Monique Slodzian, Mathieu Valette (2009) « Connaissances prescrites ou connaissances décrites ? L'apport de la sémantique des textes », Patrimoine 3.0, *Actes du 12e Colloque International sur le Document Electronique (CIDE.12)*, Khaldoun Zreik, dir., Europia Productions, Paris, pp. 129-141.

version soumise [24 juillet 2009] – merci de se reporter à la version publiée

Connaissances prescrites ou connaissances décrites ? L'apport de la sémantique des textes

Monique Slodzian (1)
mslodz@inalco.fr

Mathieu Valette (2)
mvalette@atilf.fr

CRIM-ERTIM (EA 2520) INaLCO, Paris (1)

ATILF (UMR 7118) CNRS, Nancy (2)

Mots-clés : Connaissances prescrites, Vérité forte/vérité faible, Systèmes d'organisation des connaissances, Sémantique des textes, Parcours interprétatif, Planification de l'information, Forme sémantique, Thématisation, Lexicalisation

Keywords: prescriptive knowledge, Strong/weak truth, Knowledge Organisation Systems, Text Semantics, Interpretative path, Information planification, Semantic form, thematisation, lexicalisation.

Résumé : L'article vise à montrer que le modèle collaboratif de communication des connaissances revendiqué par le Web 2.0 ne rompt pas de manière significative avec le modèle épistémologique antérieur, issu du positivisme logique, notamment par son primat référentialiste prescriptif. En postulant *in fine* l'existence de concepts primitifs partagés, il est conduit à reproduire les mêmes limites que le Web sémantique fondé sur un socle de métadonnées réputées universelles. Par ailleurs, une acceptabilité indiscutée des connaissances de vérité faible pose des problèmes de fiabilité et de garantie susceptibles de compromettre le succès du modèle. L'article entend démontrer dans une deuxième partie en quoi la sémantique des textes peut contribuer à objectiver les connaissances par la description de parcours interprétatifs. Considérant que les textes relèvent d'une planification de l'information, l'article explicite la notion de *forme sémantique*, entre le texte et le concept, et envisage la possibilité de faire émerger des *préconnaissances* non encore lexicalisées. Cette proposition théorique est illustrée à partir de discours de prévention contre le tabagisme issus du Web.

Abstract : This paper aims at showing that the collaborative communication model upheld by Web 2.0 doesn't significantly break with the previous epistemological model, stemming from logical positivism, mainly due to its prescritive referentialist primacy. By assuming the existence of shared primitive concepts, it is finally led to reproduce the same drawbacks as the Semantic Web founded on a bunch of metadata given as universal. Moreover, accepting non-expert knowledge without debate raises the issue of reliability and expertise, exposing Web 2.0 to a fatal risk of misinformation. In its second part, the paper argues that textual semantics is able to contribute to objectivating knowledge by describing interpretative scenarios. Considering that texts come under a planified pattern of information, the authors clarify the notion of *semantic form*, between text and concept, and consider the possibility of eliciting *pre-knowlede elements*, not yet lexicalized. As a concrete framework to this theoretical proposal, arguments will be supported by anti-smoking texts trawled from the Web.

Introduction

La masse de données qui constitue le Web ne relève pas pour l'essentiel de champs de connaissances homogènes et discrétisables en ontologies. En effet, ce qu'on appelle faute de mieux « connaissances générales » ou « connaissances vulgarisées » s'opposent en premier lieu aux connaissances historiquement encadrées et présumées garanties par les domaines scientifiques et techniques que des documentalistes ont organisées en classifications au fur et à mesure de leur constitution.

Monique Slodzian, Mathieu Valette (2009) « Connaissances prescrites ou connaissances décrites ? L'apport de la sémantique des textes », Patrimoine 3.0, *Actes du 12e Colloque International sur le Document Electronique (CIDE.12)*, Khaldoun Zreik, dir., Europia Productions, Paris, pp. 129-141.

version soumise [24 juillet 2009] – merci de se reporter à la version publiée

L'hypothèse que l'on puisse accorder de la valeur à des connaissances de vérité faible coïncide avec l'avènement du Web et de ses contenus surabondants offerts à tous. Le propre du tout-venant du web consiste en effet à demeurer hors du contrôle rationnel régi par les communautés d'experts et à défier ainsi toute dichotomie de type vulgarisé *vs* scientifique, voire scientifique *vs* pseudo-scientifique. Ceci par opposition aux notions d'arbre de connaissance ou d'ontologie qui supposent une finitude des données et une production raisonnée, restrictive et contrôlée des connaissances. Les théories et systèmes de classification produits tout au long du XXème siècle concevaient en effet « la connaissance » comme une structure rigoureuse (mathématique ou logique), se présentant comme un système formel. La science étant définie comme une structure logique commune à tous, le but de la connaissance consistait à orienter parmi les objets et à prédire leur comportement : on était censé y parvenir en découvrant leur *ordre* et en assignant à chacun d'eux la place qui est la sienne au sein de la structure du monde. Carnap donnera la théorisation la plus aboutie de ce modèle partagé par les positivistes logiques (Carnap, 1928). Sans nous attarder davantage sur la dichotomie entre connaissance intuitive vs connaissance scientifique comme genre supérieur de connaissance, nous retiendrons ce moment précis où le mode énonciatif de la science passe impérativement par des formes d'expression et de relations fortement *prescriptives*, au point que la conformité de la forme garantit la valeur du contenu (Carnap, 1934, Schlick, 1932). Il nous semble être le point de référence indispensable pour mesurer l'ampleur des ruptures épistémologiques accomplies et pour appréhender la question du texte et de sa description, souvent oubliée dans les plis de la philosophie de la connaissance.

Nous entendons montrer dans cet article, que les textes présentent une matérialité qui se prête à l'analyse et à l'objectivation. Par objectivation nous n'entendons pas une extraction de connaissances déliées des textes et des interprétations possibles, comme le font les approches prescriptives, mais au contraire une prise en compte des conditions de production et d'interprétation des documents. L'enjeu est de faire émerger les connaissances des textes et de les caractériser en tenant compte des conditions de leur production et de leur interprétation, de façon à évaluer leur pertinence par rapport à une tâche donnée. Notre proposition méthodologique, s'inspirant de la sémantique des textes, aura pour objectif de faire émerger des informations susceptibles d'être constituées en connaissances.

1 Problématique

1.1 L'information est une relation

La construction formelle que présupposait la science unitaire des positivistes logiques reposait notamment sur le concept de *relation***,** au point que l'on pourrait suivre Barlow (Barlow, 1994, p.13) et poser que "l'information est une relation". Les controverses sur la taxinomie et le nombre de relations logiques depuis l'élaboration de la Classification décimale universelle (DCU) montre bien le succès de cette formule lapidaire: d'une vingtaine pour Coates (Coates,1960), on passe à quelque 5000 relations dans le système CYC[1]. La polémique sur le statut épistémique des métadonnées proposées par le Dublin Core Metadata Initiative trouve ici sa place[2]. En effet, les relations classiques qui structuraient les systèmes de classification bibliographiques relevaient de l'empirisme logique et de l'heuristique de la vérité scientifique, tandis que les relations de CYC concernent des connaissances de sens commun et, plus paradoxal encore, les métadonnées du Dublin Core s'adressent aux connaissances illimitées du Web sémantique. Les folksonomies ne seraient-elles pas en dernière instance le symptôme de l'impuissance intrinsèque de tout système de métadonnées à structurer une masse illimitée de données?

En fin de compte, la difficulté théorique et pratique de concevoir des ontologies générales institutionnellement prédéfinies pour classer la masse d'informations hétérogènes aboutit à une substitution de paradigme : on passe du paradigme de l'information à celui de la communication (Web 2.0). La notion de web horizontal exprime ce renversement en invitant les usagers à indexer eux-mêmes le web à l'aune de leurs intérêts propres. Classification traditionnelle et folksonomie correspondraient ainsi à deux options philosophiques opposées. La

[1] Base de connaissances issues du sens commun ayant pour but le développement d'un système intelligent.

[2] Référence sur ladite polémique.

Monique Slodzian, Mathieu Valette (2009) « Connaissances prescrites ou connaissances décrites ? L'apport de la sémantique des textes », Patrimoine 3.0, *Actes du 12e Colloque International sur le Document Electronique (CIDE.12)*, Khaldoun Zreik, dir., Europia Productions, Paris, pp. 129-141.

version soumise [24 juillet 2009] – merci de se reporter à la version publiée

première répondrait à une structure hiérarchique *top-down* adossée à l'objectivité supposée des technosciences, la seconde assumerait la subjectivité d'un étiquetage *sui generis* issu d'un filtrage collectif.

1.2 Entre l'expert et la sagesse des foules, quelle place pour l'objectivité ?

Cependant, l'opposition entre « stockage de l'information hiérarchique » et « filtrage collaboratif » (Origgi, 2008) mérite d'y regarder de plus près. S'ils se distinguent par leur intentionnalité (informer vs communiquer), il reste à prouver qu'ils procèdent d'épistémologies différentes et qu'ils se situent de part et d'autre de la ligne référentiel /non référentiel, autrement dit qu'ils ne procèdent pas également de l'empirisme logique. L'activité de connaissance présupposée dans les deux cas n'est-elle pas conçue comme le transcodage d'un langage-objet en un métalangage et réciproquement (Rastier, 1995) ? C'est en tout cas ce que suggère l'idée de monde comme « catalogue » (Olivier Ertzscheid, 2008) qui laisse entendre que l'on peut accéder directement aux objets, l'univers de référence étant induit par l'objet lui-même. Il s'agit de saisir les mots par la dénotation et donc de représenter la relation des objets au monde. Cette vision néo-platonicienne du rapport des objets –virtuels et réels- au monde à travers le médium technologique est théorisée sous le nom d' « Internet des objets » et certains de ses tenants se risquent à présenter l'informatique comme le socle de la structure du monde (Wolfram, 2001). Leurs propositions viennent en droite ligne des travaux des positivistes, Russel, Whitehead et en particulier Carnap, dont *La Construction logique du Monde* peut être considérée à cet égard comme préfigurant la philosophie du Web.

On se demandera, par ailleurs, si le filtrage collaboratif à l'origine des folksonomies échappe à un modèle communicationnel inférant un lexique mental qui s'enracinerait lui-même dans des universaux cognitifs donnant accès à une vérité-évidence. La « sagesse des foules » dont procéderait le PageRank de Google signifierait que, selon Adam Bosworth cité par Ertzscheid, pour tout élément donné (texte, image, document), on aurait une série de mots et de termes composant le plus petit lexique commun (expression d'un consensus conceptuel) permettant de décrire l'objet ou le document. Ainsi, l'indexation sociale, dont FlikR par exemple serait le parangon, consiste à rechercher le consensus sur des valeurs phénoménologiques (opinions consensuelles) qui présupposent qu'on tient le monde pour un « pan-catalogue » d'objets. Peut-on être plus référentialiste ?

D'une certain manière, cette vérité-évidence, qu'elle soit dictée par des experts ou par la sagesse des foules, demeure l'affectation de valeurs subjectives à des contenus et procède en conséquence du prescriptif, quand même ses prescriptions sont de nature différentes, fortes lorsqu'elles sont émises par une autorité experte et faibles lorsqu'elle sont induites par des stéréotypes partagés. Ainsi, on rapporte le jugement d'individus ou de collectivité d'individus sur les contenus plus qu'on ne les décrit dans leurs contextes sociaux et culturels.

2 Faire parler les textes

2.1 La pertinence en jeu

La dichotomie objectivité *vs* subjectivité qui présuppose l'existence de « normes scientifiques » actualisées par des méthodes, des standards et des pratiques devient à son tour un critère déterminant de démarcation entre « bonne » et « mauvaise » science. Au-delà des enjeux juridiques et économiques sous-jacents à ce débat, nous nous intéressons à sa dimension épistémologique. Cette dernière est en effet déterminante si l'on considère les textes comme lieux de production de l'information. Plus particulièrement, la catégorisation des genres textuels (par exemple scientifique *vs* vulgarisé) pose directement la question de la possibilité de discriminer les textes scientifiques et pseudo-scientifiques. Autrement dit, y a-t-il des caractéristiques formelles stables et généralisables qui permettent de distinguer un texte scientifique d'un texte pseudo-scientifique? A priori, la présence de tableaux statistiques ou d'indices de quantification et de bibliographie (parmi d'autres traits) semble caractérisante de textes présentant une valeur de vérité forte. Or, la fabrication d'une argumentation pseudo-scientifique consistera précisément à exhiber ces indices, parmi d'autres, de telle sorte qu'il sera impossible de trancher tant la conformité à la forme attendue est confondante. La question du vrai/faux, qu'on la considère comme pastiche ou sorte de spam, invite à prendre la textualité au sérieux. Le cas limite du « faux » – problème général posé aujourd'hui au Web – impose que l'on s'appuie sur une sémantique des textes élaborée, tant il est vrai qu'une liste finie de mots clés (concepts homologués du domaine) et de procédés rhétoriques externes (figures de style obligées) ne suffisent pas pour produire une analyse des textes suffisamment pertinente.

Monique Slodzian, Mathieu Valette (2009) « Connaissances prescrites ou connaissances décrites ? L'apport de la sémantique des textes », Patrimoine 3.0, *Actes du 12e Colloque International sur le Document Electronique (CIDE.12)*, Khaldoun Zreik, dir., Europia Productions, Paris, pp. 129-141.

version soumise [24 juillet 2009] – merci de se reporter à la version publiée

S'il est vrai, comme le suggère Gloria Origgi, que « la vérification directe de l'information n'est tout simplement pas possible à des coûts raisonnables », ce passage à une ère d'informations de vérité faible est porteur de risques socioculturels incommensurables. Face à la crise annoncée, des outils opératoires nouveaux doivent être proposés, faisant appel à des approches transdisciplinaires demeurées à la lisière des travaux sur l'ingénierie des connaissances. En proposant la description de parcours interprétatifs assignant un ou plusieurs sens à un texte, la sémantique des textes, ouverte au document dans la perspective du numérique (RTP.DOC, 2006), affirme sa capacité à tracer et hiérarchiser les subjectivités qui traversent les textes et, en cela, à assumer leur part d'objectivation.

Par objectivation nous ne supposons pas une extraction immédiate de connaissances déliées des textes et de leurs interprétations possibles, comme le suggèrent les approches prescriptives en produisant des listes de mots censés livrer sans médiation les connaissances d'un texte. Nous posons au contraire la nécessité de passer par des procédures d'analyse pour faire émerger et caractériser les connaissances d'un texte en tenant compte de ses conditions de production et d'interprétation (ordre herméneutique), si l'on veut assurer leur pertinence par rapport à une tâche donnée.

Cette approche impliquant l'ordre herméneutique est incompatible avec la philosophie sous-jacente à l'Internet des objets qui se réduit à l'ordre référentiel ou, au mieux, à l'ordre communicationnel. Il y a là un débat de fond à mener.

2.2 La sémantique du document dans les SOC

La notion de document, défini comme "une artefact médiateur à dominante sémiotique inséré dans des flux transactionnels" qui nous vient des STIC (Zacklad *et al.*, 2007) s'accompagne d'une vision ouverte de l'ingénierie des systèmes d'information à partir d'une réflexion nouvelle sur le processus de documentarisation. La théorie du document qui en émane met en avant « la recherche d'une complémentarité entre SOC hétérogènes, impliquant un rapprochement plus grand entre champs et secteurs différents ». On y trouve une invitation à construire une approche unifiée des espaces sémiotiques ouverts par les TIC, à partir de la notion de co-production sémiotique.

Le processus de documentarisation ainsi décrit propose un couplage texte/document où les approches de la sémantique interprétative peuvent trouver leur légitimité, en même temps qu'elles s'y verront confrontées à une dimension sémiotique nouvelle susceptible de renouveler le concept de texte. Il s'agira en particulier de voir comment des approches relevant respectivement d'une sémiotique du document et d'une sémantique du texte peuvent converger.

Nous tenterons maintenant de démontrer la possibilité de cette convergence en soumettant quelques propositions méthodologiques susceptibles d'intéresser ceux qui, dans la communauté STIC, partagent avec nous une vision « constructiviste » des connaissances et confèrent au texte/document un statut herméneutique en rupture avec les descriptions strictement référentielles.

3 Le texte comme système d'organisation des connaissances ?

Dire que le texte est un SOC introduit un débat entre linguistique et ingénierie des connaissances. En effet, si la pratique de l'extraction de terminologies ou d'ontologies à partir de textes donne à penser que le texte est un espace de collecte privilégié, il serait faux de le considérer seulement comme le terrain d'actualisation des concepts : les concepts ne préexistent pas aux textes, ils sont des îlots, des zones stables de sens construits, élaborés dans les textes et par les textes. C'est pourquoi la textualité exerce des contraintes fortes sur l'élaboration des concepts.

D'une manière générale, la production et l'interprétation des textes sont soumises à des contraintes tant linguistiques que socioculturelles. Ainsi, les discours et les genres textuels configurent les textes en constituant des ensembles de règles de production et d'interprétation acquises ou apprises, parfois de manière inconsciente.

Par exemple, les chercheurs en médecine, eux-mêmes médecins, sont susceptibles de produire, à partir du même contenu informationnel, différents discours : le discours scientifique (à l'attention des chercheurs) ; le discours de la presse médicale (à l'attention des praticiens) et le discours de prévention (à l'attention des patients). Ainsi, au syntagme subtantival *« prise de poids »*, on opposera dans certains textes institutionnels « la forme verbale

Monique Slodzian, Mathieu Valette (2009) « Connaissances prescrites ou connaissances décrites ? L'apport de la sémantique des textes », Patrimoine 3.0, *Actes du 12e Colloque International sur le Document Electronique (CIDE.12)*, Khaldoun Zreik, dir., Europia Productions, Paris, pp. 129-141.

version soumise [24 juillet 2009] – merci de se reporter à la version publiée

« grossir ». Plutôt que « *surcharge pondérale »*, on lira par exemple sur un forum de discussion *« être ronde »*. En bref, les genres textuels organisent différemment la connaissance et à chaque pratique correspondent des genres particuliers. La prévention contre le tabagisme est fortement médicalisée dans les textes institutionnels, elle ne l'est que marginalement dans les forums de discussion dont l'objectif est pourtant identique[3].

Dans les textes spécialisés, le genre choisi sélectionne les concepts et les organise en fonction de contraintes textuelles précises. D'une certaine manière, il décide de son niveau de spécialisation en éliminant certains concepts et en en privilégiant d'autres. Par exemple, un texte médical sur le tabagisme utilisera le concept hyperonymique *tabac* pour *« cigarette »*, *« pipe »*, *« cigare »*, *« narghilé »*, etc. tandis qu'un texte de vulgarisation privilégiera les hyponymes en fonction de leur cible (« *cigarette »*, *« tabac à rouler »*) et de l'ethos du lecteur supposé (un fumeur de cigarette n'est pas un fumeur de cigare). D'une manière générale, des analyses statistiques révèlent que le texte institutionnel construit un discours distancié, intellectuel quand le texte informel est davantage *incarné* ; en forçant le trait, on peut dire qu'il faut de la *volonté* pour s'arrêter dans un texte institutionnel (c'est-à-dire une faculté intellectuelle) et du *courage* dans un texte informel (c'est-à-dire une actualisation sensible de la volonté)

De ces exemples rapides, on conclura que les textes relèvent d'une planification de l'information. Cette planification est différentielle dans la mesure où les textes explicitent et organisent des connaissances apparentées de manières différentes. On peut en conséquence se risquer à allouer au textuel le statut de *système d'organisation des connaissances*.

3.1 Textes, informations et connaissances différentielles

Pour illustrer notre propos, nous proposons d'étudier brièvement différents discours de prévention contre le tabagisme. Le projet général vise notamment les tabacologues et a pour objectif de mieux connaître les pratiques tabagiques. Pour cela, nous étudierons ici un corpus composé de deux ensembles : (a) un discours institutionnel composé de sites médicaux (*Ligue contre le cancer*), de sites de lobbying (*OFT*) et de site de prévention du tabagisme (*Pataclope*, qui s'adresse aux adolescents) ou d'aide au sevrage (*OFT*) et (b) un discours informel, constitué de blog et de forums contre le tabac, sur le sevrage tabagique (*Atoute*).

Sans entrer dans le détail d'une analyse textométrique qui n'est pas ici notre propos, nous tâcherons dans les paragraphes suivant de proposer des grilles interprétatives générales destinées à mieux circonscrire d'un point de vue linguistique les différences de traitement de l'information et d'organisation ou de production des connaissances, pour une thématique semblable, dans ces deux types de discours. Nous aborderons tour à tour les statuts macroscopiques du texte, de l'information et de la connaissance.

3.1.1 Le statut du texte

	Sites institutionnels	**Blogs et forums**
Statut	Objectif (*« les fumeurs »*)	Subjectif (*« moi je »*)
Zone anthropique	Distal (*« le tabac »*)	Identitaire et proximal (*« une cigarette »*)
Fonction	Exposition	Construction

Figure 1. Statut différentiel des textes des deux sous-corpus

On observe que les sites institutionnels adoptent une perspective qui se présente comme objective. Ils mettent à distance l'objet. Ainsi, les différents actants des textes sont par exemple *« les fumeurs », « le fumeur », « le tabac », « la nicotine »*, des entités abstraites correspondant éventuellement à des positions ontologiques. A l'inverse, les forums privilégient la subjectivité. On y relève de nombreux marqueurs identitaires et de coordonnées spatiotemporelles, tels que les pronoms personnels, des déictiques (*« moi », « je »*). Le tabac ou les

[3] Ces observations proviennent d'études réalisées dans le cadre du projet ANR-07-MDCO-002 C-MANTIC destiné à élaborer des méthodologies et des outils pour l'application de la sémantique de corpus au filtrage des masses documentaires.

substances sont peu actualisées, on leur préfère les objets *« clope »* ou *« cigarette »* qui correspondent à une pratique concrète (après le repas, on fume *une* cigarette, pas *du* tabac). La cigarette, enfin, est un objet personnel qui relève de l'identitaire ou, dans le cas du tabagisme socialisant, du proximal (la relation de soi à l'autre). Le tabac est à l'inverse vu comme une plante, une substance, sa conceptualisation est scientifique donc distale – elle ressortit à une mise à distance.

Enfin, la fonction des textes institutionnels est exposante. Ils exposent les risques liés au tabagisme sur un mode dysphorique (*« cancer »*, *« maladie »*, etc.) tandis que les textes informels sont davantage dans la construction d'un savoir, l'élaboration d'une connaissance à partager.

3.1.2 Le statut de l'information

	Sites institutionnels	**Blogs et forums**
Statut	Sanctionnée Haut niveau	Débattue Bas niveau

Figure 2. Statut différentiel de l'information dans les deux sous-corpus

L'information des sites institutionnels est sanctionnée par le corps médical, elle est dite de « haut niveau », les produits de substitution présentés sont par exemple ceux valider par la recherche médicale (ou pharmaceutique) : *« substitut nicotinique »*, *« patch »* ; *« aide médicale »*, *« consultation tabacologique »*, etc. Dans les forums et les blogues, l'information est débattue, dialectisée, les classes sémantiques produites sont davantage liées à des pratiques de sevrage qu'à des catégories générales. On pourrait y trouver pêle-mêle *« chewing-gum »*, *« coup de fil à une copine »*, *« verre d'eau »*, *« footing »*, etc.). L'information n'est pas sanctionnée et peut-être considérée comme de bas niveau (par exemple, un internaute rapporte avoir recouru au cannabis comme substitut nicotinique).

3.1.3 Le statut des connaissances

	Sites institutionnels	**Blogs et forums**
Statut	Exposées	Produites
Modèle	Ontologique	Praxéologique
Représentation des connaissances	Mots-clés, concepts	Passages-clés, *formes sémantiques*

Figure 3. Statut différentiel des connaissances dans les deux sous-corpus

Dans les textes institutionnels, les connaissances existent en amont de la production textuelles, elles relèvent de savoirs scientifiques, médicaux voire encyclopédiques construit dans d'autres situations énonciatives, d'autres pratiques sociales (par exemple, dans des articles scientifiques). Les connaissances sont des concepts déjà lexicalisés, des termes (qui correspondent à des mots-clés) et la textualité a pour fonction l'exposition et la mise en relation de ces connaissances. Les textes institutionnels « déploient » des ontologies, ou des mondes conceptuels similaires aux connaissances ontologiques. Dans les textes informels, blogs et forums, les connaissances sont *produites* par le texte. Elles n'existent pas préalablement aux textes mais résultent de l'élaboration ou de la collaboration des auteurs qui construisent des connaissances partagées En rendant compte de pratiques tabagiques et de scénarii de sevrage par exemple, les textes relèvent d'une praxéologie. Les connaissances dès lors ne sont pas données comme préalables à la mise en texte, elles sont élaborées par la textualité et n'accèdent pas à proprement parler au statut de concepts, mais de préconcepts ou de connaissances préconceptuelles suivant des modalités textuelles particulières que nous décrirons dans le paragraphe suivant.

3.2 Entre le texte et le concept, la *forme sémantique*

La tradition linguistique et terminologique privilégie le lexique, et plus particulièrement les groupes nominaux, dans la détermination des concepts. Or, la linguistique, depuis Saussure, pose que le versant psychique d'un signe, le *signifié*, ne se confond pas avec le concept. Un concept, au sens linguistique proposé ici, n'est donc pas systématiquement lié à un signe particulier, il peut s'actualiser dans une *forme sémantique*, c'est-à-dire un

Monique Slodzian, Mathieu Valette (2009) « Connaissances prescrites ou connaissances décrites ? L'apport de la sémantique des textes », Patrimoine 3.0, *Actes du 12e Colloque International sur le Document Electronique (CIDE.12)*, Khaldoun Zreik, dir., Europia Productions, Paris, pp. 129-141.

version soumise [24 juillet 2009] – merci de se reporter à la version publiée

ensemble de valeurs sémantiques systématiquement cooccurrentes et groupées dans différents textes, relativement stabilisées, mais non nécessairement lexicalisé.

Par exemple, si les mots *« tabac »* et *« choix »* sont en cooccurrence dans un texte et *« fumer »* et *« liberté »* dans un autre, on peut avoir deux fois la même forme sémantique composée minimalement des traits sémantiques /fumer/ et /liberté/ (à considérer que ces traits sémantiques sont contenus dans les signifiés de ces différentes unités lexicales). Ainsi dans les deux extraits ci-dessous, la forme sémantique /fumer/+/liberté/ est actualisée de façon différentes :

> Opter pour la consommation du tabac$^{/fumer/}$ relève du choix$^{/liberté/}$ personnel de l'individu. (http://www.orinfor.gov.rw/DOCS/Sante47.htm)
>
> Le citoyen est libre$^{/liberté/}$ de fumer$^{/fumer/}$ ou de ne pas fumer, de manger de la salade si ça lui chante et des rillettes s'il en a envie (http://www.le-tigre.net/Fumer-ne-tue-pas.html)

On pourrait évidemment enrichir cette forme sémantique des traits /personne/ ou /humain/ mais cette cooccurrence simple est en soi suffisante. Il ne s'agit pas d'un concept à proprement parler mais d'une connaissance préconceptuelle non lexicalisée susceptible de se stabiliser. Cette stabilisation peut mener à un figement lexical (par exemple le syntagme *« liberté de fumer »*) ou à la constitution de connaissances communes partagées.

Selon (Rastier 2008), ce que nous appelons ici connaissance préconceptuelle (ou *préconnaissance*) constitue une connaissance :

> Une connaissance est un ensemble de passages de textes (éventuellement multimédia) : dans leurs récurrences, le contenu de ces passages (les fragments) et leurs expressions (les extraits) sont en relation de transformation, ne serait-ce que par changement de position.
> Résultant de figements et de réductions de syntagmes, les mots sont une sorte très particulière de ces passages, et comme les autres passages, ils restent impossibles à interpréter sans recontextualisation.
> En somme, la connaissance est issue d'une décontextualisation de certaines formes sémantiques saillantes et des expressions qui leur correspondent.

On peut donc avancer qu'un concept est une forme sémantique lexicalisée. Mais un concept, au sens textuel que nous défendons ici, ne correspond pas forcément à une unité lexicale. La lexicalisation d'une forme sémantique, qui aboutit à la formation du concept, ne doit pas être envisagée exclusivement comme sa naissance, ni même comme l'aboutissement de la conceptualisation. Elle s'apparente davantage à un état de stabilisation provisoire, correspondant à un usage circonscrit d'un point de vue socioculturel et temporel.

4 Conclusion

En tant que Système d'Organisation des Connaissances, les textes ne permettent pas de niveler les connaissances ni des les laisser indifférenciées. Les ontologies, en globalisant les connaissances, les dé-situent, au détriment de la pertinence.

Par ailleurs, le passage du texte au document renforce la dimension sémiotique. C'est la congruence des dimensions sémantique et sémiotique qui est à même de susciter de nouveaux débats féconds sur des méthodes alternatives de constitution de connaissances à partir des textes.

Jusqu'ici, l'ingénierie des connaissances et la terminologie textuelle notamment se sont plus particulièrement consacrées à l'extraction de candidats termes dans les textes pour les expertiser et les valider ou non comme concepts ou termes. Elles se sont peu intéressées en revanche à l'émergence de ces concepts dans les textes dès lors que l'on admet que les textes en sont les lieux de production et pas seulement ceux de leur exposition. Nous faisons l'hypothèse qu'avant d'accéder au statut de signes dont les signifiés sont normés (les termes), les concepts émergents se manifestent dans les textes comme formes sémantiques qui se coaguleront ou non en unités lexicales nouvelles et en termes. On peut dès lors considérer que ces formes sémantiques ont valeur de préconcepts. L'enjeu pour la linguistique est d'être capable de décrire et de formaliser ces formes sémantiques et de les requalifier en zones de pertinence. Ce processus d'émergence intéresse d'un point de vue théorique la terminologie et, au plan pratique, l'identification et de détection pour la veille et la constitution de terminologies.

Remerciements

Monique Slodzian, Mathieu Valette (2009) « Connaissances prescrites ou connaissances décrites ? L'apport de la sémantique des textes », Patrimoine 3.0, *Actes du 12e Colloque International sur le Document Electronique (CIDE.12)*, Khaldoun Zreik, dir., Europia Productions, Paris, pp. 129-141.

version soumise [24 juillet 2009] – merci de se reporter à la version publiée

Nous remercions toute l'équipe du projet C-MANTIC, ainsi que Manuel Zacklad et Alain Giboin qui ont sollicité ce débat en organisant l'atelier *Modèles, méthodes, pratiques pour la conception de logiciels basés sur des SOC* à la plateforme AFIA 2009.

Bibliographie

[1] Barlow, J.P., 1994, *A taxonomy of information*, in Bulletin of the American Society for Information Science, 20, 13-17

[2] Coates, E.J., 1978, *Classification in Information Retrieval : Headings and Structure.* London, Library Association

[3] Carnap, R., 1928, *La Construction logique du monde*, trad. fr. Elisabeth Schwarz et Thierry Rivain. Paris : J. Vrin, 2002

[4] Carnap, R., 1934, *La syntaxe logique du langage*, trad.

[5] Ertzscheid, O., 2008, *Indexation sociale et folksonomies: le monde comme catalogue*, Journées ABES.Montpellier, 20 et 21 mai 2008, http://wwwaffordance.info

[6] Origgi, G., 2008, *Le sens des autres. L'ontogenèse de la confiance épistémique* in Raisons Pratiques n°17 – L'épistémologie sociale, Paris, ed.CNRS

[7] Rastier, F., 1995, *Le terme : entre ontologie et linguistique***,** in La banque des mots

[8] Rastier, F., 2002, « Anthropologie linguistique et sémiotique des cultures », in Rastier, F. et Bouquet, S. (éds.), *Une introduction aux sciences de la culture*, Paris, PUF.

[9] Rastier, F., 2007, « Passages », *Interprétation, contextes, codage*, B. Pincemin (éd.), Corpus, 6, 125-152.

[10] Rastier, F., 2008, « Sémantique du Web vs Semantic Web ? Le problème de la pertinence », *Textes, documents numériques, corpus. Pour une science des textes instrumentée,* M. Valette (éd.), *Syntaxe & Sémantique*, n°9, 15-35.

[11] Schlick, 1932, *Forme et contenu. Une introduction à la pensée philosophique*, Agone, Paris.

[12] Valette, M., Slodzian, M., 2008, « Sémantique des textes et Recherche d'information », *Extraction d'information : l'apport de la linguistique*, A. Condamines & Th. Poibeau (éd.), *Revue Française de Linguistique Appliquée*, volume XIII-1 – juin 2008, 119-133.

[13] Wolfram, S., 2001, *A new kind of Science*, Wolfram Media Inc.

[14] Zacklad *et al.*, 2007, *Hypertopic: une métasémiotique et un protocole pour le Web socio-sémantique,* Francky Trichet (Eds), Cépaduès