassemble l'ensemble des relations r(UL,Info), en indiquant le nombre d'occurrences rencontrées lorsque Info est identique. Les résultats obtenus sont stockés dans un fichier de texte éditable.

Nous présentons dans le tableau ci-dessous les résultats obtenus pour l'analyse d'un certain nombre de résumés d'articles médicaux publiés par le *Laennec Digest's*[32].

[32] Il s'agit d'une revue animée par le Dr B. Bugnas avec le concours d'une quinzaine de spécialistes des maladies respiratoires proposant des synthèses d'articles médicaux destinées à l'ensemble des pneumologues français. Près de 40 revues sont analysées régulièrement et environ 10.000 articles sont actuellement enregistrés dans la base de données.

asthme	adj(agressif):1 adj(léger):1 vb(traiter):1 vb(présenter):1
bénéfice	adj(clinique):1 adj(persistant):1 vb(entraîner):1 vb(persister):1 vb(conférer):1
complication	adj(acceptable):1 adj(cardiovasculaire):1 adj(respiratoire):1 vb(prévenir):1
dose	adj(fort):1 adj(petit):2 adj(moyen):1 adj(seul):2 vb(cumuler):1
fonction	adj(cardiovasculaire):1 adj(mauvais):2 adj(moindre):1 adj(pulmonaire):11 vb(altérer):1
inflammation	adj(systémique):1 adj(bronchique):1
patient	adj(âgé):3 adj(consécutif):1 adj(évaluable):1 adj(particulier):1 vb(étudier):1 vb(hospitaliser):3 vb(présenter):1 vb(traiter):1 vb(réséquer):1 vb(souffrir):19
résultat	adj(actuel):1 adj(bon):1 adj(important):1

	adj(mauvais):1 adj(thérapeutique):2 vb(augmenter):1 vb(obtenir):1 vb(permettre):1
risque	adj(élevé):1 adj(global):1 adj(grand):1 adj(haut):1 adj(indépendant):1 adj(moindre):2 adj(traditionnel):1 vb(doubler):1 vb(réduire):1
significatif	sub(association):1 sub(amélioration):3 sub(façon):1 sub(protection):1
thérapeutique	sub(approche):1 sub(décision):1 sub(résultat):2 sub(technique):1
traitement	adj(approprié):1 adj(complémentaire):1 adj(difficile):1 adj(empirique):1 adj(initial):1 vb(nécessiter):1
trouble	adj(respiratoire):1 vb(aggraver):1

Il convient d'insister sur le fait qu'il s'agit là de l'analyse de quelques résumés seulement, destinée à illustrer les résultats que la fonction de recherche de mots associés permet d'obtenir; il va sans dire que l'analyse d'un corpus plus étendu ferait apparaître bien plus de relations avec des occurrences plus élevées.

On peut constater, à titre d'exemple, que dans le corpus traité l'unité lexicale *asthme* est liée aux adjectifs *agressif*, *léger* et aux verbes *présenter* (le patient présente un asthme) et *traiter* (le médecin traite l'asthme de son patient). L'unité lexicale *bénéfice* quant à elle est associée aux adjectifs *clinique*, *persistant* et aux verbes *entraîner* (un traitement entraîne un bénéfice), *persister* (le bénéfice d'un traitement persiste) et *conférer* (une molécule ou une action thérapeutique confère un bénéfice à l'état du patient). Des observations similaires peuvent être faites pour les substantifs *complication, dose, fonction, inflammation, patient, résultat, risque, traitement* et *trouble* ainsi que pour les adjectifs s*ignificatif* et *thérapeutique.*

Comparativement aux résultats obtenus précédemment avec le logiciel ZText (Zinglé, 1998), ceux que nous obtenus ici sont de bien meilleure qualité, dans la mesure où l'on ne s'intéresse pas à l'ensemble des cooccurents dans un certain cotexte (la phrase ou éventuellement une fenêtre d'analyse inférieure à la taille de la phrase) mais seulement aux cooccurrents validables au plan grammatical. La vision du corpus qui en résulte est de nature topologique et nous estimons que celle-ci est susceptible de permettre au lexicographe de dégager plus rapidement l'ontologie d'un domaine à partir du corpus qu'il étudie. Il importe en effet lorsqu'un concept est identifié au travers d'un substantif, par exemple, de pouvoir en cerner rapidement les propriétés par les actions qui le détermine ou qu'il implique ainsi que les qualificatifs qui permettent de le délimiter éventuellement par rapport à d'autres concepts. On notera également qu'en cas d'emplois polysémiques d'une unité lexicale, l'étude des relations est de nature à dégager rapidement les critères d'identification des concepts sous-jacents.

Il s'agit pour l'instant d'un travail exploratoire destiné à établir si l'approche suivie se révèle qualitativement meilleure que les méthodes utilisées antérieurement. Nous pensons, au vu de ces premiers résultats, qu'il est intéressant de poursuivre dans cette voie. L'étape suivante du travail consiste à améliorer le traitement des relations extraites. En effet, pour le corpus testé, pas moins de 1313 relations ont été extraites pour 573 unités lexicales. Nous entendons développer une interface graphique pour permettre au lexicographe d'appréhender plus aisément

l'information utile plutôt que de proposer une liste alphabétique d'unités lexicales associées à un ensemble de relations. Par ailleurs, dans ce travail exploratoire notre intérêt s'est porté sur un nombre limité de relations. Dans une version plus élaborée, nous souhaitons également étendre la couverture syntaxique de l'analyseur, en particulier pour traiter les groupes prépositionnels, les relatives et les subordonnées introduites par une subjonction.

Ouvrages cités

Zinglé H. (1998) ZTEXT : un outil pour l'analyse de corpus. *Travaux du LILLA*, n°3, Nice, pp 69-78

Zinglé H. (1999) *La modélisation des langues naturelles. Aspects théoriques et pratiques*. Travaux du LILLA (numéro spécial), 151 p.

Les nominalisations en *–tion* dans un texte techno-administratif

Pierre Lerat
Université Paris XIII
34, rue N.D. de Recouvrance, F- 45000 Orléans
pierre.lerat@wanadoo.fr

Résumé :

Les nominalisations conduisent à des blocs d'informations quand elles sont accompagnées d'actants. Elles abondent dans les textes spécialisés, notamment techno-administratifs, en français et ailleurs.

Un règlement communautaire établit la pertinence de ce mode d'accès aux textes et l'intérêt de traiter les formes concernées (ex. : organisation) comme des valeurs d'un attribut « nominalisation combinée » dans la gestion terminographique ou ontologique des noms d'objets (ex. : marché vitivinicole).

L'application de la même méthode à un texte comparable, plus long et de contenu différent, valide globalement ces résultats.

1. Introduction

Le travail présenté ici consiste en une terminologie sélective : un accès aux textes spécialisés en vue de leur compréhension en tant que textes. Il n'existe aucune méthode simple et robuste pour entrer dans les textes. Dans le cas des textes techniques, on peut tirer parti des nominalisations, qui ont la réputation d'y proliférer[33]. Parmi elles, celles en *–tion* dominent, ce qui invite à privilégier cette terminaison facile à

[33] Voir notamment Banks, 2002 et Condamines, 2003.

identifier. Son intérêt particulier est sa fréquence dans les titres, sous-titres et tables des matières, et aussi dans les textes spécialisés en général.

Les terminologues ont surtout travaillé jusqu'ici sur les dénominations longues, qu'elles aient comme tête, indifféremment, un nom d'entité ou un nom prédicatif. De même, les spécialistes des ontologies ont longtemps négligé ce que Grabar et Hamon appellent des « relations transversales » (2004 : 63). Un exemple peut néanmoins être emprunté à Zweigenbaum : « *des relations (…) comme* manifestation_de, *non hiérarchique mais utile en particulier pour la recherche d'information* » (2004 : 117). C'est précisément à des condensés d'information que permettent d'accéder les formes en *–tion.*

2. Pourquoi un règlement communautaire ?

Les terminologies sont de plus en plus des vocabulaires techno-administratifs. Le rapprochement entre terminologie et ontologie se trouve favorisé tout particulièrement dans les textes réglementaires qui régissent un nombre croissant d'activités dans l'Union Européenne. Ces textes comportent en effet beaucoup de définitions stipulatives faisant l'objet d'un consensus entre les États, ainsi que les expressions retenues pour nommer dans chaque langue les concepts correspondants.

Ainsi, les règlements communautaires sont triplement normatifs : conceptuellement, par l'harmonisation des points de vue en dépit d'intérêts différents, lexicalement, au risque de développer une « nomenclature d'expressions longues » (Lerat, 2007b), et enfin juridiquement, puisqu'il s'agit de favoriser l'existence d'un espace juridique commun. Cette dernière particularité n'est pas indifférente : l'obligation de faire connaître aussi largement que possible ces textes les rend téléchargeables sans restrictions. Elle permet aussi de tester les méthodes d'entrée dans les textes sur chaque version officielle, ce qui favorise le travail sur les concepts, et non pas seulement sur les mots.

Le choix d'un texte concernant la viniculture[34] est relativement fortuit : il se trouve que j'ai été amené à me pencher à plusieurs reprises sur cette terminologie de domaine. Il n'est pas du tout marginal, puisque l'on évalue à plus de 2000 le nombre de textes réglementaires communautaires portant sur la vitiviniculture. Dans le cas présent, ce qui a été retenu est un règlement bref, donc facilement reproductible en entier (en annexe), et à forte pertinence terminologique et ontologique puisqu'il s'agit exclusivement de promouvoir une méthode de mesure du titre alcoométrique des vins par la balance hydrostatique.

3. Pourquoi les nominalisations ?

La définition de la nominalisation varie malheureusement selon les auteurs. Ici, le mot est pris au sens le plus classique : remplacement d'une formulation verbale par une formulation nominale, « *nominalized process* » (Banks, 2002). Exemples : passage de *séparer les blancs des œufs* à *séparation des blancs des œufs* (action), de *ils se sont séparés* à *leur séparation* (processus), ou de *les pouvoirs sont séparés* à *la séparation des pouvoirs* (état). Il existe aussi des nominalisations à partir de formulations adjectivales, par exemple *la séparabilité des blancs* à partir de *les blancs sont séparables*. Les règlements et directives communautaires comportent surtout des noms d'actions en *–tion*.

Les nominalisations dans les textes spécialisés ne sont pas nécessairement des termes. Un exemple juridique que j'ai surexploité depuis 1988 est à cet égard un arbre qui masque la forêt. Il s'agit de *promulgation*, qui en français de France appelle l'usage d' « *arguments fortement contraints, tels que* loi *et* président de la République » (Lerat, 2006 : 91). Il arrive plus souvent que les noms prédicatifs résultant de nominalisations appartiennent à la langue courante ; c'est le cas, par exemple, dans le texte étudié ici, avec *organisation* et *application*. Ce sont les

[34] Règlement (CE) n° 128/2004 de la Commission du 23 janvier 2004 modifiant le règlement (CEE) n° 2676/90 déterminant des méthodes d'analyse communautaires applicables dans le secteur du vin.

noms de leurs arguments qui sont terminologiques : respectivement *marché vitivinicole* et *méthode de mesure*.

L'idée de « schémas d'arguments spécialisés » (Lerat, 2002) a l'avantage de ne pas préjuger du caractère terminologique ou non de l'expression prédicative. Elle présente en revanche plusieurs inconvénients non négligeables. D'abord, elle n'est pas opératoire parce que le même mot peut exprimer une action ou une métonymie de cette action, comme le montre bien la polysémie d'*organisation* (fait d'organiser, donc prédicat, nominalisation verbale, ou organisme, donc entité ontologique, nominalisation substantivale). Ensuite, elle est liée à une grammaire de la phrase simple issue de Z. Harris et M. Gross, et non pas à une grammaire de l'énoncé réalisé ; autrement dit, elle privilégie *in absentia* des schémas plausibles, non des cooccurrences observables. Enfin, elle met sur le même plan le verbe et le nom dérivé, ce qui ne tient pas compte des particularités des nominalisations.

Chaque substantif issu d'une nominalisation a son histoire individuelle, plus ou moins complexe, souvent atypique. Ainsi, dans la langue du droit français, *prestation de serment* correspond à *prêter serment*, mais *prestation de service* renvoie à un état de langue ancien (Lerat, 2007a).

Pour résumer, les nominalisations vivantes ne sont pas forcément des termes, même dans les textes spécialisés, elles sont susceptibles de compléments non prévisibles, et elles ont le même contenu conceptuel que les formulations verbales correspondantes.

Il est d'observation courante que les nominalisations ainsi comprises sont particulièrement fréquentes dans les textes spécialisés. Ce qui n'a pas été exploité systématiquement, en revanche, et que je voudrais mettre en évidence, c'est qu'autour d'elles gravitent, les saturant dans un contexte étroit, non seulement des actants mais aussi des circonstants, qui désignent des éléments de ce qui est donné comme le réel pertinent, autrement dit une ontologie spécialisée.

4. Pourquoi les formes en *–tion* ?

Pour un traitement automatique comme pour un traitement manuel, *-tion* présente quatre avantages :

- cette forme n'est pas du tout bruyante
- grammaticalement, c'est toujours un nom féminin, à l'exception de *cation*, nom donné aux ions positifs
- le pluriel est rare en cas de nominalisation (moins dans les emplois concrets, en tant que noms d'arguments)
- *-tion* évite de distinguer *–ation, -ition* et *–ution*, en les englobant

Il faut aussi accepter deux inconvénients :

toute forme en *–tion* n'est pas nécessairement une nominalisation vivante ; ainsi, dans le texte considéré, l'annexe, qui n'est pas reproduite à cause de sa longueur, comporte des noms d'objets (comme *solution de 100 ml*) et des expressions non saturées (comme *introduction*)

on trouve dans le corpus d'autres nominalisations : dans le texte principal les dérivés régressifs *analyse, mesure* et *contrôle* et, dans l'annexe, des dérivés utilisant un autre suffixe, comme *étiquetage*

La pioche est très bonne quand on accède par la forme en *–tion* à des compléments tels que *marché vitivinicole, méthode de mesure, méthode de contrôle* et *Journal officiel de l'Union européenne*. Elle l'est également quand les arguments font l'objet de simples désignations anaphoriques comme *sa* renvoyant à *règlement* dans *sa publication*. Elle est moins bonne quand elle conduit à *comité de gestion des vins*, nom d'institution.

5. Résultats pour le texte français

On observe que les principaux éléments de l'univers du discours sont introduits par *–tion* : le marché vitivinicole, qui est l'enjeu, la méthode de mesure, qui est l'objet considéré, le règlement lui-même, qui

est le support. Linguistiquement, on peut noter un autre intérêt : la désambiguïsation des noms polysémiques *organisation (du marché)* et *application (de méthodes)*.

Il manque à première vue l'essentiel : le nom de la méthode de mesure. En fait, il est accessible par les segments *description de cette méthode* et *validation de celle-ci*, où le démonstratif renvoie dans les deux cas à l'expression longue *méthode de mesure du titre des vins par la balance hydrostatique*.

Le besoin d'un référentiel externe n'en demeure pas moins. C'est le cas pour le vocabulaire des institutions comme *CE, CEE, Commission, Conseil* etc.. Pour savoir ce que dénomme *titre alcoométrique volumique* ou *balance hydrostatique*, il est possible de trouver l'information dans les textes mêmes, grâce à leurs annexes et leurs renvois.

De même que les ontologies ne valent que pour une application donnée, les points de vue sur les objets varient avec les textes. Ainsi, le titre d'un vin est une simple mention obligatoire dans les textes sur l'étiquetage : c'est l'indication qui comporte *alcool, alc.* ou *% vol.* Ici, au contraire, c'est le résultat d'un calcul au moyen de la balance hydrostatique de dernière génération. Dans le premier cas[35] on a affaire à une ontologie commerciale, dans l'autre à celle des laboratoires d'œnologie.

6. Traitement terminographique

Un texte qui contiendrait explicitement à lui seul la totalité des connaissances nécessaires à sa compréhension serait un artefact illisible, avec une définition stipulative par dénomination. Il faut donc distinguer les définitions, qui pour toute matière spécialisée sont nécessairement encyclopédiques et s'acquièrent par synthèse des ressources d'une

[35] Voir Sánchez Nieto, 2006.

« terminologie transtextuelle » (Lerat, 2006 : 96), et les liens sémantiques entre termes, qui peuvent en partie être repérés dans le texte lui-même.

Que faire quand on est en présence de relations syntagmatiques, comme celles entre les nominalisations et leurs arguments ? Les traiter paradigmatiquement. En matière de relations transversales en général, Grabar et Hamon notent qu' « *il n'existe pas d'outils dédiés spécifiquement à l'acquisition de ces relations. Toutefois, l'acquisition de ces relations peut profiter de techniques existantes et dédiées à l'origine à d'autres relations* » (2004 : 80).

Le mode de fonctionnement des bases de données fournit une ressource : le jeu des attributs et des valeurs d'attributs. Un attribut étant « *une caractéristique d'un objet ou d'une entité* » (ISO 11179), la combinaison d'un nom d'objet avec une nominalisation peut être considérée comme une propriété de ce nom d'objet. Une valeur d'attribut étant « *une représentation d'une instance d'un attribut* » (*ibid.*), telle combinaison concrète avec telle nominalisation peut être considérée comme une instance de cette propriété.

Les nominalisations combinées effectivement à des noms d'entités dans le texte se prêtent au traitement suivant : à partir de *description de cette méthode, utilisation de cette méthode, validation de celle-ci* et *sa publication au Journal officiel de l'Union européenne* on obtient au moyen de NOMICOM (abréviation pour *nominalisation combinée*) :

méthode de mesure du titre alcoométrique des vins par la balance hydrostatique / NOMICOM : *description, utilisation, validation*

règlement / NOMICOM : *publication*

7. Résultats pour d'autres langues

En dehors du français, les versions examinées le sont dans l'ordre alphabétique des abréviations officielles : de, en, es, it, pl.

7.1. En allemand

Un rendement comparable à celui de *–tion* pour le français est obtenu à partir de *–ung*. Toutefois, les nominalisations n'apparaissent pas aux mêmes endroits, ce qui montre que la méthode est productive, mais de façon aléatoire. On peut aussi noter la présence de noms d'objets ou entités en *–ung* (*Verordnung, Verwaltung*) et d'une nominalisation non saturée (*Validierung*). Il apparaît également que le mode de composition courant de l'allemand conduit à l'intégration de la terminaison du premier composant (ici, *Validierungsparameter*). Enfin, comme en français, il existe d'autres terminaisons de noms d'actions (Scheffler, 2005), à commencer par *–tion* (ici, *Marktorganisation*).

7.2. En anglais

Les nominalisations en –tion, qui dans les textes scientifiques sont les plus fréquentes (Banks, 2002), sont fortement concurrencées, ici, par d'autres formes : formes radicales (analysis, use), formes en –ment (management, measurement), formes en –ing (measuring, using). On trouve néanmoins les homographes du français description, organisation et publication, qui conduisent aux entités pertinentes : market in wine, this method et Official Journal of the European Union.

7.3. En espagnol

Comme mesure se dit en espagnol medición, le rendement de *–ción* y est encore meilleur que celui de *–tion* en français.

7.4. En italien

Là aussi, et il en irait de même en portugais, la parenté linguistique promet des résultats voisins. Toutefois, l'histoire individuelle de chaque langue et de chaque mot joue son rôle : pas de surprise avec *organizzazione, descrizione, applicazione* et *pubblicazione*, qui sont les indicateurs de l'univers du discours, mais des différences lexicales çà et là (*misurazione* / *mesure, utilizzo* / *utilisation, convalida* . *validation*).

7.5. En polonais

Le test sur une langue slave impose la prise en compte de la déclinaison : non seulement *–nie*, mais aussi *–nia, -niem* et *–niu*. Dans les autres langues indo-européennes considérées plus haut, seule la mise au pluriel compliquait la tâche, et c'est d'autant moins gênant que le procès, qui seul importe pour les nominalisations vivantes, s'accompagne généralement du singulier (Condamines, 2003 : 104).

Le polonais met aussi en présence d'autres terminaisons (par exemple, l'équivalent d'*organisation* est *organisacja*). Il utilise volontiers un adjectif de relation à la place d'un génitif (ainsi, *comité de gestion des vins* se dit *Komitetu zazadzajacego ds. Wina*). Enfin, *-nie* ne se trouve pas que dans des noms (par exemple, *zgodnie z* veut dire « en accord avec »).

Inutile de multiplier les exemples de langues : la preuve est faite que la terminologie et l'ontologie textuelles ne sont productives que dans les limites autorisées par la typologie des langues. Ainsi, s'il existait une version arabe de ce texte, la méthode explorée ici ne vaudrait rien, puisque le nom d'action dans cette langue (le masdar) fait partie de la conjugaison du verbe et varie donc selon le schème et la racine à la fois.

8. Validation de la méthode par un autre test

Il faudrait multiplier les expériences sur des textes spécialisés diversifiés thématiquement et discursivement (recherche, didactique, vulgarisation, mode d'emploi etc.). On se contentera ici de prendre un texte comparable (un règlement communautaire) mais portant sur un sujet tout autre[36] et d'une longueur plus importante (5 pages).

[36] Règlement (CE) n° 1820/2003 du Parlement européen et du Conseil du 22 septembre 2003 concernant la traçabilité et l'étiquetage des organismes génétiquement modifiés et la traçabilité des produits destinés à l'alimentation humaine ou animale produits à partir d'organismes énétiquement modifiés, et modifiant la directive 2001/18/CE.

Les résultats obtenus ne sont guère différents. En français, l'entité *OGM* a pour NOMICOMB *application, circulation, détection, dissémination, identification* et *information*, autrement dit tout ce qui est matière à débat. Le prédicat *définition* est mis au pluriel, dans un titre d'article, où il correspond à des résultats (le contenu conceptuel retenu), et *législation* désigne un ensemble concret de textes.

En allemand, *GVO* (équivalent d'*OGM*) a pour NOMICOMB des noms en *–ung*, avec un excédent par rapport au français : *Kennzeichnung* (« étiquetage »).

En anglais, la méthode est un peu moins productive, du fait que la nominalisation y a souvent une forme sans suffixe (ex. : *release* / fr. *dissémination*, de. *Freisetzung* ; *change* / fr. *modification*, de. *Änderung*). Inversement, *implementation* est une nominalisation plus repérable que son équivalent français *mise en œuvre*.

En espagnol, *OGM* bénéficie de bons NOMICOMB : *comercialización, detección, identificación* et *modificaciones geneticas*. Au demeurant, on observe par rapport au français des différences en plus et en moins. En plus, *comercialización* / *mise sur le marché* et *aplicación* / *mise en œuvre* ; en moins, *cambio* / *modification, propuesta* / *proposition* et *movimiento* / *circulation*.

En italien, les NOMICOMB *d'OGM* sont *circolazione, identificazione, modificazioni genetiche* et *rilevazione*. La méthode a contre elle *proposta* / *proposition* et *verifica* / *vérification*, mais pour elle *attuazione* / *mise en œuvre*.

En polonais, deux précautions sont nécessaires pour que la pioche soit correcte : prendre en compte, outre le nominatif en *–nie*, le datif en *–nia* et l'instrumental en *–niem*, ce que permet un analyseur morphologique, et négliger l'adverbe *genetycznie*, ce qui suppose une analyse syntaxique.

9. Conclusions

Les formes en –*tion* condensent des informations spécialisées si les textes le sont

Elles ne constituent qu'une partie (majoritaire dans les textes spécialisés) des nominalisations. On peut donc étendre l'investigation à d'autres terminaisons, mais avec des risques de bruit (notamment dans le cas de –*ment*). Pour tirer le meilleur parti terminologique et ontologique des nominalisations en –*tion*, il paraît recommandable d'opérer en trois temps : inventaire de ces formes, repérage de leurs arguments, exploitation des concordances de ces derniers pour accéder à leurs autres NOMICOMB

L'élaboration d'une ontologie ascendante reste d'abord une affaire de dénomination des objets pertinents, mais les NOMICOMB constituent un vocabulaire d'opérations qui est de grande importance : il s'agit de ce qui peut et doit être fait par tel agent sur tel objet dans telles conditions, autrement dit d'une langue du travail

Au-delà du cas particulier des textes techno-administratifs, qui ne sont pas seulement une langue de bois mais un encadrement juridique de pratiques existantes, comme la commercialisation des vins ou la prolifération des OGM, on trouve des gisements de nominalisations dans les tables des matières et les résumés de travaux scientifiques, dans les contrats etc.. Les formes en –tion constituent donc de bons matériaux pour la terminologie, l'ontologie spécialisée et la documentation.

Bibliographie

D. Banks, « Types of Nominalization in Scientific English », 2002, *www.univ-brest.fr/erla/membres/bankdocs/FEST02.DOC*

A. Condamines, *Sémantique et corpus specialises: constitution de bases de connaissances terminologiques*, mémoire d'habilitation, Université de

Toulouse 2, 2003, w3.univ-tlse2.fr/erss/textes/pagespersos/acondami/HDR-2.pdf

N. Grabar et T. Hamon, « Les relations dans les terminologies structurées : de la théorie à la pratique », *Revue d'intelligence artificielle*, 18-1, p. 57-85

ISO, norme 11179 « Technologies de l'information – Spécifications et normalisation des éléments de données » (en révision), *www.services.gouv.qc.ca*

P. Lerat, « Vocabulaire juridique et schémas d'arguments juridiques », *Meta*, 47-2, p. 155-162, 2002, *www.erudit.org/revue/meta/2002/v47/n2*

P. Lerat, « Terme et microcontexte. Les prédications spécialisées » in *Mots, termes et contextes*, D. Blampain, P. Thoiron et M. Van Campenhoudt edd., Paris, AUF, 2006, p. 89-98

P. Lerat, *Vocabulaire du juriste débutant*, Paris, Ellipses, 2007 (Lerat 2007a)

P. Lerat, « Langue et production de sens dans un texte communautaire », à paraître dans les mélanges offerts à Leandro Schena, Paris, L'Harmattan, 2007 (Lerat 2007b)

M.T. Sánchez Nieto, « La terminologia jurídica del etiquetado y el embotellado en español y en alemán en la legislación comunitaria » in *El lenguaje de la vid y el vino y su traducción*, M. Ibánez Rodrigues et M.T. Sánchez Nieto edd., Université de Valladolid, 2006, p. 195-214

T. Scheffler, « Nominalization in German », 2005, *www.ling.upenn.edu/~tatjana/papers/scheffler-nom.pdf*

P. Zweigenbaum, « L'UMLS entre langue et ontologie : une approche pragmatique dans le domaine médical », *Revue d'intelligence artificielle*, 18-1, p. 11-137

Corpus[37]

Règlement (CE) n° 128/2004 de la Commission

modifiant le règlement (CEE) n° 2676/90 déterminant des méthodes d'analyse communautaires applicables dans le secteur du vin

LA COMMISSION DES COMMUNAUTÉS EUROPÉENNES,

vu le traité instituant la Communauté européenne,

vu le règlement (CE) n° 1493/1999 du Conseil du 17 mai 1999 portant organisation commune du marché vitivinicole, et notamment son article 46, paragraphe 3,

considérant ce qui suit :

La méthode de mesure du titre alcoométrique des vins par la balance hydrostatique a été mise à jour et validée selon des critères internationalement reconnus. La nouvelle description de cette méthode a été adoptée par l'Office International de la Vigne et du Vin lors de son Assemblée Générale de 2003.

(1) L'utilisation de cette méthode de mesure peut assurer un contrôle plus simple et plus précis du titre alcoométrique volumique des vins et éviter les litiges dus à l'application de méthodes de contrôle moins précises.

[37] Corpus : « collection de textes (éventuellement un seul texte) constitué à partir de critères linguistiques ou extralinguistiques pour évaluer une hypothèse linguistique ou répondre à un besoin applicatif » (Condamines, 2003 : 32). Dans le cas présent, le texte unique (sans son annexe) est utilisé pour évaluer une hypothèse linguistique, en vue de fournir un mode d'emploi pour répondre à des besoins applicatifs.

(2) Il convient d'introduire au chapitre 3 de l'annexe du règlement (CEE) n° 2676/90 de la Commission la description mise à jour de cette méthode accompagnée des valeurs expérimentales des paramètres de validation de celle-ci.

(3) Il y a lieu de modifier le règlement (CEE) n° 2676/90 en conséquence.

(4) Les mesures prévues au présent règlement sont conformes à l'avis du comité de gestion des vins,

A ARRÊTÉ LE PRÉSENT RÈGLEMENT :

Article premier

À l'annexe du règlement (CEE) n° 2676/90, le chapitre 3 « Titre alcoométrique volumique » est modifié comme suit :

1) Au paragraphe 2, le point 2.3.2. est supprimé.

2) Après le paragraphe 4, le texte figurant à l'annexe du présent règlement est inséré en tant que paragraphe 4 bis.

3) Au paragraphe 5, le point 5.2 « Densimétrie par la balance hydrostatique » est supprimé.

Article 2

Le présent règlement entre en vigueur le septième jour suivant celui de sa publication au Journal officiel de l'Union européenne.

Le présent règlement est obligatoire dans tous ses éléments et directement applicable dans tout État membre.

Fait à Bruxelles, le 23 janvier 2004

Par la Commission

Franz Fischler

Membre de la Commission

Portage linguistique d'applications de gestion de contenu

Najeh HAJLAOUI, Christian BOITET
GETALP, laboratoire LIG, Université Joseph Fourier, CNRS, INPG, INRIA
385 rue de la Bibliothèque, BP 53
38041 Grenoble, cedex 9, France
Najeh.Hajlaoui@imag.fr
Christian.Boitet@imag.fr

Résumé

Nous nous intéressons à la multilinguïsation, ou "portage linguistique" (plus simple que la localisation) des services de gestion de contenu traitant des énoncés spontanés en langue naturelle, souvent bruitée mais contrainte 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. Trois stratégies de portage vers une langue L2 ont été étudiées : (1) traduction des énoncés de L2 vers L1 ; (2) localisation "interne", i.e. adaptation à L2 de l'EC, donnant `EC-App-L2` ; (3) 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)`. Le choix de la stratégie est contraint par la situation traductionnelle : types et niveau d'accès possibles, ressources disponibles, compétences langagières et linguistiques des intervenants pour la multilinguïsation des applications. Les stratégies (2) et 3) ont été expérimentées sur le portage d'arabe en français de la partie de CATS concernant l'occasion automobile. CATS est une application de e-commerce construite par D. Daoud et déployée en Jordanie sur le réseau FastLink. 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). 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.

Introduction

La multilinguïsation des services de e-commerce traitant des énoncés spontanés en langue naturelle est un problème difficile, et de ce fait très peu de services le font. Des facteurs principaux dépendant de la situation traductionnelle interviennent :

- le niveau d'accès aux ressources des applications, avec quatre cas possibles: accès complet au code source, accès limité à la représentation interne, accès limité au dictionnaire, et aucun accès.

- le niveau de compétence langagière et linguistique des intervenants dans la « portage » vers une nouvelle langue : connaissance des deux langues, source et cible, et compétences en TALN.

La multilinguïsation ou le « portage linguistique » dont nous parlons n'est pas une « localisation », qui implique une adaptation à un autre contexte culturel. Il s'agit uniquement de permettre l'accès dans plusieurs langues à un service de e-commerce, tel qu'il est et où il est.

Nous présentons d'abord le besoin en localisation d'application d'extraction de contenu suivi d'une analyse des méthodes possibles. Ensuite, nous illustrons notre étude par le portage linguistique (arabe vers français) d'une application d'e-commerce déployée, pour laquelle les facteurs présentés ci-dessus sont assurés, et plusieurs stratégies de multilinguïsation sont alors possibles. Nous présentons deux stratégies de localisation, dites « interne » et « externe » et nous évaluons leurs résultats.

1. Contexte du problème

1.1. Application munies d'un extracteur de contenu

Les applications qui nous intéressent sont celles qui donnent de la valeur ajoutée par le traitement du contenu de messages spontanés en langue naturelle. Les principaux types d'applications et de services susceptibles de le faire sont : la catégorisation de documents divers (Bélanger 2003) (dépêches de l'AFP « Agence France-Presse », brèves des différentes bourses, messages de clients à un serveur de SAV), l'extraction d'informations pour nourrir ou consulter une base de données (petites annonces ciblées, FAQ intelligentes, indexation ciblée à un domaine/métier), les hotlines automatisées, et plus généralement l'interaction en langage naturel avec des bases de données (Califf 1998).

Ces applications reposent en général sur un « extracteur de contenu » (EC) (Cardie 1997) plus ou moins puissant, produisant une représentation formelle du contenu extrait. Il peut s'agir d'une liste de propriétés (couples attribut-valeur), ou d'une forme logique, ou d'une forme arborescente plus ou moins « plate » (IF de CSTAR/Nespole !) (Besacier, Blanchon et al. 2001), etc.

1.2. Importance croissante de la multilinguïsation des services

La plupart des services à destination d'utilisateurs finals, et dans une seule langue, sont en anglais. Par exemple, CISCO (*http://www.cisco.fr/*) ne distribuait sa documentation qu'en anglais jusqu'à un passé récent. En Asie (Chine, Corée, Japon), bien que tous les utilisateurs visés aient étudié l'anglais 8 ou 9 ans, l'anglais technique n'est pas du tout bien compris, et les centres d'appel, qui coûtent très cher, étaient débordés. La production de traductions automatiques (Systran) a permis de diminuer notablement le recours aux centres d'appel, malgré leur assez mauvaise qualité.

En ce qui nous concernent, nous visons non pas à créer des services de traduction comme dans cet exemple, mais à multilinguïser l'accès à des services comme pourrait l'être un centre d'appel ou un service de SAV automatisé.

Notre exemple principal sera un système de petites annonces déployé à Amman en arabe (Daoud 2005) : notre but sera alors de le rendre accessible en français à des francophones résidant à Amman, puis dans d'autres langues pour les locuteurs de ces langues.

La nécessité de services multilingues sur place est très claire dans des pays multilingues (Canada, Inde, USA), mais elle apparaît aussi dans des pays monolingues (France) à cause du tourisme et de la nouvelle mobilité.

1.3. Malgré l'intérêt des énoncés spontanés, peu de services les traitent

L'interaction avec un service au moyen de formulaires présente des limites :

- l'interaction n'est pas naturelle ;
- surtout, les formulaires et les menus, à caractère modal et figé, ne permettent pas aux utilisateurs d'exprimer ce qu'ils veulent, comme par exemple de décrire le contexte du dysfonctionnement d'un logiciel ou d'un graveur de DVD.

En e-commerce, la navigation par mots-clés pilotée par des menus, telle qu'on la trouve dans la plupart des sites commerciaux, tend à accabler et frustrer les utilisateurs avec des interactions prolongées et rigides (Ritchie 1995).

L'intérêt de l'utilisateur pour un site particulier diminue exponentiellement avec l'augmentation du nombre de clics de souris

(Huberman, Pirolli et al. 1998). Par conséquent, le raccourcissement du chemin d'interaction pour fournir des informations utiles devient important.

Beaucoup de sites de e-commerce essayent de résoudre ce problème en fournissant des possibilités de recherche par mots-clés. Cependant, il s'agit du grand public, et outre le défaut d'ergonomie signalé plus haut, il y a un problème de compétence, car il faut que les utilisateurs connaissent le jargon spécifique du domaine.

Peu de services traitent des énoncés spontanés et cela, même en contexte monolingue. En interrogeant plusieurs moteurs de recherche (Google, Altavista, Tiscali…) sur les applications traitant des énoncés spontanés en langues naturelles, avec des requêtes variées[38], nous avons obtenu très peu de résultats positifs, et très peu de renseignements sur le fonctionnement interne de tels services et leur multilinguïsation, quand on en trouve. Il semble qu'il y en a encore très peu ! Nous avons cependant trouvé :

- Pertinence Summarizer (Lehman 1996), un logiciel de résumé automatique de textes multilingues ;
- Amilcare (Ciravegna 2001), un système adaptatif d'extraction d'information ;
- NLSA « Naturel Language Sales Assistant », un système basé sur le dialogue à travers le Web déployé par IBM) ;

[38] Requêtes : localizing natural language message processors, localization NLP free text, localization NLP interfaces, multilingual customer message processing, multilingual customer messages tools, multilingual, customer relationship processing, multilingual NLP e-commerce, multilingual online sales customer support, multilingual online sales NLP customer support, categorizing natural language messages, handling natural language messages in business, Natural Language Conversational Interface in Online Sales….

- CATS « Classifieds Ads Through SMS » (Daoud 2006), un système d'achat et de vente de voitures d'occasion et d'immobilier basé sur l'utilisation des SMS en arabe.

Notre hypothèse quant aux raisons possibles qui font que peu de services traitent les ESLN (Enoncés Spontanés en Langues Naturelles) est que ce travail présente des difficultés inhérentes et que la multilinguïsation est perçue comme un gros problème. On rencontre en effet des problèmes un peu analogues à ceux de l'oral : grammaire « non standard » (plus ou moins proche de l'oral), abondance d'erreurs (fautes de frappe, d'orthographe), utilisation de conventions typographiques propres au contexte (abréviations propres aux SMS et à la langue "tchatée", utilisation d'émotions pour noter les émotions et l'affect).

Assez souvent, on est dans un sous-langage relativement éloigné de la langue générale, comme les petites annonces ou des alarmes/avertissements (trafic routier, catastrophes naturelles). On ne peut pas utiliser des outils faits pour du langage écrit général et « propre ». De plus, il faut « traduire » les ESLN dans un formalisme de représentation de contenu (Content Representation Language ou CRL), et chaque application possède son propre formalisme.

1.4. Nécessité d'une approche spécifique à chaque sous-langage et inefficacité des outils faits pour les langues générales

Les applications traitant les ESLN utilisent en général une représentation du contenu. On trouve plusieurs formes de représentation de contenu : listes <attribut, valeur(s)>, structures de traits typés, expressions logiques (Prolog), expressions logico-fonctionnelles, objets (classes (méthodes, attributs), instances).

Par exemple, le système CATS utilise une représentation de type `Propriété=couleur{objet=saloon, valeur=bleu}`, dans le domaine de l'occasion automobile, pour exprimer que la couleur d'une voiture (saloon) est bleue.

Souvent, une application possède son propre formalisme et l'adaptation ou le portage d'un « extracteur de contenu » d'une application à l'autre, même pour la même langue, est difficile car il faut pouvoir garantir un niveau minimum de qualité, (exactitude et complétude) de l'extraction de contenu, de pertinence des réponses produites (traitement et gestion) et de l'adéquation linguistique de ces réponses.

La donnée primaire à fournir pour un portage linguistique d'une application traitant les ESLN est un corpus d'ESLN relatif à la même tâche et dans la langue cible, ce qui n'est pas toujours facile à trouver. Il faut travailler le plus souvent par adaptation, simulation et imagination. Par exemple, il nous faudra construire un corpus, au départ nécessairement imaginaire, de SMS supposés écrits par des francophones désirant acheter ou vendre de l'occasion automobile en Jordanie.

Généralement, le « sous-langage » (Sekine 1994) des énoncés spontanés en langue naturelle associés à un service donné est relativement éloigné de la « langue générale ». Par conséquent, les outils existants faits pour la langue générale ne marchent pas, qu'il s'agisse d'outils de TA (Traduction Automatique) ou d'EC (Extraction de Contenu). Une approche spécifique à chaque sous-langage (Slocum 1986) s'impose alors.

2. Approches possibles

2.1. Traduction automatique des énoncés vers la langue originale

Une première idée consiste à traduire les ESLN de la « nouvelle langue » L2 vers la langue originale L1 de l'application à localiser. L2 est « cible » du portage, mais « source », pour la traduction !

Quelle que soit l'approche linguistique choisie pour cette TA (Traduction Automatique), il faut créer un système spécialisé, et donc

disposer d'un corpus parallèle L2//L1. On peut l'obtenir en traduisant le corpus des ESLN, disponible par hypothèse en L1, sachant bien que le corpus parallèle obtenu sera « à l'envers » et donc nettement moins représentatif qu'un corpus L2//L1. Mais enfin c'est un début.

La question qui se pose ensuite est la taille du corpus nécessaire. Si l'on utilise une approche calculatoire de la TA « fondée sur des corpus » (TA statistique (Koehn 2004), TA par analogie (Lepage 2006)), on sait qu'il faut d'énormes corpus s'il s'agit de langue générale (entre 50 et 200 millions de mots d'après K. Knight et Ph. Koehn), bien plus grands que ceux disponibles après deux ou trois ans de fonctionnement d'un e-service. Il est possible que, dans le cas de sous-langages restreints, des corpus beaucoup plus petits suffisent, mais ce n'est qu'une hypothèse, et nous n'avons trouvé aucune étude sur le sujet. Nous avons commencé à travailler sur ce point, mais n'avons pas encore de résultat.

Si l'on utilise une approche calculatoire « par règles », il faut disposer de linguistes computationnels, ce qui est rare.

En résumé, le portage par réalisation d'un système de TA L2→L1 est possible en théorie, mais nous ne sommes pas encore en mesure de déterminer si on peut le faire, sans linguistes qualifiés, par des méthodes d'apprentissage automatique.

2.2. Réalisation d'un nouvel extracteur de contenu (EC) pour chaque langue visée

La réalisation d'un nouvel EC peut se faire par plusieurs méthodes.

La première solution est d'adapter l'EC existant de L1 à L2, mais cela n'est viable, que si

- les développeurs acceptent d'ouvrir leur code ou leur boîte à outils (BàO) à des collaborateurs nécessairement éphémères ;

- ce code ou cette BàO est assez facile à maîtriser ;
- les ressources ne sont pas trop lourdes à créer (en particulier le dictionnaire s'il existe) ;
- la maintenance peut ensuite se faire à coût faible, par des collaborateurs épisodiques.

Cette méthode d'adaptation interne de l'EC natif nécessite bien sûr une formation de l'équipe de localisation aux outils et aux méthodes utilisées.

La deuxième solution consisterait, pour une société voulant offrir des services de multilinguïsation/portage, à implémenter un EC générique et à l'adapter à chaque situation (langue, sous-langage, domaine, CRL — « Content Representation Language », tâche, contraintes). On verra ci-dessous que cela semble très difficile à envisager actuellement.

Dans beaucoup de contextes, les « multilinguïseurs » n'auront donc pas accès à l'EC de l'application, ni à un EC « universel ». Une troisième solution pourrait alors être de rechercher et d'adapter un EC existant et disponible, soit pour la langue L2, pour un domaine et/ou une tâche différents, soit pour le même domaine et la même tâche, pour une autre langue (différente de L2).

Dans ce qui suit, nous illustrons la première et la troisième méthode par le cas du portage d'arabe en français du système CATS. Cette expérience vise à permettre à des francophones vivant en Jordanie (à Amman) et disposant d'un mobile d'envoyer des SMS pour vendre et acheter des voitures d'occasion.

3. Illustration : localisation du système CATS

3.1. Présentation de CATS

CATS est un système d'achat et de vente basé sur l'utilisation des SMS en arabe (Daoud 2006). Il est déployé en Jordanie par la société FastLink, le plus gros opérateur local de téléphonie mobile. Bien qu'il n'y ait pas de transaction directe, CATS aide les utilisateurs à vendre et acheter sans avoir à se déplacer, en les mettant en contact.

Les SMS sont envoyés à un numéro spécial unique [39]. Leur contenu est extrait dans le langage CRL-CATS, puis transformé en requêtes SQL. Une réponse est envoyée automatiquement à l'expéditeur du SMS en cas de correspondance de la demande avec l'une des propositions. Si rien n'est trouvé, le système le dit dans sa réponse, et réessaie plus tard quand la base de données change.

CATS a deux principaux composants : un EC (Extracteur de Contenu) et un gestionnaire de requêtes QM « Query Manager ».

Voici un exemple de SMS et de sa représentation CRL-CATS produite automatiquement.

[39] Ils sont enregistrés, ce qui nous fournit un corpus d'ESLN en arabe pour CATS.

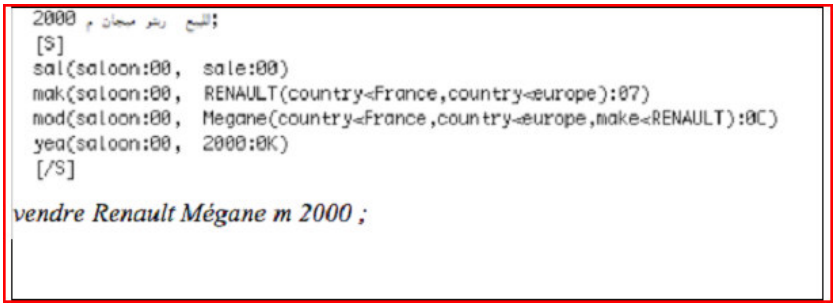


Figure 1 : exemple de représentation CRL-CATS

Dans cet exemple, la propriété est le type `(make)`, l'objet est une voiture `(saloon)` et la valeur est égale à `(RENAULT(country<France, country<europe))`. Pour la propriété modèle `(mod)`, l'objet est une voiture `(saloon)` et la valeur est égale à `(Megane(country<France, country<europe, make<RENAULT))`. Pour la propriété année `(yea)`, l'objet est une voiture `(saloon)` et la valeur est `(2000)`.

Le QM permet de convertir la représentation CRL-CATS vers un texte SQL (requête de sélection pour l'achat et/ou requête la d'insertion pour la vente). Il traite aussi les situations dans lesquelles aucune réponse n'a été trouvée.

3.2. Besoin d'un corpus de démarrage

Pour toutes les méthodes de localisation de CATS vers le français, la première chose à faire est de constituer un « corpus de démarrage » en français, analogue à celui utilisé par D. Daoud au départ de son projet pour l'arabe. Cela est évidemment nécessaire pour étudier la forme syntaxique des SMS à traiter en français, et aussi pour voir à quelles variantes lexicales il faut s'attendre.

Une première idée pour fabriquer un corpus français de démarrage est de partir du corpus CATS en arabe et de le traduire en

« SMS français spontanés », supposés être envoyés (en Jordanie) par des francophones.

Une traduction « brute » produite par un non Français est généralement très différente d'une traduction naturelle et fonctionnelle produite par un Français, c'est-à-dire de ce que dirait un Français d'une façon spontanée dans la même situation. Nous avons évalué cette différence entre traduction brute (ou littérale) et naturelle (ou fonctionnelle) en calculant la distance d'édition entre les deux traductions. La mesure de distance moyenne trouvée est de 21,88 (Hajlaoui 2006), sachant que les SMS ne sont pas très longs (moins de 100 caractères en moyenne). La distance moyenne trouvée est le nombre minimal de suppressions, insertions ou remplacements de lettres nécessaires pour transformer une traduction brute (ou littérale) en une traduction naturelle (ou fonctionnelle).

Nous avons montré dans un article antérieur (Hajlaoui 2006) que les traductions directes d'un corpus réel constitué de phrases naturelles ne donnent pas de résultats naturels en français. Tenant compte de ce résultat, nous avons essayé de produire un petit corpus français fonctionnellement équivalent au corpus arabe initial. Afin de développer ce corpus, nous avons adopté la technique suivante : à partir d'un ensemble de 50 SMS révisés et jugées fonctionnels, nous avons construit un ensemble plus grand en formant des combinaisons différentes des arguments utilisés (type, modèle, année, couleur, prix…). Par exemple, on remplace une année par une autre (*je cherche une voiture modèle 98*) → (*je cherche une voiture modèle 99*) ou une marque par une autre, une couleur par une autre (*A vendre BMW rouge*) → (*A vendre PEUGEOT noire*), etc.

3.3. Adaptation interne

L'extracteur de contenu de CATS est écrit avec l'outil EnCo, un LSPL[40] développé par H. Uchida dans le cadre du projet UNL (Uchida,

[40] LSPL :Langage Spécialisé pour la Programmation Linguistique.

Zhu et al. 2005-2006) pour écrire des « enconvertisseurs » vers le langage pivot UNL.

Cet outil a été utilisé par D. Daoud pour produire une représentation syntaxiquement semblable à UNL, mais qui ne correspond pas du tout à la représentation UNL (Uchida and Zhu 2003) standard, liée à une expression linguistique en anglais (même si elle en est une représentation profonde). En effet, CRL-CATS est une représentation de type `Propriété {objet, valeur}`, et pas un graphe représentant l'analyse sémantique d'un énoncé.

Le langage spécialisé EnCo et l'extracteur de contenu de l'arabe

EnCo (Uchida and Zhu 1999) attend en entrée :

- un dictionnaire et une grammaire (linguiciel).
- un texte découpé en phrases.
- Il compile le linguiciel, puis traite successivement chaque phrase. Les structures de données manipulées par EnCo sont :
- une liste de nœuds avec deux têtes de lecture/écriture placées sur deux nœuds successifs (LW « Left Windows », RW « Right Windows ») et deux têtes de lecture (LC, RC) pour les contextes gauche et droit.
- un graphe de nœuds, initialement vide, pouvant contenir des nœuds de la liste, et dont les arcs portent des « relations » identifiées par des symboles à trois caractères alphabétiques.

Au départ, la liste comporte trois nœuds : la limite gauche, le nœud courant et la limite droite. Le nœud courant contient comme chaîne la phrase à traiter.

De façon générale, un nœud peut contenir quatre éléments : une chaîne, un ensemble d'attributs « de chaîne » (initialisés lors des appels au dictionnaire), une UW (référence lexicale, venant du dictionnaire ou créée par une règle), et un ensemble d'attributs « de graphe » (préfixés par « .@ »). Les attributs sont booléens, et ne sont pas déclarés. Seul « .@entry » a un rôle spécial.

La syntaxe des règles à appliquer est la suivante (Uchida, Zhu et al. 2005-2006) :

```
<TYPE>...(<PRE2>)(<PRE1>){<LNODE>} {<RNODE>}
          (<SUF1>) (<SUF2>)… P<PRI>;
```

avec

```
<LNODE>:="{“ [<COND1>] ":" [<ACTION1>] ":"
          [<RELATION1>]":" [<ROLE1>] "}"

<RNODE>:="{“ [<COND2>] ":" [<ACTION2>] ":"
          [<RELATION2>]":" [<ROLE2>] "}"
```

Une règle peut s'appliquer si sous la fenêtre d'analyse gauche (LW) se trouve un nœud qui satisfait la condition <COND1> et sous la fenêtre d'analyse droite se trouve un nœud qui satisfait la condition <COND2>. Quand les nœuds à gauche et à droite de la fenêtre d'analyse répondent aux conditions trouvées dans <PRE> et <SUF>, les propriétés grammaticales dans la fenêtre d'analyse sont réécrites selon les actions <ACTION1> et <ACTION2>.

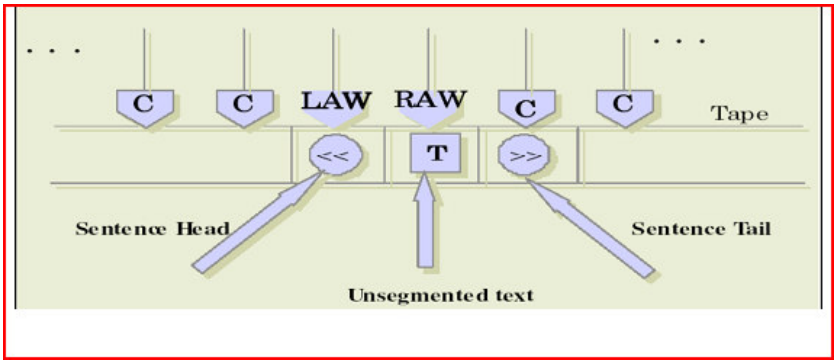


Figure 2 : Configuration initiale d'EnCo (Daoud 2006)

Voici un petit exemple. Initialement, LW contient le symbole '<<' appelé SHEAD et RW contient le premier mot de la phrase qui est « recherche »:

- SMS en entrée : *recherche voiture*
- Articles du dictionnaire utilisés :
- `[chaîne] {} "UW" (traits) <priorité>`
- `[recherche]{} "wanted" (want) <F,1,1>;`
- `[voiture] {} "saloon" (vech) <A,1,1>;`
- Plusieurs règles sont appliquées, dont la première est : `R{SHEAD:::}{wanted:::}P20;`

Cette règle fait un "shift right" désigné par `R` car sous la fenêtre gauche il y a `SHEAD`, et sous la fenêtre de droite, il y a le mot « recherche » qui est mis en correspondance avec l'UW `"wanted"` et le trait `wan` dans le dictionnaire. `P20` indique la priorité affectée à cette règle par rapport aux autres.

Le résultat final est :

```
;========== UNL =================
;Recherche voiture.
[S]
wan(saloon:0A,      wanted:00)
[/S]
;================================
```

Le dictionnaire utilisé dans la version arabe a environ 30.000 entrées, dont 20.000 ont été générées automatiquement (grâce à un répertoire de variantes et de fautes d'orthographe fréquentes). Il relie les mots et les expressions arabes des domaines de CATS (Cars, RealEstate, Jobs, Misc) aux concepts de CRL-CATS en précisant les propriétés sémantiques, syntaxiques et morphologiques utilisées dans l'analyse des SMS arabes. La structure des entrées inclut des abréviations, différentes écritures pour la même entrée, différentes formes orthographiques et d'autres formes de jargon utilisées dans le sous-langage en question.

Les 710 règles EnCo utilisées dans le système CATS extraient les informations utiles, et ne font pas l'analyse linguistique au sens classique. Elles affectent des valeurs à des objets préfinis dans le dictionnaire pour construire des relations de type `Propriété{objet, valeur}`. L'ensemble de ces relations forme la représentation CRL-CATS.

Adaptation au français de l'extracteur de contenu écrit en EnCo

Contrairement à la difficulté de trouver un corpus fonctionnel en français, la bonne surprise de ce travail a été que nous n'avons modifié que légèrement les règles fabriquées initialement pour la version arabe, et que l'EC obtenu fonctionne bien pour le sous-langage correspondant du français, celui des SMS spontanés pour l'achat et la vente de voitures d'occasion.

Cela confirme la théorie linguistique de (Kittredge and Lehrberger 1982) selon laquelle deux sous-langages équivalents dans deux langues

différentes sont proches (très proches ici) l'un de l'autre, même si leurs deux langues mères sont éloignées.

Le dictionnaire fabriqué est destiné à tout type d'utilisateurs, de tous niveaux. Il faut donc s'attendre à recevoir des erreurs de frappe, des abréviations étranges, des mots étrangers, et des fautes. Par exemple, quelqu'un écrira « Alfa Roméo » au lieu de « Alpha Roméo ». De plus, le dictionnaire doit évoluer suivant l'usage.

L'exemple ci-dessous montre qu'il faut tenir compte dans le dictionnaire du sous-langage français : un Français peut dire « *je cherche une A3*… » au lieu de dire « *je cherche une voiture AUDI A3*… » (comme on le dit de préférence en arabe). Ainsi, on doit ajouter les entrées suivantes qui doivent converger vers le même concept CRL-CATS.

```
[AUDI]{}"AUDI(country>germany,country>europe)"(make,car)
<A,3,3>;
[A3]  {}"AUDI(country>germany,country>europe)"(make,car)
<A,3,3>;
[A4]  {}"AUDI(country>germany,country>europe)"(make,car)
<A,3,3>;
[A6]  {}"AUDI(country>germany,country>europe)"(make,car)
<A,3,3>;
```

Figure 3 : Convergence de plusieurs entrées dictionnairiques vers une même entrée

Il se trouve que l'outil EnCo fait la différence entre les minuscules et les majuscules. Pour l'arabe, le problème ne s'est pas posé vu qu'il n'y a pas cette distinction. Le nombre d'entrées dans certains cas est beaucoup plus réduit en français qu'en arabe, car une seule entrée peut être écrite en arabe de plusieurs façons, avec ou sans ECHAKEL, avec ou sans ELHAMZA et avec ou sans voyelles diacritiques. Dans d'autres cas, le nombre d'entrées doit augmenter, car il faut tenir compte de la casse et des abréviations utilisées dans le sous-langage des SMS en français. Par exemple, « *cse départ* » à la place de « *cause départ* ».

Nous sommes partis d'un ensemble de 638 d'entrées de base en arabe pour le domaine *Cars,* avec un coefficient d'expansion égal à 3, dû

aux erreurs, aux formes diacritiques, et aux transcriptions multiples de noms étrangers etc., ce qui fait un total de 1914 lexèmes. Nous avons étendu les 638 entrées de base correspondantes en français à 1761 lexèmes français, pour la prise en compte d'erreurs, d'alternance masculin/féminin, singulier/pluriel, minuscule/majuscule, et de transcriptions multiples de noms étrangers. Cela donne un coefficient d'expansion de 2,7, presque égal à celui de l'arabe (3).

3.4. Adaptation externe

Il s'agit d'adapter un extracteur de contenu d'un autre système, destiné à la même langue et à un autre domaine. Nous avons ainsi adapté à CATS l'extracteur du français développé pour le projet Nespole! par H. Blanchon en Tcl/tk, en utilisant des transducteurs réguliers. Cela représente un peu moins de 32.000 lignes de code (Blanchon 2004).

La représentation de contenu obtenue est en IF (« Interchange Format »), un pivot sémantico-pragmatique utilisé pour des domaines restreints.

La figure ci-dessous montre les composants d'une représentation en IF : actes de parole, concepts et arguments.

Au début de ce travail, nous disposions du code (celui du second démonstrateur de Nespole!) et de la version papier et électronique de la spécification de l'IF (version du 18/08/2002).

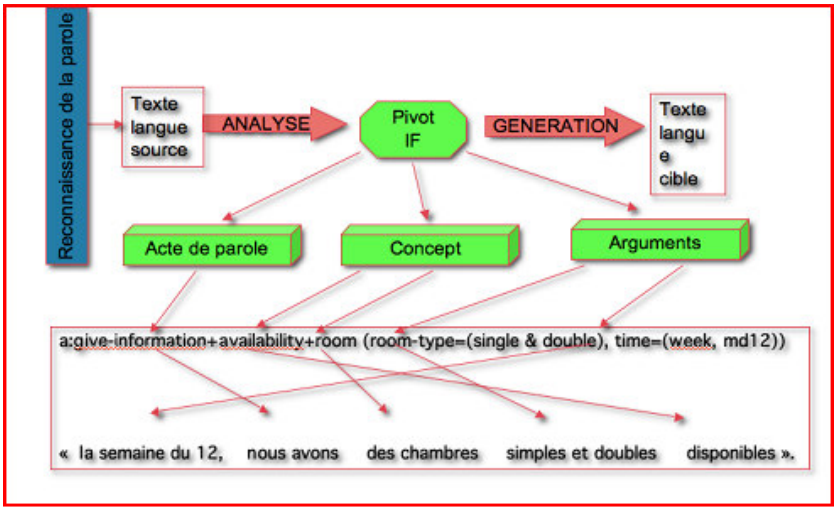


Figure 4 : extracteur de contenu pour français dans le domaine du tourisme

Méthode d'extraction de contenu dans Nespole !

Comme le montre le schéma ci-dessous, la méthode utilisée pour l'analyse du français vers l'IF est composée des étapes suivantes :

- Segmentation des SDU (Unités Sémantiques de Dialogue).
- Détection du domaine.
- Construction d'un préfixe de l'acte de dialogue et instanciation des arguments liés.
- Instanciation des arguments liés au domaine et gestion des subordinations.
- Complémentation de l'acte de dialogue

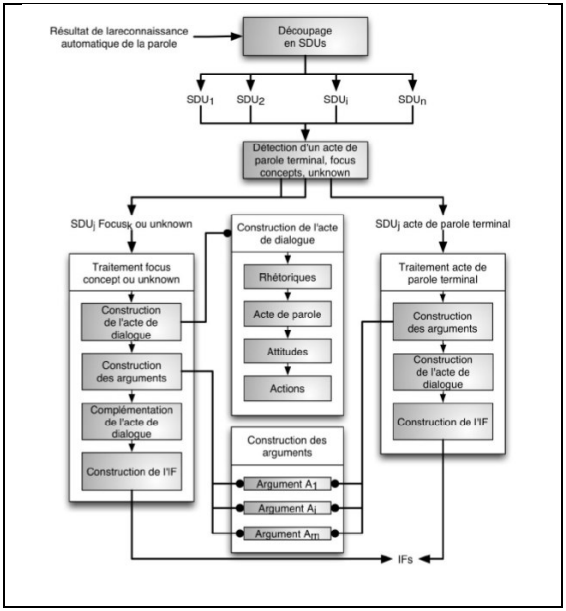


Figure 5 : Architecture du module d'analyse français vers IF du second démonstrateur NESPOLE! (Blanchon 2004).

Méthode pour transformer l'IF en IF-CATS

Nous avons adapté la spécification de l'IF au domaine de l'automobile, et l'avons enrichie en ajoutant de nouveaux arguments comme `vehicle-motor-type`, `vehicle-hand`… et de nouvelles actions, essentiellement l'action d'achat `e-buy` et l'action de vente `e-sell`. Nous avons utilisé les mêmes étapes que celles du second démonstrateur pour adapter la spécification IF au domaine de l'automobile, en essayant d'éliminer les instructions qui posent problème et/ou qui ne sont pas nécessaires, pour réduire le temps de calcul, et d'ajouter de nouvelles instructions…

Un travail important a été fait dans l'étape d'instanciation des arguments liés au domaine des véhicules (`vehicle`) : on instancie

essentiellement la spécification d'un véhicule « *vehicle-spec* », ainsi que d'autres arguments moins intéressants tels que : `theDistance`, `theLocation`, `theDuration`, `theDestination`, `theTime`, `thePrice`… La nouvelle fonction `VehicleSpec2If` permet la recherche et la construction des arguments liés au « focus concept » `vehicle` : le seul argument qui existe et qui était programmé dans le code du second démonstrateur est `frenchvehicle`, qui peut avoir comme valeur `voiture`, `ski`, `camion`, `bus`, `train`, `avion`… D'autres arguments existent dans la spécification, mais qui ne sont pas programmés tels que : `makevehicle`, `modelvehicle`, `sizevehicle`, `frenchcolor`, `agevehicle`, `pricevehicle`…

Afin d'adapter la spécification IF au domaine de l'automobile, nous avons ajouté d'autres arguments liés à ce domaine, tels que `motortypevehicle`, `handvehicle`, `conditionvehicle`… De la même façon, des fonctions `Argument2if` construisent les valeurs IF associées. La figure suivante est un exemple du résultat obtenu après adaptation.

Entrée 1 = je veux vendre une grande voiture française BM 325 4 portes diesel bleue TBE première main assurance complète avec CT sans climatisation TB prix dernier mod
Sortie1 = `{c:give-information+disposition+vehicle(disposition=(desire, who=i), action=e_sell, vehicle-spec=(car, vehicle-make=BMW, vehicle-model=325, vehicle-size=4 door, vehicle-shape=big, vehicle-motor-type=diesel, vehicle-hand=first_hand, vehicle-color=blue, vehicle-condition=good, vehicle-assurance=inssured, vehicle-controle=total_check, vehicle-air-condition=no_air_condition, vehicle-nationality=french, age-vehicle=new_mod, price-vehicle=good_price))}`

Figure 6 : extracteur de contenu pour le français pour l'occasion automobile

Nous avons appelé le résultat obtenu IF-CATS (`sortie1` dans l'exemple précédent).

Compilateur IF-CATS_CRL-CATS

Nous avons construit un compilateur qui analyse la sortie IF-CATS et la transforme dans le format CRL-CATS en utilisant un dictionnaire IF-CRL lié à cette structure qui permet la substitution des arguments. La figure suivante montre qu'en passant par cette transformation, on arrive à la même sortie que celle donnée par l'outil EnCo, à l'exception des symboles `00, 0J, 0R` ajoutés par ledit outil.

```
;======================== CRLcats ========================
; je veux vendre une Renault clio mod 1998
;{a:give-information+disposition+vehicle(disposition=(desire, who=i),
;action=e_sell, vehicle-spec=(, vehicle-make=RENAULT, vehicle-model=clio,
;vehicle-size=, vehicle-shape=, vehicle-motor-type=, vehicle-hand=,
;vehicle-color=, vehicle-condition=, vehicle-assurance=,
;vehicle-controle=, vehicle-air-condition=, vehicle-nationality=,
;vehicle-age=1998, vehicle-price=, vehicle-mileage=))}
S
sal(saloon, sale)
mak(saloon, RENAULT(country>France,country>europe))
mod(saloon, clio)
yea(saloon,  1998)
/S
;================================================
```

Figure 7 : exemple de sortie du compilateur IF-CATS

3.5. Résultats et évaluation par rapport à la version originale

Méthode d'évaluation

Nous avons traduit manuellement le corpus d'évaluation utilisé pour l'évaluation de la version arabe (originale) du système. C'est un corpus constitué de 200 SMS réels (100 SMS d'achat + 100 SMS de vente) envoyés par des utilisateurs réels en Jordanie. Nous avons mis 289 mn pour traduire les 200 SMS arabes (2082 mots équivaut à 10 mots/SMS, environ 8 pages standard[41]) de l'arabe vers une traduction

[41] Une page standard contient 250 mots.

"brute" (littérale), soit 35 mn par page. Nous avons mis 10 mn par page standard pour passer d'une traduction brute à une traduction fonctionnelle. Nous avons obtenu 200 SMS français jugés fonctionnels (1361 mots, soit 6,8 mots/SMS, environ 5 pages standard).

Pour évaluer les résultats d'extraction, nous avons calculé le rappel R, la précision P et la F-mesure F pour chacune des propriétés les plus importantes (action de vente ou d'achat, marque, modèle, année, prix) définis comme suit :

P = Nombre d'entités correctes identifiées par le système / Nombre total d'entités identifiées par le système ;

R = Nombre d'entités correctes identifiées par le système / Nombre d'entités identifiées par l'humain ;

F = 2*P*R / P+R

Résultats

Nous avons fait des évaluations pour les propriétés les plus importantes. Les pourcentages de portage par adaptation interne (par rapport à la version originale) varient entre 95% et 100%, avec une moyenne de 98 %. Les pourcentages de portage par adaptation externe (par rapport à la version originale) varient entre 46% et 99%, avec une moyenne de 77 %. Notons que ce sont les propriétés traitant les chiffres comme prix et années qui rendent faible la valeur du pourcentage du portage par adaptation externe, mais son avantage c'est qu'elle ne nécessite qu'un simple accès à la représentation interne de l'application.

	EnCoAR			EnCoFR (adaptation interne)				RegExpFR (adaptation externe)			
Propriété	Précision	Rappel	F-mesure (EnCoAR)	Précision	Rappel	F-mesure (EnCoFR)	% portage	Précision	Rappel	F-mesure (RegExpFR)	% portage
Achat/vente	0,956	0,970	0,963	0,994	0,855	0,919	95	1,000	0,835	0,910	95
Année	0,817	0,960	0,883	0,879	0,819	0,848	96	0,828	0,271	0,409	46
Prix	0,800	0,822	0,811	0,789	0,822	0,805	99	0,955	0,288	0,442	55
Marque	0,978	0,963	0,970	0,961	0,961	0,961	99	0,994	0,928	0,960	99
Modèle	0,901	0,837	0,868	0,842	0,903	0,872	100	0,965	0,661	0,785	90
Moyenne	0,890	0,910	0,899	0,893	0,872	0,881	98	0,948	0,597	0,701	77

Tableau 1 : Comparaison entre les résultats d'EC

La figure suivante permet de mieux visualiser la comparaison entre les valeurs de F-mesure trouvées pour chacune des versions du système.

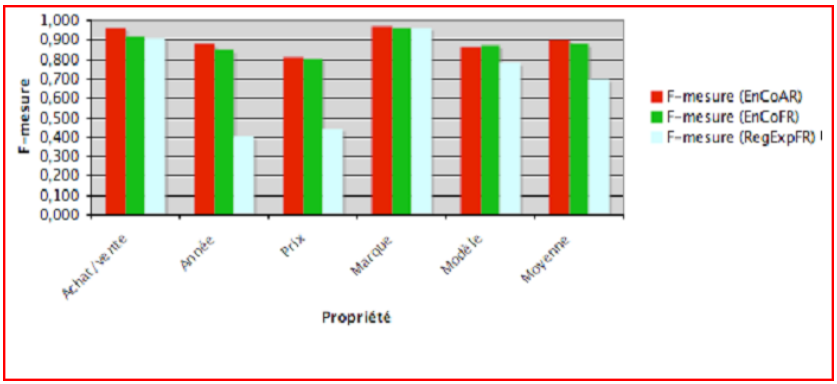


Figure 8 : comparaison entre F-mesures

4. Conclusion

Nous avons choisi CATS comme application à localiser car c'est une plate-forme qui traite des ESLN et qu'on a accès à toutes ses ressources. Nous avons présenté une première méthode de localisation « interne » qui nécessite un accès total au code source et aux ressources linguistiques de l'application. Malgré la grande distance qui existe entre le français et l'arabe, cette méthode donne de bons résultats à cause de la proximité des sous-langages. Nous avons présenté une deuxième méthode, dite « externe », qui demande un simple accès à la représentation interne de l'application, et consiste à adapter un EC existant pour la langue "cible". Enfin, nous sommes en train de faire une expérience pour déterminer si on pourrait construire un système de TA statistique, de qualité suffisante non pas pour la compréhension, mais pour l'extraction de contenu, à partir de corpus beaucoup plus petits que dans le cas général, en profitant du fait qu'il s'agit de sous-langages

restreints, et qu'on peut "injecter" des dictionnaires spécifiques relativement faciles à construire.

Bibliographie

Besacier, L., H. Blanchon, et al. (2001). *Speech Translation for French in the NESPOLE! European Project*. Eurospeech, Aalborg, Denmark.

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, UJF. HDR: 380. Grenoble, Université Joseph Fourier: 380.

Bélanger, L. (2003). *Le traitement automatisé des courriels pour les services aux investisseurs: une approche par la question-réponse*, Université de Montéréal: 48.

Califf, M. E. (1998). *Relational learning techniques for natural language information extraction.* Artificial Intelligence Laboratory, The university of Texas at Austin: A198-276.

Cardie, C. (1997). "*Empirical methods in information extraction."* AI Journal 18, 4: 65-79.

Ciravegna, F. (2001). *Adaptive information extraction from text by rule induction and generalisation*. 17th International Joint Conference on Artificial Intelligence (IJCAI), Seattle.

Daoud, D. M. (2005). "B*uilding SMS-based System using Information Extraction Technology*". ACIDCA-ICMI, Tozeur, Tunisia.

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

Hajlaoui, N. (2006). *Recherche et production de corpus de messages pour la multilinguïsation de sites de e-commerce en SMS*, initialement en arabe. 6th international IBIMA (International Business Information Association) Conference, Bonn, Allemagne.

Huberman, B., P. Pirolli, et al. (1998). S*trong Regularities in World Wide Web* Surfing. Science.

Kittredge, R. and J. Lehrberger (1982). *Sublanguage: study of language in restricted semantic domain.*

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

Lehman, A. (1996). *Construction d'un système de résumé automatique de textes de type scientifique et technique. RECITAL* (Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues), Paris.

Lepage, Y. (2006). *Traduction par analogie*. MTS, Pukhet, IAMT, ed.

Ritchie, I. A. G. D. (1995). "*Natural Language Interfaces to Databases Ð an Introduction. Natural Language Engineering.*" Cambridge University Press: 29-81.

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

Slocum, J. (1986). *How one might automatically identify and adapt to a sublanguage. Analyzing language in restricted domains.*

Uchida, H. and M. Zhu (1999). *Enconverter Specifications,* UNU/IAS UNL Center.

Uchida, H. and M. Zhu (2003). *The Universal Networking Language specification*, UNL Center UNDL Foundation.

Uchida, H., M. Zhu, et al. (2005-2006). *Universal Networking Language.*

De la variation des usages au consensus terminologique : vers un dictionnaire de l'ingénierie nucléaire

Marie Calberg-Challot

AREVA NP et UMR 7597 CNRS – Université Paris 7
Tour AREVA, Bal. 1027A, 1, place de la Coupole, 92 084 Paris La Défense Cedex
marie.calberg@areva.com
http://www.arevagroup.com

Danielle Candel

UMR 7597 CNRS – Université Paris 7
C 7034, 2, place Jussieu, 75 251 Paris Cedex 05
dcandel@linguist.jussieu.fr
http://htl.linguist.jussieu.fr

Christophe Roche

Equipe Condillac - Listic - Université de Savoie
Campus scientifique, 73 376 Le Bourget du Lac Cedex
christophe.roche@univ-savoie.fr
http://ontology.univ-savoie.fr

Résumé :

Cet article propose une méthode pour l'élaboration d'un dictionnaire de l'ingénierie nucléaire au sein de l'entreprise AREVA NP. Dans cette expérience, nous nous intéresserons plus particulièrement au sous-domaine « Réacteurs ». Après avoir présenté le secteur de travail et les raisons de l'élaboration d'un tel projet, nous décrivons les étapes suivies pour l'élaboration d'un dictionnaire. Tout en évaluant l'importance de la variation terminologique dans le choix des termes et des domaines, nous exposons le travail conjoint et complémentaire

des experts du domaine, des experts terminologues et lexicologues, et des experts en ingénierie des connaissances. Le but de l'étude étant un consensus terminologique, nous mettons en avant l'importance de croiser les démarches et la nécessité d'une approche pluridisciplinaire dans la construction d'une ressource terminologique et dictionnairique telle que le dictionnaire en construction.

1. Introduction

Nous voudrions relater une expérience lexicographique menée dans le domaine de l'ingénierie nucléaire, et nous nous intéressons plus particulièrement dans cet article aux vocabulaires scientifiques et techniques du sous-domaine « Réacteurs[42] ». Cette expérience se place au cœur de l'élaboration du dictionnaire de l'ingénierie nucléaire d'AREVA NP.

Ce travail sur l'enrichissement du vocabulaire d'un domaine de spécialité met en lumière la variation terminologique, que ce soit chez un seul expert ou chez plusieurs, et montre les difficultés à trouver un consensus terminologique. Ce consensus est tout aussi difficile à trouver en ce qui concerne la couverture des domaines d'application. Pour illustrer la façon dont on approche d'un consensus lors de la sélection de termes dans le cadre plus large de la pratique terminologique définitoire, retracer le travail réalisé en amont par les experts et le linguiste terminologue lexicologue est essentiel dans la construction d'une telle ressource.

Trois experts ont participé à diverses étapes du projet. Nous les nommerons expert E1, expert E2 et expert E3. Deux autres experts nous ont apporté des témoignages, ce sont l'expert E4 et l'expert E5. Cette étude s'appuie sur deux types de corpus : d'une part, un ensemble

[42] Dénommé dorénavant Réacteurs, sans guillemets.

de textes techniques, le corpus RCC-P[43] et d'autre part, une liste de termes relevant du sous-domaine « Réacteurs » élaborée par l'expert E1. Ce travail relate l'état du projet en février 2007.

La liste de termes élaborée par l'expert E1 a été soumise à deux experts pour plusieurs lectures successives, avant d'être proposée à un comité participant à l'élaboration du dictionnaire (le « Comité dictionnaire »), et d'être enrichie de définitions. Que sont alors devenus les termes de la liste initiale avec les relectures successives de nos experts ? Ont-ils été acceptés ou refusés ? Quelles sont les raisons qui poussent les experts à garder ou refuser un terme ? Après avoir donné lieu à une deuxième liste, on se demandera ce que sont devenus les termes au sein du Comité dictionnaire, s'ils ont été définis, si l'élaboration des définitions a mis en avant différents niveaux de spécialisation des termes qui entraîneraient à leur tour un changement dans la dénomination des domaines. Finalement, le travail sur le vocabulaire d'un domaine de spécialité et les réflexions qui y sont associées nous amèneront à tester une démarche de nature plus nettement conceptuelle.

2. Présentation de la ressource terminologique dans le domaine de l'ingénierie nucléaire

2.1. Présentation du secteur d'activité et du sous-domaine de spécialité

Le sous-domaine Réacteurs (aussi dénommé « Plants ») constitue un sous-ensemble du domaine scientifique et technique de l'ingénierie nucléaire. Il couvre l'ensemble des activités allant de la conception à la mise en service des chaudières nucléaires. Les équipes d'AREVA NP

[43] *Règles de conception et de construction applicables aux procédés.* Ce document, édité par l'AFCEN (Association française pour les règles de conception de l'énergie nucléaire), a pour objectif de définir les règles de conception et de construction des systèmes des centrales nucléaires à eau sous pression construites en France.

interviennent également sur l'amélioration des performances et de l'extension de la durée de vie des réacteurs à eau sous pression (REP) en service. Nous pourrions définir une centrale nucléaire comme l'ensemble des bâtiments et des équipements d'un site dédié à la production d'électricité comportant une ou plusieurs unités ou tranches, équipée chacune d'un réacteur nucléaire.

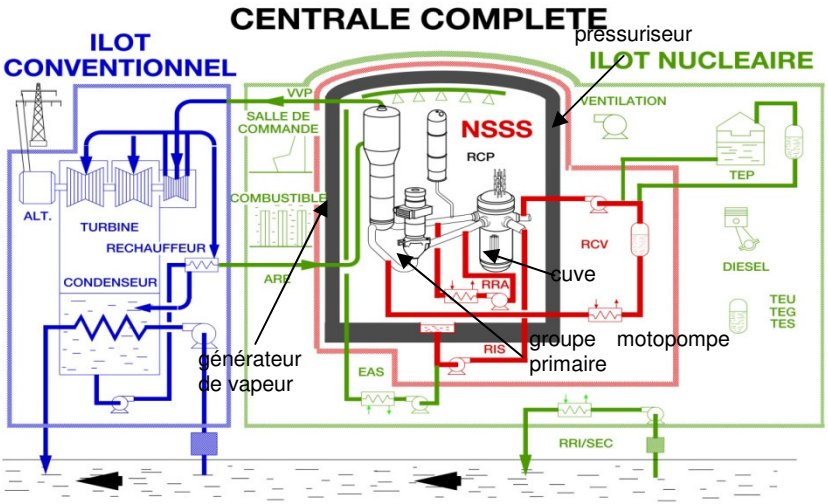


Figure 1 : Schéma d'une tranche nucléaire (Source AREVA NP)

Pour expliquer le schéma d'une tranche nucléaire telle que représentée à la figure 1, nous pourrions partir du circuit primaire principal (CPP) qui est composé de trois ou quatre boucles connectées à la cuve abritant le combustible nucléaire, chaque boucle étant composée d'un générateur de vapeur, d'un groupe motopompe primaire (GMPP) et de tuyauteries primaires assurant la liaison entre ces équipements. A une des boucles est connecté un pressuriseur, permettant le contrôle de la pression de l'ensemble du circuit.

L'îlot nucléaire est, pour l'essentiel, composé du circuit primaire principal, des « systèmes auxiliaires », des « systèmes de sauvegarde » et

des bâtiments qui les abritent, c'est-à-dire le bâtiment réacteur (BR), le bâtiment combustible (BK), les « systèmes auxiliaires », les « systèmes de sauvegarde » et les autres bâtiments associés.

Les « systèmes auxiliaires » sont le « système de contrôle volumétrique et chimique (RCV) », le « système d'appoint en eau et bore (REA) », le « système de refroidissement du réacteur à l'arrêt (RRA) », le « système de traitement des effluents primaires (TEP) », le « système de traitement des effluents liquides (TEU) », le « système de traitement des effluents gazeux (TEG) » et le « système de traitement des effluents solides (TES) ».

Les « systèmes de sauvegarde » regroupent quant à eux le « système d'alimentation de secours des générateurs de vapeur (ASG) », le « système d'injection de sécurité (RIS) », le « système d'aspersion de l'enceinte (EAS) », le « système de refroidissement intermédiaire (RRI) » et le « système d'eau brute secourue (SEC) ».

Au sein de l'ilot nucléaire, l'ensemble, « livré en kit », constitué du circuit primaire principal, des principaux équipements de quelques systèmes auxiliaires (RRA et RCV) et du système de sauvegarde (RIS), ainsi que des quelques systèmes de contrôle-commande essentiels au pilotage et à la sûreté du réacteur, forme le « Nuclear steam supply system (NSSS) ». La traduction stricte de ce terme d'origine américaine est « chaudière nucléaire » ; cependant une distinction entre les deux termes est à rappeler. Elle porte sur les limites de fourniture. Le « NSSS », selon la pratique contractuelle américaine du passé, est un ensemble d'équipements non montés, non essayés, non testés, sans tuyauteries ni câbles électriques de liaison ; dans la pratique contractuelle française, la « chaudière nucléaire » est ce même ensemble d'équipements complété de quelques autres, notamment ceux portant sur la manutention du combustible, mais avec, en plus, les tuyauteries et les câbles de liaison, le tout monté, essayé et testé.

Le schéma ci-dessous constitue une représentation de la notion de « système » de l'îlot nucléaire au sein d'AREVA NP, en cohérence avec

les témoignages des experts et la littérature du groupe à notre disposition (Coppolani et *al.*, 2004).

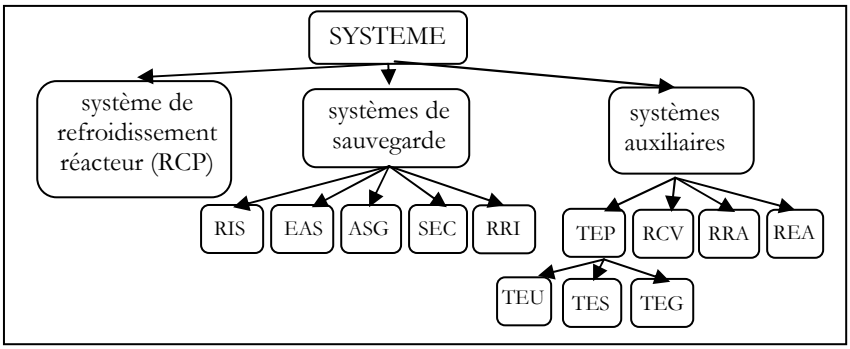


Figure 2 : Représentation de la notion de « système » de l'îlot nucléaire

Les « systèmes auxiliaires » sont des systèmes utilisés en fonctionnement normal de la tranche nucléaire, c'est-à-dire ceux requis pour la production d'électricité alors que les « systèmes de sauvegarde » sont ceux requis pour ramener ou maintenir le réacteur nucléaire dans un état sûr après un incident ou un accident ou pour en limiter leurs conséquences. Quand le réacteur nucléaire est en fonctionnement normal, les « systèmes de sauvegarde » sont en veille ou à l'arrêt. On pourrait aussi dire que les « systèmes auxiliaires » sont des systèmes n'ayant pas d'impact sur la sûreté alors que les « systèmes de sauvegarde » sont des systèmes ayant un impact sur la sûreté (témoignage des experts E4 et E5).

2.2. Les motivations d'un tel projet

A ce jour, il existe bien évidemment différents dictionnaires, lexiques, vocabulaires, répertoires ou même listes de termes dans le domaine du nucléaire[44] et il convenait donc de s'interroger sur l'intérêt

[44] Voir aussi SFEN, 2007.

d'ajouter un nouveau dictionnaire aux publications déjà à notre disposition. Il est vrai qu'il n'est pas facile de disposer de l'ensemble de ces publications qui permettraient après une recherche plus ou moins longue de recueillir une réponse satisfaisante. Consulter successivement l'ensemble des publications peut s'avérer fastidieux. Beaucoup de ces publications sont de simples lexiques, sans définitions, et rédigés en langue étrangère. Les trouver incomplètes, obsolètes, dépassées peut ajouter à l'insatisfaction.

Ces publications sont nombreuses, souvent partielles, limitées à un sous-domaine de l'ingénierie nucléaire. En outre, le domaine du nucléaire, comme toute autre activité humaine, évolue. Les dictionnaires du passé définissent des termes et expressions inutilisés aujourd'hui, et n'incluent pas les termes des nouvelles technologies. Il est vraiment apparu nécessaire de disposer d'une autre sorte de document.

L'idée s'est fait jour d'un dictionnaire complet, qui intégrerait l'ensemble des termes et expressions employés par les professionnels du domaine de l'ingénierie nucléaire d'AREVA NP, avec des définitions suffisamment explicites pour que toute personne en relation avec cette discipline, même non spécialiste, ou d'une spécialité connexe, puisse disposer d'un service complet et fiable. Ce domaine est lui-même composé des sous-domaines que sont le « Combustible », les « Réacteurs », les « Equipements » et les « Services ».

En raison des lacunes que présentent les divers ouvrages passés ou actuels, il existe de la part des ingénieurs du domaine et plus spécialement le personnel des bureaux d'étude une demande forte pour un dictionnaire de l'ingénierie nucléaire.

Il semble également essentiel d'assurer la transmission des connaissances entre les générations du nucléaire : la génération qui a participé à la réalisation du parc actuel des centrales nucléaires va se retirer dans les toutes prochaines années, il convient donc de profiter des volontés encore présentes et disponibles pour mener à bien cette tâche.

Il est judicieux d'enrichir la terminologie de ce domaine, puisque des développements ont lieu dans différentes branches de l'ingénierie nucléaire. Il est donc essentiel d'en tenir compte dans le dictionnaire.

Enfin il est essentiel de mettre à disposition de ceux travaillant dans ce domaine, au quotidien ou occasionnellement, un dictionnaire de référence validé par les acteurs du nucléaire et apportant aux lecteurs une information validée.

2.3. Les corpus de travail

Etablir le paysage dictionnairique des données existantes était le point de départ pour mener à bien notre projet. Après avoir évalué les répertoires existants (dictionnaires, lexiques, glossaires, vocabulaires), leurs formats et leurs contenus, après les avoir répertoriés, rassemblés et nettoyés (Calberg, 2003), nous avons constitué le support permettant d'élaborer le dictionnaire de l'ingénierie nucléaire pour AREVA NP.

La deuxième étape de ce travail fut de valider l'existant et de mettre à jour les données grâce à l'extraction de termes candidats à partir de corpus (Calberg-Challot *et al.*, article soumis pour publication[45]). Nous voyons ce travail comme une étape intéressante dans la construction d'une ressource terminologique avec la validation de l'existant, l'enrichissement de la nomenclature d'un domaine de spécialité et la mise en avant de la néologie du domaine d'étude.

Enfin et en parallèle de ce travail, nous avons demandé à l'expert E1 de lire le corpus *RCC-P*[46] et de relever les termes méritant une définition dans le dictionnaire de l'ingénierie nucléaire[47] et appartenant au sous-domaine « Réacteurs ». C'est plus particulièrement à ce dernier volet du corpus que nous nous intéresserons plus bas.

[45] V. Aussi, pour des éléments de méthode, Delavigne 2001.

[46] *Règles de conception et de construction applicables aux procédés.*

[47] Dorénavant « Dictionnaire ».

3. Travail sur la ressource terminologique

3.1. Interventions des experts E2 et E3

Il a été demandé à l'expert E1 de lire les *Règles de conception et de construction applicables aux procédés* (RCC-P) et de relever les termes du domaine Réacteurs qui lui semblaient mériter une définition dans le cadre du Dictionnaire de l'ingénierie nucléaire d'AREVA NP en construction. Ce travail réalisé, l'expert a proposé une liste de 408 termes, qui a ensuite été elle-même soumise à deux autres experts. L'expert E2 l'a relue en juin 2005 et l'expert E3 est intervenu en décembre 2005. Les lectures par les experts E2 et E3 ont été successives et indépendantes l'une de l'autre. Après l'intervention des deux experts, la liste a été soumise au Comité dictionnaire.

Les 408 termes proposés par l'expert E1 ont suscité 138 interventions de la part de l'expert E2 : 75 termes ont été proposés pour être supprimés, sans explication, 13 termes ont reçu une marque de domaine (12 fois « Contrôle-commande » et une fois « Combustible ») et huit termes apparaissent inconnus de l'expert E2. Certains termes sont donc pour lui trop spécialisés, d'autres au contraire appellent une spécialisation avec l'ajout d'une marque de domaine. Enfin il propose de regrouper plusieurs termes sous une même entrée, et « déspécialiserait » ou « déterminologiserait » alors les termes en leur attribuant une définition commune et non spécifique (Calberg & Candel, 2005).

L'expert E3, quant à lui, a effectué 66 interventions sur les 408 termes proposés par l'expert E1. En dehors de neuf termes supprimés sans explication, l'expert E3 a effectué trois types d'interventions, portant respectivement, et en nombre égal, sur la trop grande généralité de termes, sur la grande spécialisation d'autres termes, et enfin, sur le fait qu'il s'agit de termes extérieurs aux secteurs des métiers d'AREVA NP.

Les remarques ou interventions des experts E2 et E3 ne concernent pas les mêmes termes mais pour chacun d'entre eux, c'est 50% des termes qui restent non commentés. Ce silence peut être

considéré comme une mise en valeur des termes retenus pour définition et ceci constitue un point commun entre nos experts.

On notera que depuis ces interventions, cet ensemble de termes est effectivement entré dans le Dictionnaire en construction. Mais cette expérience souligne surtout la variation terminologique chez trois experts et met en exergue, à ce point du développement du projet, la difficulté de trouver un consensus terminologique.

3.2. Intervention du Comité dictionnaire

La liste élaborée par l'expert E1 (liste des 408 termes) a été soumise au Comité dictionnaire : ces termes sont-ils alors maintenus et définis dans le Comité ou bien, au contraire, supprimés ?

Le Comité dictionnaire est composé de onze experts[48] et d'une linguiste terminologue lexicologue. Les experts travaillant sur ce projet couvrent, par leurs compétences, divers domaines. Les réunions mensuelles sont consacrées au travail sur les définitions proposées par les experts sur des termes préalablement sélectionnés en commun. Ces réunions sont le lieu idéal pour observer la mise en place d'une terminologie, son évolution et la néologie (Sablayrolles 2003, Candel & Tombeux, à paraître) - même si ce pan du travail n'est pas abordé dans la présente étude. Ainsi relevons-nous quatre statuts distincts pour les termes présents dans le Dictionnaire. Le premier statut représente les termes validés et définis dans le Dictionnaire, le deuxième, les termes présents mais non définis, le troisième, les termes non présents dans le Dictionnaire, et le quatrième, les termes proposés pour suppression.

Par besoin de précision et pour définir le plus justement les termes sélectionnés et par là-même leur conférer un niveau de spécialisation supplémentaire, le sous-domaine Réacteurs a été

[48] Nous entendons par « expert » toute personne ayant une expérience importante dans le domaine du nucléaire et ne possédant pas forcément le titre d'expert au sein du groupe AREVA.

décomposé en nouveaux sous-domaines : « Procédés », « Mécanique », « Contrôle-Commande ». Ces nouveaux sous-domaines ne sont pas exhaustifs et se sont créés en fonction des compétences des experts. D'autres sous-domaines, comme « Thermohydraulique », « Neutronique », « Science des matériaux », appellent de nouveaux experts et donc de nouveaux termes à définir. Un sous-domaine transverse, « Sûreté », couvrant les différents secteurs d'AREVA NP, serait également à mettre en place.

Sur les 408 termes proposés par l'expert E1, 128 termes ne se trouvent pas dans les entrées du Dictionnaire de l'ingénierie nucléaire en février 2007. Ce fait, à cette date, montre que 31% des termes de l'expert E1 n'ont pas été retenus. Sur ces 128 termes non présents dans le Dictionnaire, on note que pour plus de la moitié, l'expert E2 ou l'expert E3 avaient proposé les termes pour suppression. Malgré tout, rien n'est définitif puisque le Dictionnaire se construit en même temps que nous en parlons.

Sur les 280 termes de l'expert E1 restants et présents dans le Dictionnaire, 22 changent complètement de domaine en allant au « Combustible », soit environ 9%. Cette expérience met encore en lumière la variation terminologique d'un expert à l'autre.

3.3. Deuxième intervention de l'expert E2

Une deuxième intervention par l'expert E2 a été effectuée en septembre 2006. Cette liste comportait alors les remarques de l'expert E3 et un retour sur les termes qui étaient déjà passés en Comité dictionnaire. Les commentaires de l'expert E2 lors de sa première intervention n'étaient pas présents. L'expert E2 est intervenu autant de fois que lors de sa première intervention. Pour 44% des interventions, les commentaires concernent les 75 termes qui avaient été supprimés sans motivation en première relecture et nous trouvons alors les motifs de la suppression.

Le reste des interventions peut être regroupé en trois sous-ensembles. L'expert considère les termes comme trop généraux pour 21% de ces interventions, comme ne correspondant pas au domaine d'AREVA NP pour 17% de ses interventions et comme trop spécialisés pour 11% des termes commentés. Ces trois ensembles d'interventions montrent la complexité de positionner le niveau de spécialisation adéquat à la réalisation de notre travail.

Enfin l'expert, qui avait initialement supprimé certains termes, revient sur sa décision dans 7% des cas et propose de les définir ultérieurement.

Cette dernière expérience met cette fois en lumière la variation terminologique chez un même expert et démontre une nouvelle fois la difficulté pour arriver au consensus terminologique.

3.4. Vers un consensus terminologique

Un projet, si bien défini soit-il au départ, évolue avec l'activité des experts. Les différentes étapes de ce travail montrent que l'aboutissement au consensus terminologique est un travail long et qui demande l'intervention de nombreux experts dans le choix des termes, dans leur classification par domaines et dans l'élaboration même des définitions. Par ailleurs, ce consensus est toujours relatif à un moment donné car la variation est toujours possible. Les éléments donnant une légitimité à notre travail sont principalement le travail en équipe et l'expertise des membres du Comité dictionnaire, l'expertise terminologique (Humbley 2000) et lexicologique permettant, quant à elle, les supports méthodologiques et l'analyse critique des résultats.

4. Apports mutuels de la terminologie et de l'ingénierie des connaissances

4.1. Analyse des résultats du Comité dictionnaire

Nous allons maintenant analyser quelques articles du Dictionnaire et voir si les termes sont en adéquation avec la représentation de la tranche nucléaire telle que nous l'avons jusqu'à ce point du développement.

De la liste de départ contenant les 408 termes proposés par l'expert E1, nous avons retenus 34 termes relatifs à la notion de « système ». Les articles retenus, dont certains sont en cours d'élaboration et requièrent encore des définitions, inspirent plusieurs remarques. En ce qui concerne les « systèmes de sauvegarde », tous les systèmes de sauvegarde de la figure 2 (RIS, EAS, ASG, SEC, RRI) sont représentés dans le Dictionnaire en cours et ont pour définisseur « système de sauvegarde », sauf le « système d'alimentation de secours des générateurs de vapeur (ASG) », terme non encore défini mais qui est déjà entré dans le Dictionnaire. Le cas des « systèmes auxiliaires », demeure, quant à lui, plus complexe. Il est aussi à noter que le terme de « système auxiliaire » n'est pas forcément un terme bien choisi du fait que ces systèmes sont en fait des systèmes utilisés en fonctionnement normal (témoignage de l'expert E4). L'article « système de traitement des effluents » et le « RCV » ont pour définisseur « système » et le « RRA », « système de sûreté ». Voici donc le schéma que nous pourrions tracer de la notion de « système » à partir des définitions du dictionnaire.

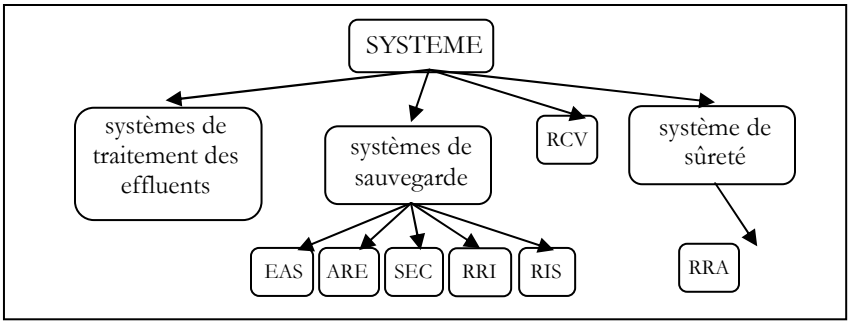


Figure 3 : Représentation de la notion de "système"
(Etat du Dict. AREVA NP en fév. 2007)

Enfin, lorsqu'on s'intéresse à l'ensemble des articles traitant des effluents, sept entrées y sont consacrées : « système de recueil d'effluents (RPE) », « système de traitement des effluents », « système de traitement des effluents gazeux (TEG) », « système de traitement des effluents liquides (TEU) », « système de traitement des effluents primaires (TEP) », « système de traitement des effluents solides » et « système de rejet des effluents liquides (TER, KER, SEK) ». Si l'on considère l'article « système de traitement des effluents » qui est en cours d'élaboration et qui renvoie vers l'ensemble des six autres entrées, on note que l'on distingue « cinq types de systèmes : systèmes de recueil d'effluents (RPE), systèmes de traitement des effluents liquides primaires et usés (TEP, TEU), systèmes de traitement et rejet des effluents gazeux (TEG), système de traitement des effluents solides (TES), et les systèmes de rejet des effluents liquides (TER, KER, SEK)». Or, le corps de la définition explique que le système « a pour rôle de recueillir, de traiter puis de recycler ou de rejeter les effluents… ». Deux représentations sont alors possibles, l'une en fonction du type de système (« systèmes de recueil », « systèmes de traitement », « systèmes de rejet »), l'autre en fonction de l'action (« recueillir », « traiter », « recycler », « rejeter »), toutes deux également utilisables.

Des représentations sont utiles pour mettre en avant le manque de clarté des définitions et les lacunes dues aux termes manquants dans le dictionnaire (témoignage de l'Expert E5). Ces représentations constituent un apport méthodologique et gagnent à être poursuivies afin de progresser dans l'élaboration du dictionnaire.

4.2. De la définition du mot à la représentation du concept de « système »

Etudier, pour un dictionnaire de langue générale construit à partir d'un grand corpus textuel, une unité lexicale telle que « système »[49], donne naturellement lieu à la rencontre d'une vaste gamme de sens et emplois généraux et spécialisés, d'emplois concrets et abstraits. La lecture de l'article du *TLF* permet de distinguer deux catégories de sens et emplois : abstraits comme « Construction de l'esprit, ensemble de propositions (…) qui forment un corps de doctrine », « méthode… », ou concrets et physiques comme « systèmes mécaniques » ou « système cristallin ».

Comment, parmi l'ensemble des expressions langagières liées à une pratique, distinguer ce qui relève du terme par rapport à ce qui relève de la variation terminologique, du trope ou du mot d'usage ? Se référer à la conceptualisation du domaine peut être une autre manière d'apporter des éléments de réponse (Roche, 2007).

Réseau conceptuel

La modélisation d'un domaine consiste à identifier les concepts décrivant les objets du *monde réel* (ici les types de systèmes) et à les structurer selon différentes relations, qu'elles soient générales comme celles de *généralisation-spécialisation* (« est-un »), d'*instance* (« est un exemple »), de *composition* (« est composé de ») ou spécifiques au domaine

[49] L'un des auteurs a eu l'occasion de s'y employer, pour le *Trésor de la langue française.*

d'application (« sur », « pour »). Le résultat est un réseau de concepts[50] (figure 4).

« Ensemble » et « concept »

Si une telle modélisation est une aide précieuse à la structuration des significations, elle reste imprécise et porteuse d'ambiguïtés. Ainsi, quels rapports existe-t-il entre les différents systèmes de refroidissement ? Les liens qu'entretiennent entre eux les différents types de systèmes de sauvegarde et ceux qui lient les différents systèmes auxiliaires sont-ils de même nature ? Est-il possible de définir plus précisément les systèmes de traitement des effluents primaires TEC, TEU, etc. ? Autant de questions qui requièrent des principes épistémologiques plus structurants.

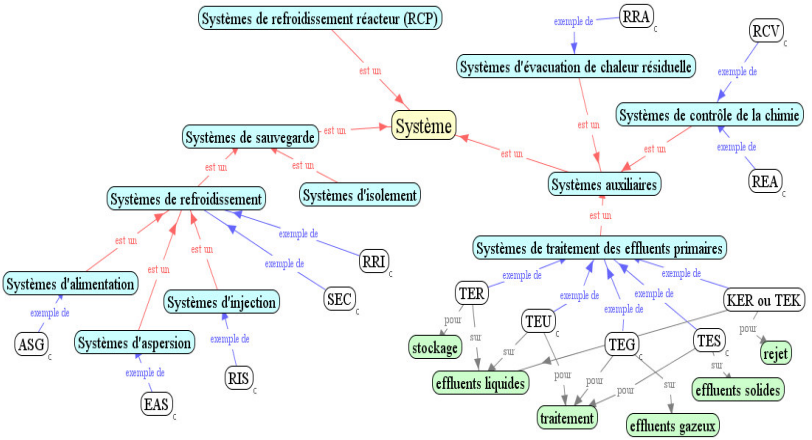


Figure 4 : Réseau conceptuel de « système »

[50] Réseau conceptuel réalisé à l'aide de SNCW (Semantic Network Craft Workbench), éditeur de schémas (concepts définis par un ensemble d'attributs) et de relations (Ontologos corp.).

La notion de *système*, au sens de *sustêma* (assemblage, composition), permet de regrouper sous une même appellation des entités qui ont en commun le fait d'être structurées, d'être composées d'éléments. Cela ne signifie pas nécessairement que ces entités soient de même nature et donc comparables entre elles. Prise dans ce sens, la notion de *système* correspond davantage à un « ensemble » – si l'on considère qu'un « ensemble » regroupe des entités pouvant être de nature différente –. C'est le cas de « système auxiliaire » alors que « système de sauvegarde » correspond davantage à un « concept » – si l'on considère qu'un « concept » regroupe des objets de même nature – qui se spécialise en différents types. La définition par *différenciation spécifique*[51] permet à la fois de distinguer ce qui relève de la notion d'ensemble ou de concept et de définir, tout en les différenciant, les divers concepts (figure 5).

[51] Définition en genre-espèce. Les exemples sont construits à l'aide d'OCW (Ontology Craft Workbench), éditeur d'ontologies par différenciation spécifique (Ontologos corp.).

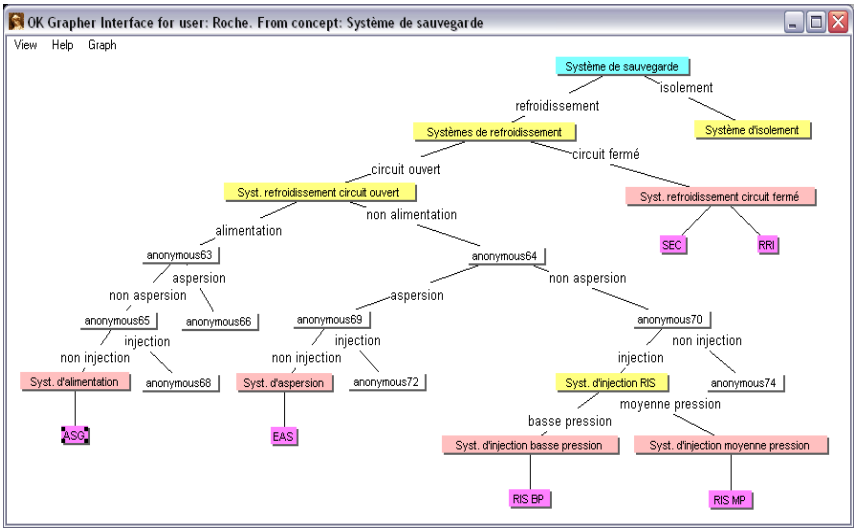


Figure 5 : Représentation conceptuelle de « système de sauvegarde »

Notons que cette figure, comme les trois suivantes, propose des cases vides, signalées au moyen des étiquettes « anonymous » : ces dernières révèlent des cases terminologiques elles-mêmes virtuelles. Les résultats des pratiques de dénomination, de néonymie et de définition analysées chez nos experts, une fois traités, permettraient donc de repérer des étapes de la représentation conceptuelle auxquelles pourrait correspondre une vacance voire un besoin de dénomination.

« Concept simple » et « concept composé »

Il existe différents types de « systèmes de traitement des effluents primaires » qui se différencient, pour les uns, selon l'action effectuée (traitement, stockage, rejet) et, pour les autres, selon l'état des effluents (gazeux, liquide, solide). Il est alors possible de les définir par différenciation spécifique au sein d'une même catégorie – ensemble de

concepts sémantiquement liés – (figure 6). Mais il est également possible, dans la mesure où l'action à effectuer et l'état des effluents sont indépendants, de définir les différents systèmes de traitement comme autant de « concepts composés » définis à partir de la catégorie des actions (figure 7) et de la catégorie des états (figure 8).

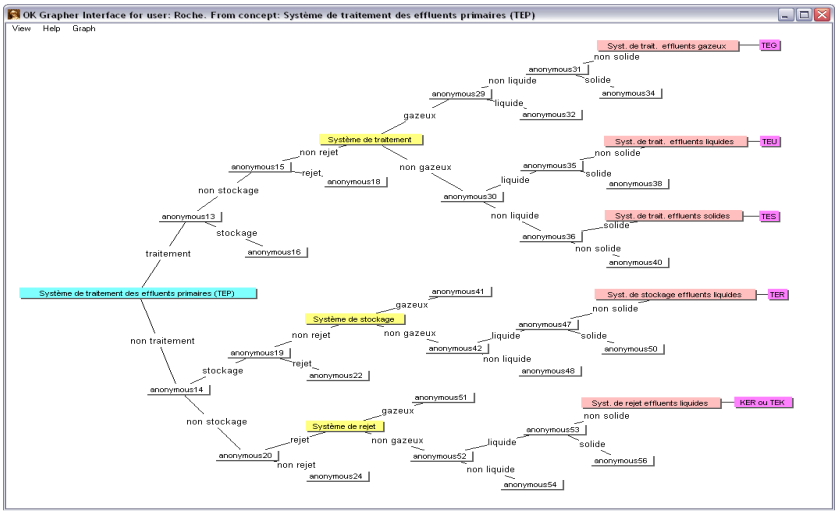


Figure 6 : Catégorie des systèmes de traitement des effluents primaires

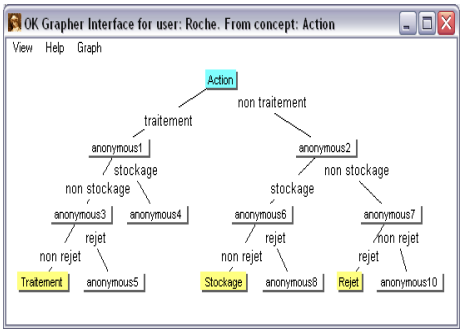


Figure 7 : Catégorie des actions.

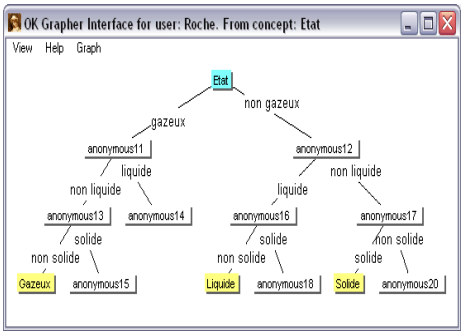


Figure 8 : Catégorie des états.

Ainsi, un "Système de traitement des effluents gazeux" (figure 6) est le résultat de la composition du "concept" "Traitement" de la catégorie "Action", et du "concept" "Gazeux" de la catégorie "Etat" (figures 7 et 8) ; l'opérateur de composition permet de définir un nouveau concept à partir de la définition, c'est-à-dire la liste des "différences" entre plusieurs concepts préalablement définis.

Les définitions obtenues grâce à une somme d'expertises – collaboration entre experts des domaines de l'ingénierie nucléaire, de la terminologie et de la lexicographie – ont donné lieu à une réappropriation des résultats et à leur transformation par une expertise en ingénierie des connaissances[52]. Les résultats tendent à donner une image de ce que les experts cherchent à exprimer et de tels résultats pourraient être réutilisés pour l'accès aux connaissances dans les bases documentaires. Sans doute l'expérience a-t-elle aussi permis de révéler des cases conceptuelles (encore) vacantes en étiquettes terminologiques (figures 5 à 8). Nous avons suivi un ensemble de démarches de nature onomasiologique (de quels domaines se compose le secteur des Réacteurs, quels sont les concepts qui composent le sous-domaine des Réacteurs ?) puis sémasiologique (quelles sont les définitions de tels

[52] Pour ce dernier aspect, voir Després 2001.

termes dans tels contextes ?) pour revenir confronter un ensemble de résultats avec une construction de nature plus conceptuelle.

5. Conclusions et perspectives

Le choix des termes, leurs figements, les définitions qui leur sont associées et leur affectation dans leurs domaines n'aboutissent qu'après l'obtention du consensus par plusieurs experts (Candel & Calberg-Challot 2007). Mais, se situant au cœur de l'élaboration du projet, on se rend compte du caractère temporaire du consensus trouvé. La variation continuera de fluctuer au cours du projet en fonction de divers paramètres et le consensus est donc tout relatif. Il faudra ainsi réfléchir à la mise en place d'un processus permettant une validation plus définitive.

Cette étude montre l'importance de croiser les méthodes et les approches pour être au plus près des experts et de leurs attentes. Il nous semble nécessaire de montrer l'apport de la terminologie en tant que travail préliminaire et indispensable pour l'élaboration de représentations conceptuelles des réalités étudiées. Cela, en vue de réutilisations transdisciplinaires, et par des usagers plus ou moins spécialistes des notions étudiées et de leurs dénominations. En contrepartie, la représentation conceptuelle d'une réalité permet, au travers d'une approche différente, un retour sur le travail terminologique en le validant ou en mettant en avant les incohérences ou imprécisions dans les définitions. En effet, toute représentation conceptuelle impose de découper le champ de travail en questions élémentaires et de formaliser des savoirs jusqu'ici non explicites.

Deux démarches se superposent et se croisent tout au long de cet article, une démarche sémasiologique et une démarche onomasiologique. Pourquoi n'a-t-on pas la même représentation d'une même réalité pour les deux approches ? Quelle est la bonne représentation ? A ces questions, seule une réponse pluridisciplinaire, confrontant les approches de manière itérative tendant vers un consensus, peut être jugée satisfaisante.

Il nous importait enfin de comprendre, au-delà des besoins terminologiques déjà décrits, quelles sont les applications exploitant les terminologies d'un point de vue industriel. Celles-ci, à n'en pas douter, sont tournées vers la gestion des connaissances pour laquelle l'élaboration, grâce à un travail préliminaire de terminologie, d'une représentation conceptuelle satisfaisante pour un grand nombre, permet une représentation intuitive de la réalité pour le public concerné et ainsi une meilleure valorisation des savoirs de l'entreprise et une meilleure valorisation du travail terminologique réalisé.

6. Remerciements

Nous tenons à remercier les experts consultés pour ce travail ainsi que toutes les personnes ayant contribué à ce projet par l'apport de documents ou d'information, et tout particulièrement Claude Cauquelin, Alain De Tonnac, Xavier Dumont, Jacques Joseph, Yves Meyzaud, Jean Oullion, Emilio Raimondo, Philippe Revel, Philippe Rouiller et Francis Tétreau. Philippe Revel nous a par ailleurs exprimé de précieux commentaires sur cette étude et nous l'en remercions.

Bibliographie

M. Calberg « Recherche méthodologique pour l'élaboration d'un dictionnaire de l'ingénierie nucléaire » mémoire pour l'obtention du DESS Ingénierie multilingue, CRIM-INaLCO, 2003

M. Calberg, D. Candel « *Pratiques terminologiques: un exemple dans le domaine du nucléaire (résultat d'enquête)* » EA 1483, Séminaire de l'Ecole doctorale, Marie-Annick Morel, Université Paris 3, 14 mars 2005

M. Calberg-Challot, D. Candel, D. Bourigault, X. Dumont, J. Humbley et J. Joseph « *Une analyse méthodique pour l'extraction terminologique dans le domaine du nucléaire* », *Terminology*, à paraître, 2008

D. Candel, M. Calberg-Challot « *Regards sur la néologie en marche, motivations et réactions. Comment la terminologie du nucléaire évolue* », in Juhani Härmä, Eva

Havu, Mervi Helkkula, Meri Larjavaara, Mari Lehtinen et Ulla Tuomarla (eds), *Actes du XXIX^e Colloque international de Linguistique Fonctionnelle*, Helsinki 2005 (21-24 septembre 2005), Publications du Département des langues romanes de l'Université de Helsinki, 18, pp. 45-49, 2007

D. Candel, V. Tombeux « Aspects de la néologie de spécialité en lexicographie générale : à propos du nucléaire dans le *Trésor de la langue française*», in Jean-François Sablayrolles (éd.), *La néologie et les dictionnaires*, Champion, à paraître

P. Coppolani, N. Hassenboehler, J. Joseph, J.-F. Petetrot, J.-P. Py & J.-S Zampa *La chaudière des réacteurs à eau sous pression*, Coll. Génie atomique, EDP Sciences, 2004

V. Delavigne *Les mots du nucléaire, Contribution socioterminologique à une analyse des discours de vulgarisation*, Thèse de Doctorat, Université de Rouen, 2001

S. Després *Une comparaison raisonnée des apports de la terminologie et de l'intelligence artificielle pour servir et améliorer la construction d'ontologies*, TIA-2001, *inist*, Nancy, 2001

SFEN (Société française d'énergie nucléaire) *Vocabulaire de l'ingénierie nucléaire*, Paris, SFEN, 2007

J. Humbley "*La terminologie*", In G. Antoine & B. Cerquiglini (éds.), *Histoire de la langue française 1945-2000,* Paris, Editions du CNRS, 315-338, 2000

C. Roche « *Dire n'est pas concevoir* », IC 2007 : 18e Journées Francophones d'Ingénierie des Connaissances, Grenoble 2-6 juillet 2007

J.-F. Sablayrolles « Le sentiment néologique », dans J.-F. Sablayrolles (éd.), *L'innovation lexicale*, Lexica, Mots et dictionnaires 11, Champion, 279-295, 2003

Trésor de la langue française (*TLF*), Centre National de la Recherche Scientifique, Paris, CNRS Editions, 16 vol., 1971-1994

Peut-on faire confiance aux outils de terminologie ?

L'évaluation entre un souci de normalisation et une complexité de modélisation -

Ismaïl Timimi

Laboratoire Geriico[53],
Université Charles de Gaulle, Lille3
BP 60 149 - 59 653 Villeneuve d'Ascq cedex - France
ismail.timimi@univ-lille3.fr

Résumé :

Cet article s'inscrit dans les travaux de l'appel à projets Technolangue sur l'évaluation des outils de traitement automatique des corpus écrits et oraux. Nous nous intéressons dans une première partie à une réflexion sur la pratique de l'évaluation en tant que méthodologie d'observation fortement liée aux processus d'innovation socio-technologique. On tentera ensuite de déterminer les critères d'observation et d'appréciation d'outils de terminologie, mais aussi, comment ces critères ne peuvent être déterminants que pour un besoin informationnel et un usage particulier. Notre questionnement final nous permettra de refocaliser le débat sur la méta-évaluation et la complexité de l'évaluation dans le contexte des outils de terminologie.

[53] Groupe d'Etude et de Recherche Interdisciplinaire en Information et Communication.

1. Les sous-entendus d'une évaluation

A l'opposé des autres méthodologies d'observation comme l'audit, le questionnaire et les sondages…, l'évaluation, dans le cas des systèmes de traitement de corpus, reste un exercice multidisciplinaire assez préoccupant. Le manque de normalisation ajouté à la difficulté de modélisation rend l'évaluation comme forme de jugement très problématique…

1.1. Le relativisme de l'évaluation

Si l'évaluation est un exercice d'appréciation des performances d'un système assorti d'un indice de satisfaction, cet indice ne peut être calculé que si l'observation est effectuée par rapport à des besoins et des référentiels de type intrinsèque ou extrinsèque. La diversité des besoins et des référentiels génère plusieurs types et degrés d'évaluation.

Dans une évaluation dite *de progression (verticale)*, le système est comparé à ses versions antérieures pour une tâche déterminée, en vue d'une étude diachronique de ses performances. C'est une démarche très courante dans les activités de conception et de développement de systèmes.

Une seconde démarche d'évaluation, dite *d'appariement (transversale)*, consiste à comparer les performance d'un système par rapport soit à d'autres systèmes conçus pour des applications similaires, soit à des résultats (besoins) prédéfinis, établis manuellement ou autrement, et surtout validés.

Dans une autre optique d'évaluation, dite *de diagnostic*, l'évaluateur expert cherche à déterminer à partir d'une série de tests les sources de performance ou non d'un système conçu pour une tâche précise. Ce mode est, lui aussi, orienté conception dans la mesure où ces tests de diagnostic permettent de développer par progression les performances d'un système.

Quel que soit le mode adopté, l'évaluation de systèmes de traitement automatique de corpus, ne peut être exercée que dans une démarche comparative... par rapport à d'autres systèmes, à des référentiels préétablis, à une application bien déterminée.

1.2. Les dichotomies de l'évaluation

Dans le cas particulier des outils d'ingénierie linguistique, on peut relever deux démarches dans la méthodologie d'évaluation, qui ne sont pas forcément exclusives. Il s'agit de l'évaluation avec *interface statique* par opposition à *l'interface dynamique* d'une part ; et de l'évaluation de type *boîte noire* par opposition à la *boîte transparente* d'autre part.

L'évaluation d'un système avec *interface statique* consiste à juger ses performances, sans faire appel à des interventions ou à des enrichissements extérieurs. A l'inverse, l'évaluation avec *interface dynamique* permet de calculer l'amélioration des performances d'un système suite à une intégration de ressources extérieures (par enrichissement terminologique par exemple). (Chaudiron, 2001)

Parallèlement, l'activité évaluative dans sa forme classique peut être menée sur le concept de la *boîte noire*. Elle porte sur le jugement des performances globales du système à partir seulement des ressources fournies en entrée (Input) et des résultats produits en sortie (Output), sans examiner le traitement intermédiaire des données effectué par les divers modules du système. A l'opposé, l'évaluation orientée *boîte transparente* s'intéresse à l'étude du fonctionnement interne du système à travers ses différents modules et prétraitements. Elle rejoint dans cet aspect *l'évaluation de diagnostic* précitée.

2. La campagne Cesart : un terrain d'expérimentation

La campagne Cesart s'inscrit dans le cadre du Projet Technolangue Chapitre Evalda, co-organisée par le laboratoire Cersates de l'université Lille3 (devenu Geriico) et Elda. Elle consiste à élaborer un

protocole *normalisé* pour l'évaluation de systèmes d'acquisition de ressources terminologiques.

2.1. Le modèle multidimensionnel de l'évaluation en Cesart

Les différentes formes d'évaluation (citées brièvement ci-dessus) sont combinables et adaptables suivant les contextes et les enjeux de la campagne d'évaluation d'une part et la typologie des systèmes participant d'autre part. Dans le cadre de notre projet Cesart, nous avons abandonné certaines formes d'évaluation telles que *l'évaluation de progression et de diagnostic,* ainsi que *l'évaluation à la boîte transparente.*

L*'évaluation de progression* a été abandonnée dans la mesure où nous ne pourrions disposer de tous les logiciels, et encore moins de leurs versions antérieures pour pouvoir étudier l'évolution diachronique de leurs performances. Nous avons abandonné également l'évaluation sur le principe de la *boîte transparente*, car il s'agit d'un mécanisme difficile à mettre en place, il requiert une connaissance des processus internes et des fondements théoriques de chacun des systèmes participant au projet. Il réclame l'accès à l'architecture et à la stratégie du système, ce qui risque d'être compromettant lorsque l'évaluateur est un intervenant extérieur (Cavazza, 1993). Pour les mêmes raisons, il ne nous était pas possible d'adopter *l'évaluation de diagnostic*, qui est aussi orientée conception et proche dès lors de l'évaluation de type *boîte transparente*.

Nous avons adopté dans le projet Cesart, le principe de l'évaluation *boîte noire.* Ce choix est justifié du fait qu'il s'agit d'une démarche d'expertise facile à mettre en œuvre, dans un consortium composé d'universitaires et d'industriels et pose le moins de problèmes méthodologiques. Sans nécessiter l'accès au fonctionnement interne des systèmes, elle permet une étude comparative malgré la différence des architectures employées. (Cavazza, 1993) (Sparck, 1996)

Afin de combler les limites de l'évaluation *boîte noire*[54], nous avons renforcé le protocole de nouveaux critères d'appréciation et consignes de jugement, dont une partie est inspirée des autres formes d'évaluation.

- Nous avons opté pour une évaluation d'appariement, grâce à des métriques quantitatives calculées à partir d'un algorithme d'appariement des outils à des référentiels terminologiques préétablis. L'évaluation qualitative est aussi envisagée grâce à des classifications de pertinence proposées par des experts humains, sur des critères fixés préalablement ;

- L'évaluation est en adéquation à des contextes prédéterminés tenant compte des besoins de l'utilisateur (*interface dynamique*) et des domaines d'applications (usages) ;

- Enfin, pour étudier l'extensibilité des systèmes, leur réserve de performance et la possibilité de leurs maintenances (Cavazza, 1993), une partie du principe de l'évaluation avec *interface dynamique* est introduite dans le projet Cesart. Un questionnaire des prétraitements (sous forme de tableau de bord) est pris en considération dans le calcul par les experts des coûts[55] de l'usage prévu. Ce coût dépend des besoins en ressources d'enrichissement internes et externes, des performances requises, du nombre et de nature d'interventions de l'utilisateur du système et du temps de traitement.

2.2. Un protocole orienté applications

Les systèmes dont il est question dans ce projet sont issus du milieu universitaire et industriel. Ils proposent des niveaux de traitements et des applications assez variés dans lesquels les termes occupent une place centrale. Leur point commun est donc d'être fondé sur le traitement des termes et des connaissances terminologiques.

---

[54] Par exemple, une des limites de l'évaluation *boîte noire* est de ne pas prendre en compte les choix et les renseignements apportés au système par son utilisateur dans les étapes préliminaires.

[55] En conformité avec la norme ISO 9126 (King, 1996).

Au delà des différences dans leurs modèles théoriques et leurs architectures, nous avons réparti les systèmes en trois catégories [56] (Extracteurs de termes, de relations morpho-syntaxiques et de relations sémantiques) qui sont davantage liées aux types de tâches pour lesquelles un protocole d'évaluation en concertation pouvait être mis en place.

Certains systèmes extracteurs de relations produisent également des « classes de termes » reliées. L'évaluation consisterait dans ce cas à l'observation de la cohésion ou non des classes produites et leur adéquation à l'application préétablie.

Si dans certains projets d'évaluation, les outils participants ont des applications clairement identifiées (résumé, traduction, question-réponse...), et sur lesquelles porte l'activité d'évaluation proprement dite, les outils participant au projet Cesart présentent, de par leur fonctionnement, une particularité rendant l'évaluation partielle. Les systèmes ne sont observés et examinés que dans une phase prématurée de leur fonctionnement qu'est l'extraction (de termes, de relations ou de classes). Les applications finales telles que l'indexation, l'enrichissement, la mémoire de traduction ou la veille... suivent dans une phase ultérieure qui nécessite d'autres connaissances, non mises à la disposition des organisateurs.

Cela explique que les organisateurs et les fournisseurs de systèmes sont bien conscients que l'évaluation serait partielle et ne porterait que sur une première phase du fonctionnement des systèmes. Une évaluation plus globale nécessiterait la mise à disposition des outils ; point non envisagé dans cette campagne. Malgré cette restriction, nous avons essayé de dégager trois tâches définies en termes *d'applications*.

L'appréciation (ou non) des performances d'un système ne peut être indépendante de l'application industrielle ou langagière pour laquelle le système a été conçu. L'application doit guider la conception même de

[56] Ces catégories ne résument donc pas à elles seules les traitements et les applications cibles de tous ces systèmes.

l'outil ; c'est l'un des enseignements majeurs de la campagne Arc A3, et que nous avons toujours maintenu, ici dans le projet Cesart. Trois applications sont fixées (description ci-après) : Extraction de termes pour l'enrichissement de ressources terminologiques ; Indexation contrôlée et enrichissement et Extraction de relations.

3. La méta-évaluation… et quelques questionnements

Dans cette partie, nous présentons avec discussion les différents paramètres (alinéas) du protocole Cesart : le corpus et l'échantillon comme ressources textuelles d'entrée, puis les référentiels et les experts comme repères de comparaison et de jugement. Nous montrons dans une deuxième partie les métriques employées dans le projet ainsi que des extraits représentatifs des résultats[57]. Pour d'autres informations sur la typologie de l'évaluation dans le projet Cesart, voir (Timimi, 2006).

3.1. Le corpus et la langue : un matériel d'entraînement et de test

Pour des raisons institutionnelles, le consortium a décidé dès le départ de la campagne de ne traiter que la langue française. Trois domaines spécialisés ont été ciblés (politique, éducation, et médecine).

Une réflexion sur la constitution de corpus spécialisés est un élément nécessaire dans tout projet d'évaluation d'outils linguistiques. Le corpus doit vérifier d'après (Pincement, 1999), trois types de conditions : *signifiance, acceptabilité et exploitabilité* en plus de la *pertinence* par rapport à un objectif d'analyse. L'ensemble de ces conditions est nécessaire pour sa réutilisabilité. De même, la production de ces corpus doit respecter la règle d'*homogénéité :* les documents retenus doivent être

[57] Les résultats complets seront disponibles dans le rapport final de la campagne.

homogènes, c'est-à-dire obéir à des critères de choix précis et ne pas présenter trop de singularité en dehors de ces critères de choix.

Sous l'ensemble de ces considérations, trois corpus en rapport avec les domaines précités ont été alors construits : un corpus *politique* composé de textes tirés du Journal Officiel de l'Union Européenne, (*1477 documents, 9 024 segments et 240 000 mots*) ; un corpus de *l'éducation* contenant des articles provenant de la revue Spirale, une revue spécialisée dans la recherche en Sciences de l'éducation, (*149 documents, 12 109 segments et 535 000 mots*) ; un corpus *médical* composé de pages web aspirées du site Santé Canada[58], (*7 514 documents, 255 161 segments et 9 000 000 mots*).

Le premier corpus (politique) a été utilisé dans les sessions d'entraînement et comme corpus de masquage tandis que les deux autres ont été utilisés comme corpus de test de la campagne officielle. Différents prétraitements (formatage, nettoyage…) ont été effectués. Les corpus ont été encodés en UTF-8 et en XML et ont été présentés en versions Dos et Unix.

3.2. L'échantillonnage : un outil assez fiable mais reste discutable

Dans une campagne d'évaluation basée essentiellement sur une appréciation humaine, et dès lors, sur un jugement subjectif, il n'est pas évident que les experts effectuent un travail de validation sur tout l'ensemble des résultats donnés par les systèmes. On procède alors par échantillonnage, comme dans la plupart des campagnes d'évaluation. Seulement, ceci n'est pas sans interrogations : de quelle ressource faut-il sélectionner l'échantillon, du corpus d'entrée ou des résultats de sortie ? Comment choisir sa taille ? Quels critères observer pour garantir une représentativité de l'échantillon ?

[58] http://www.hc-sc.gc.ca/index_f.html.

Dans (Pincement, 1999), cette règle de représentativité est bien commentée : « on peut, lorsque le matériel s'y prête, effectuer l'analyse sur échantillon. L'échantillonnage est dit rigoureux si l'échantillon est une partie représentative de l'univers de départ ». Dans ce cas, les résultats obtenus sur échantillon seront généralisables à l'ensemble de l'univers.

Dans le cas de l'Arc A3, l'échantillon était choisi dans le corpus d'entrée. Il s'agissait en l'occurrence d'un numéro particulier de la revue, noyé dans le reste des numéros de la collection. En l'occurrence, il concerne le thème de la documentation dans le milieu éducatif ; un thème assez connu par les experts humains sollicités, ce qui a facilité leur tâche de juges.

Dans Cesart, nous avons privilégié de travailler sur un échantillon issu plutôt des résultats (et non plus du corpus) ; d'autant plus que les experts disponibles ont une connaissance plus de domaines (médical et éducatif) que d'un thème spécial couvert par une partie du corpus.

Il nous a été difficile de trancher sur la taille de l'échantillon, nous avons toutefois opté pour expertiser les 1000 premiers termes donnés par chaque système. Cela peut présenter d'ailleurs un début de surcharge cognitive et générer ainsi des choix parfois arbitraires.

Pour atténuer cette limite, nous avons procédé également par comparaison automatique. Nous avons utilisé un automate qui permet de comparer les données issues des systèmes à des référentiels préétablis. Certes, cette procédure a le défaut de ne traiter que des chaînes de caractères, mais elle a le mérite de traiter l'intégralité des résultats donnés par les systèmes (et de ne pas se contenter des échantillons) et reste toutefois un indicateur sur le comportement des systèmes face à l'ensemble du corpus. L'évaluation automatique permet également d'approcher la valeur du *rappel*, une mesure impossible avec un travail humain.

3.3. L'évaluation par référentiels : un débat non encore achevé

Dans une campagne d'évaluation en terminologie, le choix du *domaine* ne peut être arbitraire. Il faut s'assurer de la disponibilité des ressources textuelles (corpus adéquats et référentiels validés) et des partenaires humains (juges experts) en rapport avec le domaine sélectionné.

Dans Cesart, nous avons pris en considération ces contraintes et nous avons utilisé deux listes référentielles, construites à partir de deux terminologies de domaine couramment utilisées :

- une liste de termes de l'éducation (36 081 entrées) basée sur le thésaurus *Motbis*[59] entrepris par le CNDP. *Motbis* est un volume complet couvrant les sciences de l'éducation et correspond au vocabulaire utilisé pour l'indexation des notices en Sciences de l'Education. Dans notre évaluation des extracteurs de termes, seules les entrées du thésaurus ont été prises en compte, les relations n'ont pas été observées pour des raisons évidentes.

- une liste de termes médicaux (22 861 entrées), basée sur la terminologie de l'équipe CISMeF [60] . Il s'agit d'une ressource terminologique médicale du service documentation du CHU de Rouen.

Si le recours à des référentiels humains préétablis nous a été de grand intérêt pour développer un cadre méthodologique normalisé d'évaluation, nous nous sommes posés quelques questionnements sur le statut de ces référentiels.

Le débat que nous avons à traiter ici, dans le cas des outils de terminologie, tient en la nécessité de procéder à l'évaluation des systèmes à travers la comparaison de listes : une liste produite par chaque système

[59] http://www.cndp.fr/motbis
[60] http://www.chu-rouen.fr/terminologiecismef/

avec une liste produite manuellement, appelée traditionnellement « liste de référence » ou « référentiel ».

Dans le cas des outils d'acquisition de terminologie, la constitution de ces listes de référence et la méthode d'évaluation elle-même, posent différents types de problèmes (L'homme, 2000) :

- d'une part ces listes contiennent des éléments qui ne sont pas forcément extraits par les systèmes (verbes et collocations par exemple) ;

- d'autre part il y a des différences entre les listes produites par les utilisateurs humains (variabilité des pratiques)[61].

A ces problèmes généraux s'ajoutent des problèmes plus spécifiques, relatifs aux applications visées par les outils de terminologie :

- en indexation : nous constatons des différentiels numériques entre indexation humaine, nécessairement sélective, et extraction automatique, nécessairement " exhaustive " (au regard des principes d'extraction retenus).

- en terminologie : nous constatons des différentiels numériques selon la nature du travail (constitution d'un vocabulaire de domaine ou enrichissement d'un vocabulaire de domaine existant).

La variation quantitative des termes nécessaires dans un domaine se double d'un autre type de variation, relatif au statut des unités. Le statut des unités lexicales est très différent selon, encore une fois, les applications visées :

- en terminologie, les unités lexicales représentent les « concepts » d'un domaine auquel appartiennent les textes dont elles sont issues.

[61] Par exemple, il y a une différence entre le thésaurus français Motbis et le thésaurus européen de l'éducation.

- en indexation, les unités lexicales représentent les « concepts » spécifiques à un document (éventuellement, mais pas toujours, les concepts distinctifs d'un document pris dans une collection).

Cette liste des principales variables rencontrées dans les pratiques montre à quel point la notion de « listes de référence » peut avoir une portée discutable. Enfin, le recours à des listes de référence ne permet pas de prendre en compte la caractéristique des logiciels étudiés : des logiciels d'aide ou d'assistance au traitement de textes. Les listes de référence constituent des listes finalisées (listes comportant uniquement les expressions retenues à l'issue du processus d'interprétation) que l'on compare à des listes non finalisées (listes destinées à aider/assister le praticien).

En dépit de ces limites, la méthode d'évaluation par comparaison sur la base de listes de référence reste la plus plébiscitée et la plus souvent utilisée pour les raisons suivantes :

- elle est particulièrement adaptée aux cas des évaluations de type " boîte noire " dans lesquelles on s'intéresse uniquement aux résultats fournis par les logiciels et non pas à leur mode de production : dès lors ne reste plus à évaluer que la " valeur d'usage " des logiciels, l'utilisation pratique/professionnelle qu'ils peuvent permettre.

- elle est particulièrement adaptée aux cas des campagnes d'évaluation ne disposant pas de logiciels : comme on ne dispose que des résultats issus d'un ensemble de textes donnés, on ne peut mettre facilement au point et entièrement corrects des modes d'évaluation des logiciels qui tiennent compte d'un travail, éventuellement, collaboratif entre le praticien et l'outil.

3.4. L'expert : entre le statut d'un juge et celui d'un usager

La campagne d'évaluation consiste, dans une de ses phases, en une analyse qualitative des résultats réalisée par des experts spécialistes

des domaines correspondant aux corpus de la campagne, à savoir les sciences de *l'éducation* (pour le corpus de Spirale) et la *médecine* (pour le corpus de Mesh). Nous avons tenu compte également dans le profil de ces experts leurs connaissances dans le domaine de l'indexation et de la documentation.

Le rôle des experts était d'examiner le degré de pertinence des listes fournies par les systèmes, présentés de façon anonyme. Chaque système étant doté d'un code d'identification que seuls les organisateurs connaissent. Ils se servent d'une grille pour accorder à chaque terme proposé une note allant de 1 à 5 selon son exactitude.

Pour le corpus de Spirale, les experts sont trois employés du CNDP rattachés au service de la documentation et chargés de l'entretien de Motbis. Pour le corpus de médecine, les experts sont deux employés du service de documentation du CHU de Rouen.

Si la mission principale des experts, en tant que juges, était d'évaluer des systèmes de terminologie, cela nous a permis de plus de les considérer comme des usagers potentiels de ces systèmes, en tant qu'outils d'assistance et d'aide, dans une activité de terminologie (constitution ou enrichissement de vocabulaire…).

3.5. Le cadre méthodologique des métriques

Dans plusieurs campagnes d'évaluation d'outils de traitement automatique de corpus, l'usage de métriques comme approche quantitative est très fréquent. La plupart des grilles d'évaluation se basent sur des notes (degrés) d'appréciation convertibles en métriques. Le recours à des métriques se présente ainsi comme une démarche scientifique rigoureuse, qui offre un outil de jugement et d'appréciation cadré et normalisé.

S'il est certain que cet outil permet de manière fiable des comparaisons entre des données dans un but de classification, la diversité des métriques possibles pour une même tâche de systèmes rend

l'exercice d'évaluation relativement complexe. Il suffit parfois d'affiner légèrement une métrique pour que le classement de systèmes change.

Dans les débats de Cesart, d'autres interrogations ont été soulevées au sujet des métriques, dues principalement au statut des référentiels utilisés dans le calcul d'une métrique (métrique entre un système et un référentiel) :

- est-il suffisant de procéder par comparaison de résultas donnés par des systèmes automatiques à des référentiels élaborés par des experts humains, pour en déduire de la qualité des systèmes ? Cette forme de comparaison n'est-elle pas réductrice dans la mesure où les résultats des outils, souvent conçus dans un but d'assistance, sont ici mal perçus (ou peu appréciés) face à la qualité pertinente d'un travail humain validé.

Il nous a semblé plus rationnel d'associer à cette métrique quelques éléments de cadrage :

- il faut surveiller la subjectivité des juges et croiser leurs appréciations pour ne pas biaiser les conclusions ; il faut calculer le degré de corrélation entre juges et ne tenir compte que des appréciations données par les juges en corrélation acceptable[62].

- si la comparaison à un référentiel humain peut paraître contestable dans la mesure où les systèmes sont souvent mal classés derrière les listes de référence, on peut procéder autrement et comparer les systèmes entre eux (évaluation inter-systèmes), càd créer un référentiel de consensus à partir des résultats communs à la plupart des systèmes (vote majoritaire).

- il faut diversifier les référentiels humains pour observer la stabilité des résultats de systèmes. Un système peut être proche par rapport au thésaurus européen de l'éducation mais ne l'est pas par

[62] On a utilisé par exemple le coefficient de Pearson, qui a montré une parfaite corrélation entre deux juges, et un écart léger par rapport à un troisième.

rapport au thésaurus Motbis, comme il peut être bien classé par rapport au référentiel médical et non par rapport à un référentiel en éducation.

Notes et grilles d'évaluation

Pour l'extraction des termes, les experts devaient attribuer à chaque CT (candidat-terme) proposé par un système une note allant de 0 à 4 correspondant au degré de pertinence :

- la note **4** signifie que le CT est présent dans le thésaurus ; la note **3** signifie que le CT n'est pas dans le thésaurus mais pertinent[63]; la note **2** signifie que le CT n'est pas dans le thésaurus mais pertinent sans former un terme complet, par exemple, le CT "***vasculaire cérébral***" dont les composants sont pertinents individuellement mais ne forment pas un terme complet ; la note **1** sous-entend que le CT n'est pas dans le thésaurus et peu pertinent, par exemple, "***matière de santé public***" est un ensemble de deux composants dont un est pertinent et l'autre non ; la note **0** est réservée au CT non pertinent.

La mesure de base retenue est celle de précision (rapport entre le nombre de termes pertinents trouvés et le nombre de termes proposés) : ***Précision = nbre de CT corrects (note 4) / nbre total de CTs extraits.***

Par affinement, nous avons élaboré d'autres variantes de cette métrique (*précisions progressives, précisions cumulées*…) voir ci-dessous.

Pour l'extraction des relations synonymiques, l'évaluation manuelle par expert est appliquée de la même manière pour mesurer la pertinence de la relation : la note 2 pour les relations présentes dans le référentiel (relations pertinentes), la note 1 pour les relations non présentes dans le référentiel mais pertinentes (relations pour enrichissement), et la note 0 pour les relations non pertinentes.

[63] Terme utile pour l'enrichissement.

Présentation de quelques résultats

Dans la partie suivante, nous nous contentons de donner quelques extraits des résultats des systèmes, le détail des résultas est disponible dans le rapport final de la campagne. Nous avons préféré, dans ce papier et ailleurs (Timimi, 2006), focaliser notre étude et réflexion sur la problématique de l'évaluation.

Les statistiques sur les sorties des systèmes sont présentées dans le tableau ci-dessous. Rappelons que le corpus Santé Canada est composé de 9 000 000 mots et Spirale de 535 000 mots.

CT extraits	**syst 1**	**syst 2**	**syst 3**	**syst 4**
Corpus Santé Canada	26 053	108 074	286 018	10 000
Corpus Spirale	3 447	60 695	41 377	10 000
CT extraits + variations extraites	**syst 1**	**syst 2**	**syst 3**	**syst 4**
Corpus Santé Canada	37 936	168 484	286 022	11 620
Corpus Spirale	4 609	67 944	41 377	10 000

Nous pouvons constater dans un premier temps que plusieurs systèmes sont surproductifs. Deux systèmes proposent plus de 41 000 candidats-termes pour une couverture terminologique du domaine de l'éducation.

Pour le calcul de la précision, nous avons observé les appréciations des experts humains sur un échantillon progressif des N premiers termes proposés par les systèmes (N = 10, 100, 1000), puis nous avons étudié grâce à l'appariement automatique l'ensemble des termes extraits par rapport à un référentiel.

Le tableau ci-dessous présente les résultats d'évaluation des termes extraits du corpus Santé Canada et de l'appariement automatique avec le référentiel issu du vocabulaire Cismef (il s'agit de pourcentage) :

Pourcentage de précision	syst 1	syst 2	syst 3	Syst 4
R10 (sur les 10 premiers CT)	0	40	10	0
R100 (sur les 100 premiers CT)	15	30	12	0
R1000 (sur les 1000 premiers CT)	09.3	26.7	06.0	0.5
Sur tous les CT	02.9	04.2	0.05	0.12

Tableau 1. *Calcul de précisions progressives*

Hormis le système 4, les autres systèmes restent bien appréciés sur les 100 premiers termes, ils le sont moins quand l'échantillon passe aux 1000 premiers termes. Quand à l'appariement automatique, nous constatons que tous les systèmes ont un taux de précision très faible, ce qui est normal vu la quantité massive des termes candidats proposés (voir supra).

Quant au classement des systèmes, l'appariement automatique basé principalement sur les chaînes de caractères (syst2 > syst1> syst4 > syst3) rejoint globalement les appréciations humaines (syst2 > syst1 > syst3 > syst4).

Dans un autre registre et pour tenir compte des autres notes d'appréciation données par l'expertise humaine (notes 3, 2 et 1) et qui correspondent aux termes non présents dans le référentiels mais susceptibles d'être pertinents, à des degrés différents, nous avons procédé à un autre calcul de précisions. Nous avons étudié l'évaluation humaine appliquée aux N premiers termes proposés par chaque système (N = 10, 100, 1000), et nous avons calculé la mesure basée sur la précision cumulée sur ces N premiers CT en considérant les différents critères d'évaluation décrits précédemment.

Ainsi les valeurs P4 proviennent des CT ayant la note 4, P3 des CT ayant les notes 3 ou 4, P2 des CT ayant les notes 2, 3 ou 4. etc. Par exemple : dans les 100 premiers termes, si nous trouvons 08 candidats-

termes pertinents (note 4) et 11 candidats-termes d'enrichissement (note 3), alors nous aurons 19 candidats-termes de précision cumulée P3.

Le tableau ci-dessous présente les résultats d'évaluation humaine avec la prise en compte de différents critères :

Précisions cumulées	S1	S2	S3	S4	Précisions cumulées	S1	S2	S3	S4	Précisions cumulées	S1	S2	S3	S4
R10 (P4)	0	50	20	0	R100 (P4)	17	31	20	0	R1000 (P4)	10.5	28.8	08.5	0
R10 (P3)	20	60	30	0	R100 (P3)	41	37	29	02	R1000 (P3)	34.2	34.1	14.6	03.4
R10 (P2)	20	60	30	0	R100 (P2)	48	38	33	03	R1000 (P2)	47.2	35.7	20.7	10.3
R10 (P1)	30	70	60	10	R100 (P1)	52	43	59	15	R1000 (P1)	52	38.5	36.1	29.1

Tableau 2. *Calcul de précisions cumulées*

D'après ce tableau des précisions cumulées observées dans Spirale, on peut reconfirmer nos constatations précitées concernant les systèmes face au corpus Santé Canada. Le système 4 ayant toujours un taux de précision très faible, les autres systèmes gardent un taux assez acceptable sur les 100 premiers termes, et moins sur les 1000 premiers termes. Le système 2 reste toujours très apprécié par l'ensemble des juges. Si les systèmes ne sont pas assez performants dans leur taux de précision (note 4 et ici score P4), ils restent assez appréciés dans leurs score P3, P2, voire P1, où les chiffres avoisinent les 40%, cela confirme une constatation soulevée par l'ensemble des experts : les systèmes sont plus appréciés comme outils d'enrichissement (notes 3 et autres) qu'outils d'acquisition. Enfin, même si le but final d'une campagne d'évaluation n'étant pas de classer les systèmes, on remarque que le système 2 reste le plus apprécié par l'ensemble des juges vu son taux de précision, par contre le système 1 est plus performant comme outil d'enrichissement. Le classement global reste presque le même que dans le corpus précédent.

Pour la tâche d'extraction des relations synonymiques, un seul système participant a été évalué sur le corpus Santé Canada. L'évaluation a été effectuée sur deux échantillons représentatifs de la sortie du système, établis en fonction des critères de la distribution des fréquences et de la spécificité des relations. Le premier échantillon a été établi en fonction de la distribution des fréquences des termes dans le corpus. 102 synonymes jugés pertinents, sur l'échantillon de 2115 renvoyés (tableau ci-dessous). Le deuxième échantillon a été construit à partir d'une liste réduite des synonymes extraits par le système en sélectionnant ceux qui contiennent au moins un des termes amorces (issus des synonymes du thésaurus). 96 termes jugés pertinents sur l'échantillon de 1451 extraits. Il semble que l'évaluation des extracteurs de relations pose problème. Par rapport à des référentiels, les systèmes sont mal notés vu la rareté des relations pertinentes proposées face au grand nombre de relations extraites… faudrait-il orienter les systèmes vers d'autres applications ?

En conclusion de cette méta-évaluation (évaluation du protocole d'évaluation), nous nous demandons toujours jusqu'à quel point il faut faire confiance à l'évaluation comme pratique nécessaire de normalisation mais difficilement modélisable, encore plus dans le cas des outils de terminologie. Au-delà de la valeur ajoutée que peut avoir un modèle adopté dans un exercice d'évaluation, ce modèle n'est-il pas, lui-même, source de troncature d'information ?

4. Références

(Chaudiron, 2001) S. Chaudiron. « *L'évaluation des systèmes de traitement de l'information textuelle : vers un changement de paradigmes* ». Mémoire pour l'habilitation à diriger des recherches en sciences de l'information, Université de Paris 10, novembre 2001.

(Cavazza, 1993) M. Cavazza. « *Méthodes d'évaluation des logiciels incorporant des technologies d'informatique linguistique* ». Paris, Rapport MRE-DIST, 1993.

(Sparck, 1996) K. Sparck-Jones, J.R. Gallier. « *Evaluating Natural Language Processing Systems: An Analysis and Review* ». Springer, Berlin, 1996.

(Timimi, 2006) I. Timimi. « *L'évaluation des systèmes d'acquisition d'outils de terminologie : nouvelles métriques, nouvelles pratiques* ». *Colloque Jadt'06*, Besançon, 19-21 avril 2006

L'évaluation des outils d'acquisition de ressources terminologiques : problèmes et enjeux

Widad Mustafa El Hadi
Laboratoire GERiiCO/Université de Lille 3
BP 60149
59653 Villeneuve d'Ascq
Widad.mustafa@univ-lille3.fr

Stéphane Chaudiron
Laboratoire GERiiCO/Université de Lille 3
BP 60149
59653 Villeneuve d'Ascq
stephane.chaudiron@univ-lille3.fr

Résumé :

Cet article décrit l'approche retenue pour l'évaluation des systèmes d'acquisition de ressources terminologiques. Après un rappel concernant la place centrale qu'occupe la terminologie en tant que ressource dans les processus de traitement de l'information spécialisée, nous avons mis l'accent sur la diversité des applications qui impliquent l'usage d'outils spécifiques à la constitution de ces ressources. Les types de ressources terminologiques se distinguent d'abord selon leur usage, c'est-à-dire par le type d'application dans lequel elles sont utilisées. Ce fait fondamental influe directement sur la définition des protocoles pour ce type d'outils. L'évaluation des outils d'acquisition doit donc s'effectuer selon une approche usage.

1. Introduction

Dans l'univers informationnel mouvant qui caractérise Internet, la terminologie occupe une double fonction : elle permet d'une part de borner des domaines de connaissance facilitant ainsi une appropriation par les usagers des techniques de gestion de connaissance et d'autre part d'améliorer la performance des logiciels de traitement avancé de l'information (TAI) en les adaptant au(x) contexte(s) d'usage des professionnels de l'information (recherchistes, analystes, documentalistes, veilleurs…).

La terminologie constitue donc une ressource essentielle dans le traitement de l'information spécialisée (qui comprend en particulier l'information scientifique et technique mais aussi les référentiels « métier » qu'il est nécessaire de modéliser pour adapter les logiciels aux différents contextes d'usage).

Après un rappel du double rôle de la terminologie dans le cycle de l'information spécialisée, l'article se focalisera sur la question de l'évaluation des logiciels d'acquisition de ressources terminologiques (ART) en montrant la nécessité de prendre en compte les dimensions d'usage. L'article mettra en évidence les enjeux théoriques et méthodologiques sous-jacents aux différents protocoles d'évaluation possibles et présentera l'approche retenue dans le cadre du projet CESART (programme Technolangue).

2. Terminologie et information spécialisée

2.1. Le cycle du traitement de l'information

Le cycle du traitement de l'information est fréquemment schématisé en trois étapes : l'acquisition de l'information qui comprend la recherche et la collecte des informations jugées pertinentes par rapport à un besoin informationnel, l'exploitation de cette information c'est-à-dire son analyse (détection et suivi de thèmes, extraction des entités

nommées, suivi temporel des thèmes et/ou entités, visualisation de l'information…) et la diffusion de l'information stratégique

Dans ce cycle, la terminologie est considérée sous deux approches complémentaires : la première approche, qualifiée d'*informationnelle*, considère la terminologie comme la ressource essentielle de la gestion de l'information. C'est elle qui permet de borner conceptuellement un domaine d'intérêt à travers les termes que l'on trouve dans les corpus relevant du champ. Ainsi, dans le domaine de l'information scientifique et technique, et plus généralement dans les champs de l'information spécialisée (ou professionnelle) les terminologies ont toujours occupé une place centrale à côté des nomenclatures, des taxonomies et des nombreux outils classificatoires. La seconde approche, qualifiée de *logicielle*, considère la terminologie comme une ressource linguistique nécessaire au bon fonctionnement des outils de traitement avancé de l'information tels que les logiciels de veille.

Ainsi, à partir des différents univers informationnels, notamment la mémoire et le patrimoine immatériel des entreprises, les nombreux corpus textuels produits dans les organismes, scientifiques, techniques, réglementaires, etc., des terminologies spécialisées sont produites qui alimentent en ressources linguistiques (dictionnaires, thésaurus, réseaux sémantiques…) les logiciels de traitement avancé de l'information en leur permettant de fonctionner et de répondre aux besoins des usagers. Ces logiciels permettent à leur tour de traiter les corpus relevant des différents univers informationnels. La figure 1 représente ce fonctionnement cyclique.

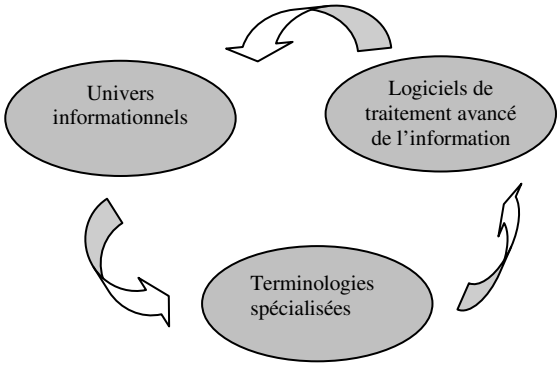


Figure 1 : La terminologie au coeur du traitement informationnel

En tant que corpus délimitant des univers informationnels, les terminologies sont des outils de description de champs de connaissance au même titre que les nomenclatures, les thésaurus ou les ontologies. A ce titre, elles interviennent dans différents processus de gestion de l'information comme le *Knowledge Management*, la veille stratégique, l'intelligence économique, la valorisation de la mémoire d'entreprise ou la gestion du patrimoine informationnel de l'entreprise.

La figure 1 visualise le double rapport qu'entretient la terminologie avec l'ingénierie linguistique. Dans de nombreuses applications (la traduction, l'analyse de contenu, le résumé, la recherche et le filtrage d'informations notamment) la terminologie est utilisée comme une ressource linguistique nécessaire au fonctionnement des applications de traitement automatique des langues (TAL). Dans d'autres applications, les terminologies sont à l'inverse des résultats produits par les outils d'ingénierie linguistique et, plus largement des applications de traitement informatique des langues. Dans ce cas, elles sont construites à partir de grands volumes de données textuelles afin de produire des listes de termes qui permettront ensuite d'élaborer des dictionnaires, des terminologies, des thésaurus…

2.2. Les applications du traitement avancé de l'information

L'utilisation de ressources terminologiques et ontologiques se fait à des degrés variables. Un domaine clé pour l'utilisation de ces ressources est la recherche d'informations multilingues (RIM). Ce type d'application fait appel à une panoplie de ressources lexicales, terminologiques et ontologiques et qui est par ailleurs tributaire des progrès de la traduction automatique (TA).

Parmi les principales applications des nouvelles technologies de l'information utilisant la terminologie, citons notamment :

- Le résumé automatique avec la méthode d'extraction de termes dans ce processus ;
- Les systèmes de questions-réponse, qui nécessitent l'extraction de concepts, de définitions ainsi que l'extraction des entités nommées ;
- Les outils de classification automatique (construction d'agrégats de termes) ;
- La génération des ontologies à partir de ressources terminologiques pour des applications d'accès à l'information dans l'environnement du web sémantique ;
- Ces nouvelles applications ont eu un impact sur les langages de représentation et les formats d'échange de données terminologiques (cf. notamment Samson-Alt *et al.* 2006).

3. L'acquisition de ressources terminologiques (ART)

3.1. Les enjeux de l'ART

Les flux de l'information spécialisée sont à la base de la constitution des stocks de connaissances et, dans ce contexte, les

terminologies sont des vecteurs importants d'information et de connaissances. L'évolution de leur rôle dans l'accès à l'information spécialisée est à la fois dictée par les nouveaux besoins informationnels des usagers mais rendue possible par les mutations technologiques. Plusieurs facteurs ont favorisé l'évolution de la terminologie : de la révolution informatique à l'Internet, en passant par le sociocognitivisme et la linguistique de corpus, les méthodes du traitement de la terminologie ont subi de profondes mutations. En raison de la nécessité de manipuler de plus en plus de termes, un certain nombre d'applications proposant une reconnaissance automatique des termes ont été développées et tout particulièrement dans le domaine de la recherche d'information. Alors que la recherche dans ce domaine a commencé dans les années cinquante la reconnaissance automatique de termes visant à constituer des corpus de référence pour la traduction automatique, la rédaction technique, les systèmes à bases de connaissances et l'accès à l'information multilingue demeurent un développement récent.

L'extraction terminologique suppose deux activités essentielles : l'acquisition terminologique qui implique la découverte de nouveaux termes pour des applications langagières et l'indexation[64]. Le but de l'acquisition terminologique est la constitution de ressources terminologiques telles que les vocabulaires contrôlés, index et thesaurus, ontologies. Les thesaurus sont par exemple utilisés depuis longtemps pour assister le processus de recherche d'information lors de la phase d'expansion de la requête, pour la recherche interlingue d'information ou pour l'interrogation de bases de données. La constitution automatique de thesaurus est une tradition déjà ancienne et les différentes études menées dans ce domaine concernaient soit la découverte de nouveaux termes ou l'établissement de relations sémantiques. Les outils d'extraction

[64] Ce terme est à prendre dans un sens générique et évolutif. Autrement dit, il couvre l'ensemble des opérations telles que l'indexation du contenu de documents ainsi que les diverses formes d'indexation « fine » dans les contextes de résumé automatique, question-réponse, recherche thématique, classification automatique de document…

terminologique bilingue, fondés sur une analyse des parties du discours et des technologies d'alignement visant à extraire des termes candidats et leurs traductions peuvent être utilisées pour ces deux tâches. Les composants monolingues de ce type de systèmes peuvent être utilisés pour l'indexation et pour l'extraction terminologique alors que les composants bilingues sont utilisés pour la traduction ainsi que pour la recherche d'information dans des bases documentaires multilingues. Bien que ce type de système ait été conçu initialement pour l'extraction et la traduction de termes, il a été ensuite utilisé également et adapté pour l'indexation.

3.2. La dimension de l'usage dans l'ART

La pratique terminologique a toujours consisté à valider et à normaliser les terminologies produites manuellement et/ou automatiquement. Par ailleurs, les types de ressources terminologiques se distinguent d'abord selon leur usage, c'est-à-dire par le type d'application dans lequel elles sont utilisées. Ce fait est non seulement fondamental mais il a de plus une incidence sur la définition des protocoles d'évaluation pour ce type d'outils.

Il va sans dire que la qualité de la terminologie est un élément fondamental quel que soit le contexte d'usage dans lequel elle est utilisée. Il existe quelques pistes pour détecter la qualité de la terminologie extraite. Dans les répertoires classiques, il existe des moyens explicites d'évaluation qui guident l'utilisateur dans la prise en compte des informations ainsi que d'autres indicateurs qui, tout en n'étant pas explicites, agissent également comme pôles d'orientation des utilisateurs (Cabré, 1998). Les marques d'usage dans les dictionnaires spécialisés sont peu fréquentes et, si elles apparaissent, sont limitées à des catégories déterminées. Cette approche, bien qu'elle soit insuffisante, constitue cependant un moyen d'aborder la question de l'usage.

Une terminologie de qualité doit être fondée sur le travail descriptif : elle doit offrir d'abord les formes documentées dans l'usage réel du milieu professionnel traité (*ibid.* p. 20), ce qui peut favoriser la

qualité du travail terminologique. Cette qualité doit être jugée dans un contexte d'usage bien déterminé. La terminologie extraite doit, autrement dit, répondre à une contrainte de pertinence par rapport à l'usage qui est fait. Nous pouvons énumérer[65] quelques contextes d'usage dans lesquels les ressources terminologiques sont utilisées, en particulier :

a) les applications langagières comme la traduction, la rédaction technique et la localisation… ;

b) les applications d'accès à l'information : la construction de thesaurus pour l'indexation et la recherche d'information (applications classiques) la recherche d'information précise et thématique (Ferret *et al.*, 2006), la construction d'index pour les documents techniques (Lallich-Boidin *et al.*, 2006) la veille et intelligence économique (Ibekwe-Sanjuan, 2006), construction d'ontologies pour le web sémantique **(**Aussenac-Gilles *et al.*, 2006), résumé automatique et condensation de textes (Minel, 2002), la recherche d'information multilingue (Fluhr, 2006).

Dans le projet CESART[66], nous avons d'abord tenté de définir le protocole par type d'outils puis avons ensuite défini les usages possibles. Sachant que les extracteurs de termes sont des systèmes génériques, nous avons défini deux types d'application, l'enrichissement de référentiels et la mise à jour de référentiels (Mustafa El Hadi *et al.*, 2006b).

[65] Les diverses applications de la terminologie sont énumérées et documentées dans une publication récente. Nous renvoyons le lecteur à (Mustafa El Hadi, 2006a).

[66] Campagne d'Évaluation des Systèmes d'Acquisition de Ressources Terminologiques (CESART) organisée dans le cadre du programme Technolangue <http//www.technolangue.gouv.fr> mis en place par les ministères de la recherche, de l'industrie et de la culture. Le projet consiste à concevoir et valider un modèle d'évaluation de logiciels d'acquisition de ressources terminologiques.

4. Vers des protocoles d'évaluation centrés sur l'usage

4.1. Paradigme usager *versus* paradigme technique

L'évaluation d'outils d'acquisition de ressources terminologiques tire ses fondements méthodologiques et théoriques du paradigme d'évaluation. Ce paradigme s'est construit et repose essentiellement sur les travaux d'évaluation des systèmes de recherche d'information (SRI). Ces travaux ont fait l'objet de diverses publication mais pour une synthèse complète et actualisée nous renvoyons le lecteur aux travaux de S. Chaudiron (Chaudiron, 2001 & 2004). La majorité des modèles d'évaluation utilisés pourrait être classée sous le modèle générique de qualimétrie[67]. (Chaudiron, 2001, p. 73). D'après l'auteur, ce modèle est connu et éprouvé mais présente deux principaux défauts : bien qu'il soit conçu dans le souci légitime d'améliorer les performances intrinsèques des matériels et logiciels informatiques, ce paradigme d'évaluation n'a pas permis, à de rares exceptions près, de sortir d'une vision exclusivement techno-centrée. Le deuxième défaut est la non prise en compte des phénomènes d'acceptabilité ou de satisfaction des utilisateurs. Ces phénomènes prennent une importance grandissante dans les évaluations d'outils d'acquisition de ressources terminologiques.

Les modèles d'évaluation orientés usagers[68] caractérise l'ensemble des pratiques évaluatives qui mettent l'usager au cœur de leurs préoccupations, quelle que soit la diversité des méthodes et des protocoles utilisés.

[67] Le modèle qualimétrique est issu de travaux sur la qualité de produits industriels initiés aux Etats Unis dans les années 70. Cette approche telle qu'elle est le plus souvent mise en oeuvre, n'intègre pas la dimension sociale de l'usage ; elle ne prend en compte que l'utilisateur idéal du système, qui évidemment n'existe pas, et non l'usager.

[68] Pour une présentation plus détaillée de cette dichotomie approche techno-centrée et approche orientée usagers, voir (Chaudiron, 2004).

L'évaluation dans le paradigme usager étudie les utilisateurs d'un système et leur interaction avec le système. Les systèmes d'information sont alors considérés comme des systèmes de médiation entre un producteur d'information (l'auteur) et un utilisateur. Le système informatique a pour objectif de faciliter cette communication, interaction et médiation. Les facteurs humains acquièrent une grande importance et la nécessité de connaître les utilisateurs engendre de nombreux travaux que l'on peut regrouper sous le thème « études des usagers » (Chaudiron, 2001). La littérature est abondante sur la connaissance des publics, la mise au point des modèles utilisateurs, les analyses des besoins, les enquêtes de satisfaction, etc. mais peu d'études portent sur l'évaluation des outils de traitement avancé de l'information en contexte professionnel.

A l'inverse de l'approche système, le paradigme usager (ou cognitif) considère que l'attention doit être portée sur les besoins réels de l'usager et sur son environnement. Il s'agit désormais d'étudier et d'évaluer comment les usagers définissent et reconnaissant et traitent leurs besoins d'information dans différentes situations.

D'autres études récentes prennent en charge la dimension ou le rôle des usagers dans le développement d'une technologie. Elles décrivent la manière dont les utilisateurs consomment, modifient, domestiquent, reconfigurent et même résistent à une technologie. Si de nombreuses études sont actuellement menées pour étudier le comportement des utilisateurs face à une technologie, inversement, d'autres relativement récentes s'efforcent d'évaluer l'impact de la technologie sur les utilisateurs d'une technologie (Oudshoorn, 2003).

Des études de plus en plus nombreuses portent sur les usagers et les utilisateurs d'une technologie afin leurs pratiques et leurs comportements. Qui doit les définir en tant que groupe ? Doit-on les isoler comme groupe autonome de consommateurs ? Comment sont-ils perçus pars les concepteurs des technologies ?

Les concepteurs et les usagers d'une technologie ont été pendant très longtemps perçus par les chercheurs comme deux entités séparées et des études récentes tentent de les considérer comme deux faces d'un même problème. L'objectif est la co-construction des usagers et des technologies qui va au-delà des visions technologiques déterministes (*ibid.*). Le point commun des différentes approches est de proposer une modélisation des usagers et de leurs comportements. La tentative initiale d'intégrer dans le paradigme « système » certaines caractéristiques comportementales pour améliorer la performance des systèmes s'est progressivement transformée en un thème de recherche à part entière. Dans le domaine de la recherche d'information, cette évolution est fondée en grande partie sur certaines critiques portée à l'encontre du paradigme système (Chaudiron, 2004) en particulier sur le fait que les requêtes ne sont que des représentations imparfaites des besoins d'information, que la notion de pertinence est inadéquate pour rendre compte de la satisfaction des usagers et enfin sur le fait que les mesures d'évaluation utilisés dans cette approche ne sont pas appropriées. Les modèles « usager » intégrant cet apport des sciences humaines et sociales mettent l'accent sur la dimension comportementale des utilisateurs en situation de recherche d'information.

4.2. L'approche CESART

Dans le projet CESART (Mustafa El Hadi *et al.*, 2004a & 2006b) et (Mustafa El Hadi, 2006a), notre approche d'évaluation tient compte de la dimension de l'usage. La prise en compte de cet aspect serait l'apport le plus important. Dans le domaine de l'acquisition de ressources terminologiques, nos évaluations antérieures de ces outils ont mis en évidence la nécessité de prendre en compte les diverses classes d'applications (Mustafa El Hadi *et al*, 2001). Par ailleurs, nous avons fait le constat que les ressources terminologiques se distinguent d'abord selon leur usage, c'est-à-dire par le type d'application dans lequel ces ressources sont utilisées. Selon l'application, elles répondent à des besoins de conceptualisation et de spécification différents (cf. aussi Bourigault, 2002a). Tout protocole d'évaluation de ce type d'outils doit tenir compte de cette spécificité. De fait, la dimension d'usage s'impose

et elle parait particulièrement pertinente dans le contexte d'acquisition de ressources terminologiques.

L'intérêt majeur de CESART a été de valider les terminologies extraites par des spécialistes des domaines qui sont en même temps des utilisateurs potentiels de ressources terminologiques. Les contextes d'usages choisis ont été les suivants : construction de référentiels génériques (listes à plat) ; enrichissement et mise à du thesaurus. Pour le corpus médical, il s'agissait d'enrichir le *MeSH* version française (ce que nous avons appelé tâche n° 1 du protocole CESART, (Mustafa El Hadi *et al.*, 2006b) et de valider les relations de synonymie extraites puis de les comparer aux relations existantes dans le thesaurus *MeSH*. Pour le deuxième corpus, les tests ont été effectués sur le thesaurus *MotBis*.

La réalisation concrète de l'expérimentation par le groupe du CNDP a été très intéressant mais l'aurait été davantage si on avait pu utiliser un outil vidéo pour enregistrer les diverses étapes de cette évaluation pour ainsi mieux apprécier la réaction des utilisateurs spécialistes face aux produits terminologiques. Un outil tel que *Camstudio* permet en effet de filmer les réactions et appréciations de l'équipe au lieu de les noter au fur et à mesure du déroulement de l'évaluation[69].

Dans les études sur l'usage qui mettent au centre les préoccupation de l'usager, la question de l'interface est fondamentale (cf. Chaudiron, 2001). Dans le cadre de l'acquisition de ressources terminologiques, l'interface peut être un simple dispositif de manipulation du système d'acquisition mais aussi une aide à la validation et à l'évaluation. Certains dispositifs constituent des adaptations logicielles de formulaires « papier », d'autres sont des outils d'accompagnement de l'évaluation, c'est-à-dire qu'ils aident l'évaluateur à structurer et à organiser l'évaluation, d'autres enfin permettent l'évaluation automatique. Cet aspect n'a pas pu être traité dans CESART malgré son importance mais doit impérativement pris en compte dans

[69] Cet outil a été utilisé dans deux cadres différents (cf. Béguin 2001a, Béguin *et al.*, 2001b).

l'élaboration de protocole d'évaluation de cette catégorie d'outils afin de faciliter la validation des résultats obtenus par les divers outils. Il faudrait rappeler dans ce contexte que les méthodes d'acquisition de ressources terminologiques ne peuvent pas être entièrement automatisables et un travail de validation manuelle des résultats est nécessaire d'où l'importance de l'interface. Ce constat est partagé par de nombreux chercheurs (Bourigault *et al.*, 2002a, 2002b), (Nazerenko *et al.*, 2002), (Biskri *et al.*, 2000). Les outils devraient être conçus comme des outils coopératifs qui assistent l'utilisateur dans son travail. Pour qu'une « coopération » soit possible, un certain nombre de points devraient être pris en considération dont le plus important selon nous est l'interface ou « table de manipulation des résultats ».

5. Conclusion

CESART nous a permis d'avoir une bonne idée de l'offre technologique francophone en matière d'outils d'ART. Nous pensons que les extracteurs de termes ont atteint une certaine maturité scientifique ce n'est pas le cas des extracteurs de relations. Un long chemin reste à parcourir.

Concernant notre expérience d'évaluations centrées usage, même si notre réflexion est appelée à être affinée, elle pourrait être généralisée à d'autres types de dispositifs ou d'outils. Nous pensons notamment aux ontologies qui sont des structures d'organisation et de représentation des connaissances innovantes qui font appel, entre autres, aux technologies linguistiques. Les ressources terminologiques participent de plus en plus à la construction d'ontologies. L'évaluation de ces outils est encore à ses débuts et il y a un intérêt certain à les tester dans leurs contextes d'usage. Pour l'évaluation des ontologies, il serait intéressant d'adopter des modèles d'évaluation fondés sur le gain de temps, qui sont des modèles éprouvés (Minel, 2002), (Bourigault *et al.*, 2002a) : il s'agit d'une expérience qui prend en compte la mesure du temps nécessaire à la construction d'une ontologie. D. Bourigault (Bourigault *et ali.*, 2002b) signale une expérience d'évaluation d'une ontologie sur la réanimation

chirurgicale. Le référentiel utilisé était tiré d'un thesaurus utilisé dans cette spécialité. Le temps que le médecin a mis à construire l'ontologie couvrant la partie correspondant au thesaurus était estimé à 50 heures pour une ontologie de 2000 concepts. Par ailleurs, les modèles fondés sur le gain de temps sont aussi appliqués dans l'évaluation des résumeurs automatiques (Minel, 2002).

Comme on le constate, les perspectives de recherche sont multiples. On peut en citer au moins deux : d'une part, l'amélioration des protocoles d'évaluation concernant les systèmes d'acquisition terminologiques au sens large (termes, concepts, relations sémantiques, ontologies…). Cet axe de réflexion concerne en premier lieu les professionnels de l'information spécialisée et de la gestion des connaissances qui ont besoin de disposer de cadres d'évaluation stables et adaptés à leurs pratiques informationnelles. Le deuxième axe concerne plus généralement la question de l'appropriation des dispositifs techniques de manipulation des connaissances numériques par différents publics, notamment professionnels. L'enjeu concerne ici plus spécifiquement les sciences de l'information et de la communication pour lesquelles les dimensions fonctionnelles, symboliques et cognitives des dispositifs techniques sont centrales.

Bibliographie

(Aussenac-Gilles, N, Hernandez, N, Baziz, M., 2006), « Ontologies pour la recherche d'information, importance de la dimension terminologique », in W. Mustafa El Hadi (dir.), *Terminologie et accès à l'information*, Paris, Hermès, p. 211-234.

(Béguin 2001a), A. Béguin, « Le corps dans les lectures à l'écran », in *Spirale*, numéro spécial *Nouveaux outils, nouvelles écritures, nouvelles lectures*, oct. 2001, n°28, p. 145-162.

(Béguin 2001b), A. Béguin, B. Amougou, « Du laboratoire au cédérom : expérience sur cédérom pour acquérir des connaissances en chimie », in *Recherches en communication*, 2001, n°16, p. 111-130.

(Biskri *et al.*, 2000), I. Biskri, S. Delisle, « User-relevant access to textual information through flexibile identification of terms: a semi-automatic method and software based on a combination of N-Grams and surface linguistic filters », in *Actes du Colloque RIAO 2000 : Content-Based Multimedia Information Access*, Paris, CID, 2000.

(Bourigault, 2002a), D. Bourigault, D. Lame, « Analyse distributionnelle et structuration de terminologie : application à la construction d'une ontologie documentaire du droit », in A. Nazarenko et T. Hamon (dir.), *Structuration de terminologie*, Revue TAL, vol. 43- n°1, Paris, Hermès, 2002, p. 128-150.

(Bourigault, 2002b), D. Bourigault, N. Aussenac-Gilles, J. Charlet D., « Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas ». in *Revue d'Intelligence Artificielle*, vol. X – n° X/2002.

(Cabré, 1998), M.-T. Cabré, « A propos de la notion de qualité en terminologie » in *La banque des mots*; Numéro Spécial 8 *Qualité et terminologie*; CTN INaLF CNRS et Le conseil international de la langue Française, Paris; 1998, p. 7-33.

(Chaudiron, 2001), S. Chaudiron, *L'évaluation des systèmes de traitement de l'information textuelle : vers un changement de paradigmes*, Mémoire pour l'habilitation à diriger des recherches en sciences de l'information, Université de Paris 10, novembre 2001.

(Chaudiron, 2004), S. Chaudiron, «La place de l'usager dans l'évaluation des systèmes de recherche d'informations », in S. Chaudiron (dir.), *Evaluation des systèmes de traitement de l'information*, Paris, Hermès, 2004, p. 287-310.

(Ferret *et al.*, 2006), O. Ferret, B. Grau « Terminologie et accès à l'information thématique et précise » in W. Mustafa El Hadi (dir.), *Terminologie et accès à l'information*, Paris, Hermès, 2006, p. 119-136.

(Fluhr, 2006), C. Fluhr, « Le rôle de la terminologie dans les systèmes d'information multilingues », in W. Mustafa El Hadi (dir.), *Terminologie et accès à l'information*, Paris, Hermès, 2006, p.235-247.

(Ibekwe-San-Juan, 2006), F. Ibekewe-San-Juan, « Connaissances terminologiques et systèmes d'informations stratégiques », in W. Mustafa El Hadi (dir.), *Terminologie et accès à l'information*, Paris, Hermès, 2006, p. 187-207.

(Lallich-Boidin *et al.*, 2006), G. Lallich-Boidon, A. Smolczewska « Une nouvelle lecture de la structure d'un document en vue de la construction d'index », in W. Mustafa El Hadi (dir.), *Terminologie et accès à l'information*, Paris, Hermès, 2006, p. 101-117

(Minel, 2002), J-L. Minel, *Filtrage sémantique : du résumé automatique à la fouille de textes*, Paris, Hermès, 2002

(Mustafa El Hadi, 2001), W. Mustafa El Hadi, I. Timimi, A. Béguin, M. Debrito, « The ARC A Project: Terminology Acquisition Tools : Evaluation Method and Tasks », in *Evaluation Methodologies for Language ad Dialogue Systems Workshop*, ACL/EACL, Toulouse, 6-7 Juillet 2001, p. 41-50.

(Mustafa El Hadi, 2004a), W. Mustafa El Hadi, I. Timimi, M. Dabbadie, « CESART Project », in *Actes du Colloque LREC 2004*, Lisbonne, ELDA, 2004.

(Mustafa El Hadi, 2004b), W. Mustafa El Hadi, « L'évaluation d'outils d'acquisition de ressources terminologiques », in S. Chaudiron (dir), *Evaluation des systèmes de traitement de l'information*, Paris, Hermès, 2004, p. 149-169.

(Mustafa El Hadi, 2006a), W. Mustafa El Hadi, « Usages des technologies linguistiques dans les traitements de l'information : essai de réflexion », In : *Actes du XVe Congrès national des Sciences de l'information et de la communication, Questionner les pratiques d'information et de communication. Agir professionnel et agir social*, Bordeaux, 10-12 mai 2006. Bordeaux, Presses Universitaires de Bordeaux, 2006.

(Mustafa El Hadi *et al.*, 2006b), W. Mustafa El Hadi, I. Timimi, M. Dabbadie, K. Choukri, O. Hamon, Y-C. Chiao, « Terminological resources acquisition tools : towards a user-oriented evaluation », in *Actes du Colloque LREC 2006*, Gènes, 25-26 mai 2006, ELDA.

(Nazerenko *et al.*, 2002), A. Nazarenko, T. Hamon, « Structuration de terminologie : quels outils pour quelles pratiques ? », in A. Nazarenko et T. Hamon (dir.), *Structuration de terminologie*, Revue TAL, vol. 43- n°1, Paris, Hermès, 2002, p. 7-18.

(Oudshoorn, 2003), N. Oudshoorn, T. Pinch, *How users matter. The co-construction of users and technologies*, Cambridge (MA), MIT press, 2003.

(Samson-Alt, S, Kramer, Isabelle, Romary, L, Roumier, Joseph, 2006) « Terminologie : principes, méthodes, modèles », in in W. Mustafa El Hadi (dir.), *Terminologie et accès à l'information*, Paris, Hermès, p. 163-185

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 en 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

http://www.porphyre.org

Institut Porphyre
Savoir et Connaissance

ISBN : 2-9516453-7-6
EAN : 9782951645370