Analyse comparative de corpus : cas de l'ingénierie des connaissances

Nathalie Aussenac-Gilles, Didier Bourigault, Régine Teulier

Analyse comparative de corpus : cas de l'ingénierie des connaissances

Nathalie Aussenac-Gilles[1], Didier Bourigault[2] et Régine Teulier[3]

[1]Institut de Recherche en Informatique de Toulouse, CNRS - Université Paul Sabatier
aussenac@irit fr
[2]Equipe de Recherche en Syntaxe et Sémantique, CNRS - Université Toulouse Le Mirail
didier.bourigault@univ-tlse2.f
[3]Centre de Recherche en Gestion, CNRS – Ecole polytechnique
teulier@ext.jussieu.fr

Résumé : Dans cet article, nous présentons une analyse de l'évolution du domaine de recherche de l'Ingénierie des connaissances telle qu'elle peut être saisie par l'étude comparative de deux corpus de textes représentatifs du domaine et chronologiquement successifs. La méthode utilisée, une analyse lexicale constrastive, s'appuie sur l'hypothèse que les évolutions du vocabulaire utilisé dans ces deux corpus peuvent être la trace d'évolutions thématiques dans ce domaine de recherche. L'article présente les deux analyseurs utilisés (SYNTEX et UPERY) avant de détailler plusieurs des interprétations construites sur la base de différences de fréquence, de répartition et d'environnement contextuel des termes révélées par les outils.

Mots-clés : Corpus, analyse syntaxique, analyse distributionnelle, analyse contrastive de corpus, application

1 Introduction

Depuis l'émergence de sa problématique vers 1980, le domaine de l'ingénierie des connaissances évolue et fait l'objet de mutations thématiques, méthodologiques et même théoriques. Plusieurs chercheurs français ont entrepris d'en repérer les caractéristiques et les frontières, les objectifs et méthodes de travail dans des articles de synthèse (Charlet et Bachimont, 1998), dans des introductions d'ouvrages (Aussenac *et al.,* 1992) (Aussenac *et al.,* 1996) (Charlet *et al.* , 2000b) et (Teulier, 2003) ou dans des réflexions ayant comme visée de structurer le domaine (Charlet, 2001). Or un des reflets de cette dynamique est (aussi) la publication scientifique. Dans cet article, nous présentons une analyse de l'évolution du domaine de recherche de l'Ingénierie des connaissances telle qu'elle peut être saisie par l'étude comparative de deux corpus de textes représentatifs du domaine et chronologiquement successifs (1995-1998 et 1999-2001). Plus qu'une réflexion fondamentale sur le domaine de l'IC, il s'agit de montrer le type de conclusions que des spécialistes d'un domaine peuvent tirer des résultats d'une analyse systématique menée à l'aide de logiciels de traitement (statistique et linguistique) de ces corpus. Notre point de vue donc est plus pragmatique et expérimental qu'épistémologique, et nous avons bien conscience que

l'expérience mériterait d'être approfondie aussi bien du côté des outils, par une confrontation à des systèmes analogues, que du côté de l'interprétation des résultats, pour en tirer des leçons plus précises sur le domaine (résultats, points forts, limites).

La méthode utilisée est une méthode d'analyse lexicale constrastive, qui s'appuie sur l'hypothèse que les évolutions du vocabulaire utilisé dans ces deux corpus peuvent être la trace d'évolutions thématiques dans le domaine de recherche de l'IC. La méthode vise à repérer les termes dont la fréquence est très nettement différente d'un corpus à l'autre, ce qui révèle l'émergence, la disparition ou la reformulation de certains thèmes de recherche ou de notions, ainsi que les termes, fréquents dans les deux corpus, dont l'environnement textuel est suffisamment différent d'un corpus à l'autre pour que cela soit la marque d'une évolution sémantique ou conceptuelle.

Dans la partie 2, nous décrivons les outils de traitement automatique des langues qui ont été utilisés pour mener à bien l'analyse constrastive des deux corpus : l'analyseur syntaxique de corpus SYNTEX pour l'extraction de mots et de syntagmes, l'outil d'analyse distributionnelle UPERY pour leur caractérisation contextuelle. Nous présentons en partie 3 le corpus utilisé puis des commentaires et interprétations construits sur la base de différences de fréquence, de répartition et d'environnement contextuel des termes révélées par les outils. La partie 4 est consacrée à une discussion des aspects techniques et méthodologiques de la méthode adoptée.

2 Les outils

2.1 SYNTEX : extraction de candidats termes

L'analyseur syntaxique de corpus SYNTEX (Bourigault & Fabre, 2000) a été utilisé pour extraire les mots et groupes de mots de chacun des corpus. SYNTEX prend en entrée un corpus de textes étiquetés[1] (en français ou en anglais), effectue l'analyse syntaxique de chacune des phrases du corpus et produit comme résultat un réseau de mots (noms, adjectifs, verbes, etc.) et de syntagmes nominaux, adjectivaux ou verbaux extraits du corpus. Un syntagme verbal (resp. nominal, adjectival) est un groupe de mots dont la tête syntaxique est un verbe (resp. nom, adjectif). Par exemple, *construire une ontologie* est un syntagme verbal dont la tête syntaxique est le verbe *construire* et l'expansion le nom *ontologie* ; *modèle de l'activité* est un syntagme nominal dont la tête syntaxique est le nom *modèle* et l'expansion le nom *activité* ; *système coopératif* est un syntagme nominal dont la tête syntaxique est le nom *système* et l'expansion l'adjectif *coopératif*. Dans le réseau construit, dit « réseau terminologique », chaque syntagme est relié d'une part à sa tête (lien T) et d'autre part à son (ses) expansion(s) (lien E). Le lien E est étiqueté par le nom de la relation syntaxique de dépendance. Dans le contexte de la construction de ressources

[1] A chaque mot du texte est affecté une catégorie grammaticale (verbe, nom, adjectif, etc.). L'étiqueteur utilisé est le Treetagger, développé à l'Université de Stuttgart.

terminologiques ou ontologiques à partir de textes, les éléments du réseau (mots et syntagmes) sont appelés « candidats termes ».

SYNTEX fournit un certain nombre d'informations numériques associées à chacun des candidats termes, en particulier la fréquence (*freq*) et la répartition (*nbart*). La fréquence d'un candidat terme est le nombre de fois que le candidat terme a été repéré dans le corpus par SYNTEX. La répartition est le nombre d'articles différents dans lesquels le candidat terme a été repéré. Précisons d'emblée que, puisque nous nous intéressons ici aux évolutions de fond du domaine de l'IC, nous avons choisi de faire porter notre analyse sur les termes dont la répartition est suffisamment importante, ceci pour laisser de côté ceux dont la fréquence est éventuellement élevée mais qui ne sont utilisés que par un ou deux auteurs. Nous avons fixé le seuil de *nbart* à 3. Dans la partie 3, nous analyserons les différences entre les deux corpus : quels sont les termes qui sont peu ou pas présents dans l'un des deux corpus et relativement fréquents dans l'autre ? A titre indicatif, nous donnons dans le tableau 1 les syntagmes nominaux les plus fréquents dans chacun des deux corpus. On remarquera l'absence dans la liste du corpus 1 du terme *ingénierie des connaissances* (qui n'y apparaît que 17 fois).

Tableau 1 : les 10 termes les plus fréquents sur les deux corpus (*fréquence*, *nombre d'articles*)

Corpus 1
résolution de problèmes (70,19) ; *acquisition des connaissances* (63,12) ; *connaissances du domaine* (63,14) ; *modèle conceptuel* (54,10) ; *base de connaissances* (54,16) ; *candidats termes* (53,4) ; *représentation des connaissances* (49,14) ; *modèle d'expertise* (47,6) ; *mesure de similarité* (39,3) ; *système d'information* (34,5)
Corpus 2
ingénierie des connaissances (95,15) ; *résolution de problème* (68,10) ; *base de connaissances* (50,11) ; *gestion des connaissances* (48,10) ; *processus de conception* (44,7) ; *activité collective* (32,5) ; *modèle conceptuel* (30,7) ; *bases de données* (28,9) ; *construction d'ontologies* (28,7) ; *recherche d'informations* (28,7)

2.2 UPERY : analyse distributionnelle

UPERY est un outil d'analyse distributionnelle (Bourigault, 2002). Il exploite l'ensemble des données présentes dans le réseau de mots et syntagmes construits par SYNTEX pour effectuer un calcul des proximités distributionnelles entre ces unités. UPERY rapproche deux à deux des candidats termes qui se retrouvent dans les mêmes configurations syntaxiques. Les données de base de l'analyse distributionnelle sont extraites des syntagmes (verbaux, nominaux, adjectivaux) extraits par SYNTEX. Pour un syntagme, constitué d'une tête T, d'une expansion E connectées par la relation R, on construit un couple [TR, E], où TR est lui-même un couple constitué de la tête et de la relation et où E est l'expansion. Par exemple, à partir du syntagme verbal *construire une ontologie*, présent dans le réseau terminologique, est construit le

couple [(*construire*, OBJ), *ontologie*], OBJ désignant la relation « complément d'objet direct ». A partir de ces données, on définit les notions duales de « contexte Tête » et de « contexte Expansion » d'un mot ou d'un syntagme :

- Soit M un mot ou un syntagme. Un contexte Tête de M est un couple (Mt, R) tel que le couple [(Mt, R), M] a été extrait d'un syntagme. Par exemple, le couple (*construire*, OBJ) est un contexte Tête pour le mot *ontologie*, car le syntagme *construire une ontologie* a été extrait par SYNTEX.
- Soit M un mot ou un syntagme et R une relation. Un contexte Expansion du couple (M, R) est un mot Me tel que le couple [(M, R), Me] a été extrait d'un syntagme. On dit encore que Me est un contexte Expansion pour le mot M via la relation R. Ainsi, le mot *ontologie* est un contexte Expansion pour le couple (*construire*, OBJ), ou, autrement dit le mot *ontologie* est un contexte Expansion pour le mot *construire* via la relation OBJ.

Le mot ***outil***		
a pour contexte **Tête**		car SYNTEX a extrait le syntagme (dans au moins 3 articles)
le mot	relation	
Type	DE	*type d'outil*
Permettre	SUJ	*l'outil permet*
construction	DE	*construction d'un outil*
Utilisation	DE	*utilisation d'un outil*
Utiliser	OBJ	*utiliser un outil*
Développer	OBJ	*développer un outil*

Tableau 2 : contextes Tête du mot *outil* dans le corpus 1

Le mot ***outil***		
a pour contexte **Expansion**		SYNTEX a extrait le syntagme (dans au moins 3 articles)
le mot	relation	
informatique	ADJ	*outil informatique*
Analyse	DE	*outil d'analyse*
Aide	DE	*outil d'aide*

Tableau 3 : contextes Expansion du mot *outil* dans le corpus 1

Le module UPERY calcule pour chaque mot ou syntagme la liste de ses contextes Tête et de ses contextes Expansion. Dans notre expérience, n'ont été exploités que les syntagmes qui apparaissent dans au moins 3 articles. Dans la partie 3.4, nous étudierons les différences entre les deux corpus : quels sont les termes qui sont présents dans les deux corpus, mais avec des contextes Tête ou Expansion très différents. La méthode mise en œuvre ici pour repérer ces fluctuations, joliment qualifiées d' « ondoiements » dans (Habert *et al.*, 2002), consiste à comparer pour chaque terme les listes des contextes Tête et Expansion qu'il possède dans les deux corpus. Si ces contextes sont sensiblement différents d'un corpus à l'autre, nous faisons l'hypothèse que c'est un indice d'évolution sémantique. Cette méthode ne s'applique bien entendu qu'aux termes qui ont suffisamment de contextes dans l'un ou l'autre des deux corpus. La comparaison se fait sur la base de deux coefficients de proximité, l'un (proxT) pour la comparaison des contextes Tête et l'autre (proxE) pour la comparaison des contextes Expansion. Le coefficient proxT est calculé ainsi : soit un terme t. Le module UPERY a calculé la liste LcT_1 des contextes Tête du terme t dans le corpus 1, et la liste LcT_2 des contextes Tête du terme T dans le corpus 2.

Soit n_1 le nombre de contextes Tête du terme t dans le corpus 1 (n_1=card(LcT_1)), n_2 le nombre de contextes Tête du terme t dans le corpus 2 (n_2=card(LcT_2)), et a le nombre de contextes Tête du terme t communs aux corpus 1 et 2 (a=card($LcT_1 \cap LcT_2$)). Le coefficient de proxT(t) est égal au ratio entre le nombre de contextes partagés et le nombre total de contextes différents :

$$proxT(t) = a /(n_1+n_2-a)$$

Le coefficient proxE se calcule de façon analogue à partir des contextes Expansion. Ces coefficients sont utilisé dans la section 3.4.

3 Résultats et interprétations

3.1 Les corpus

Le premier corpus est constitué des 35 articles qui composent l'ouvrage Ingénierie des Connaissances édité par J. Charlet, M. Zacklad, G. Kassel et D. Bourigault en 2000 (Charlet *et al.*, 2000). Ces articles ont été sélectionnés par les éditeurs dans les actes des « Journées Acquisition des Connaissances » (JAC) de 1995 et 1996, et des « journées Ingénierie des Connaissances » (IC) de 1997 et 1998. Le nombre de mots de ce corpus est de 138 000 (soit environ 3 940 mots par article).

Le second corpus est constitué des 21 articles qui composent l'ouvrage à paraître édité par R. Teulier, J. Charlet et P. Tchounikine, (Teulier et al., 2003)et qui, selon le même principe que l'ouvrage précédent, rassemble des articles des éditions 1999, 2000 et 2001 de la conférence « Ingénierie des connaissances ». Ce corpus comprend 115 000 mots (soit environ 5 475 mots par article).

3.2 Interface pour l'analyse

Les résultats des outils présentés dans la partie 2 sont comparés entre les deux corpus. Cette analyse contrastive nous aide à repérer des évolutions thématiques ou conceptuelles entre les deux corpus. Ces résultats ne sont que des guides : nous utilisons notre connaissance du domaine ainsi que le retour aux articles eux-mêmes pour interpréter les différences révélées par les outils et décider si elles sont pertinentes ou non, si elles sont révélatrices d'évolutions de fond du domaine, ou simplement le résultat conjoncturel de la sélection des articles des corpus, d'erreurs d'analyse des outils, de tics de langage des auteurs, etc. Pour mener à bien une telle analyse dans des conditions maximales d'efficacité, nous avons développé une interface d'analyse spécifique pour ce type d'analyse contrastive. Celle-ci permet de choisir le type de comparaison (fréquences, contextes Tête ou Expansion), de fixer certains seuils pour la comparaison et d'accéder à tout moment aux occurrences des termes dans les corpus. Elle a été améliorée au fur et a mesure de l'étude.

3.3 Comparaison des fréquences

3.3.1 Résultats bruts : termes en hausse, termes en baisse

Les évolutions ne peuvent se mesurer que par rapport à un fond de stabilité. Nous donnons ci-dessous la liste des syntagmes nominaux dont la fréquence et la répartition sont élevées dans les deux corpus. Chaque terme est accompagné de 4 coefficients (freq1, freq2, nbart1, nbart2) : fréquence dans le corpus 1, fréquence dans le corpus 2, nombre d'articles dans le corpus 1, nombre d'articles dans le corpus 2.

Liste des syntagmes nominaux massivement présents dans les deux corpus *:* *activité de conception* (10,16,3,4); *base de cas* (19,17,5,4); *base de connaissance* (54,50,16,11); *base de donnée* (28,28,10,9); *concept défini* (12,12,4,4); *construction de modèle* (25,12,8,4); *cycle de vie* (14,16,3,6); *domaine d'application* (22,23,10,8); *génie logiciel* (19,16,8,3); *ingénieur de connaissance* (27,10,6,7) ; *langage de représentation* (12,11,7,5); *logique de description* (24,16,4,6); *mémoire d'entreprise* (23,11,4,3); *méthode de résolution de problème* (16,21,7,8); *modèle conceptuel* (54,30,10,7); *modèle de tâche* (18,23,7,3); *modèle générique* (29,14,5,3); *modélisation des connaissances* (26,20,13,10); *niveau de abstraction* (21,13,8,8); *ontologie du domaine* (12,16,5,8); *outil d'aide* (16,16,10,7); *prise de décision* (17,11,3,4); *processus de conception* (20,44,6,7); *recherche d'information* (14,28,6,7); *relation sémantique* (17,24,5,4); *représentation des connaissances* (49,23,14,7); *représentation formelle* (13,10,3,3); *résolution de problème* (70,68,19,10); *système d'aide* (21,11,13,3); *système d'information* (34,27,5,9).

Pour relever les évolutions, nous nous intéressons aux termes « en hausse » (la fréquence dans le corpus 2 est très nettement supérieure à celle dans le corpus 1), et aux termes « en baisse » (la fréquence dans le corpus 1 est très nettement supérieure à celle dans le corpus 2). Pour isoler les évolutions les plus radicales, nous fixons des seuils assez élevés : pour les termes en hausse (resp. en baisse), le seuil de la fréquence dans le corpus 2 (resp. corpus 1) est de 50 pour les noms et de 10 pour les syntagmes nominaux, le seuil du nombre d'articles dans le corpus 2 (resp. corpus 1) est de 4 pour les noms et de 3 pour les syntagmes nominaux, et enfin le ratio entre la fréquence dans le corpus 2 (resp. corpus 1) et la somme des fréquences dans les deux corpus est de 0,80. Les termes (noms et syntagmes nominaux) en hausse et en baisse sont présentés par ordre alphabétique dans les tableaux 4 à 7.

Tableau 4 : Les noms en forte hausse (freq2>>freq1 ; freq2 >=50 ; nbart2>=4)

*Annotation (*11, 72, 6, 6) ; *IC (*0, 59, 0, 6) ; *Index (*17, 81, 6, 4) ; *Ingénierie (*21, 180, 11, 16) ; *Métier (*19, 78, 8, 9) ; *Page (*7, 77, 3, 6) ; *Personne (*25, 103, 11, 11) ; *Veille (*1, 59, 1, 4) ; *Web (*0, 92, 0, 10)

Tableau 5 : Syntagmes nominaux en forte hausse (freq2>>freq1 ; freq2 >=10 ; nbart1>=3)

*aide à conception (*3, 12, 3, 3) ; *architecture logicielle (*0, 11, 0, 3) ; *axe sémantique (*0, 10, 0, 3) ; *création de connaissance (*1, 12, 1, 4) ; *ingénierie de connaissance (*17, 95, 10, 15) ;

livre de connaissance (0, 28, 0, 3) ; *modèle de organisation* (1, 18, 1, 3) ; *nom propre* (0, 14, 0, 3) ; *page Web* (0, 27, 0, 6) ; *projet en cours* (2, 13, 2, 3) ; *relation entre terme* (0, 16, 0, 3) ; *représentation de l'ontologie* (0, 12, 0, 4) ; *site Web* (0, 10, 0, 4) ; *source d'information* (2, 17, 2, 4) ; *Web sémantique* (0, 16, 0, 4)

Tableau 6 : Les noms en forte baisse (freq1>>freq2 ; freq1 >=50 ; nbart1>=4)

Agent (109, 22, 13, 6) ; *Algorithme* (62, 12, 10, 6) ; *Cogniticien* (51, 11, 10, 3) ; *Composant* (56, 13, 13, 7) ; *Diagnostic* (57, 14, 10, 5) ; *Espace* (74, 18, 9, 6) ; *Evénement* (52, 12, 11, 6) ; *Expertise* (177, 31, 14, 9) ; *Généralisation* (82, 9, 10, 3) ; *Graphe* (54, 4, 12, 3) ; *Incident* (118, 0, 5, 0) ; *KADS* (57, 4, 5, 3) ; *Opérateur* (143, 10, 15, 5) ; *Opération* (87, 10, 21, 7) ; *Plan* (198, 30, 17, 11) ; *Primitive* (89, 21, 8, 8) ; *Procédure* (82, 15, 17, 9) ; *Session* (78, 0, 4, 0) ; *Unité* (109, 7, 15, 6)

Tableau 7 : Les syntagmes nominaux en baisse (freq1>>freq2 ; freq1 >=10 ; nbart1>=3)

acquisition de connaissances (63, 13, 12, 10) ; *connaissance de cas* (10, 0, 3, 0) ; *contexte partagé* (25, 0, 3, 0) ; *graphe conceptuel* (22, 2, 4, 2) ; *information contenue* (10, 0, 5, 0) ; *instance de classe* (10, 1, 4, 1) ; *mesure de similarité* (39, 6, 3, 2) ; *méthode de résolution* (30, 0, 9, 0) ; *mise en correspondance* (15, 2, 3, 2) ; *mode opératoire* (13, 0, 3, 0) ; *modèle d'expertise* (47, 7, 6, 2) ; *modèle de raisonnement* (24, 3, 5, 2) ; *mot clé* (23, 2, 4, 2) ; *objet du domaine* (17, 3, 6, 3) ; *primitive de modélisation* (13, 2, 3, 2) ; *problème à résoudre* (12, 3, 6, 2) ; *processus d'acquisition de connaissances* (12, 1, 5, 1) ; *réalisation de système* (11, 2, 6, 2) ; *représentation de cas* (14, 0, 4, 0) ; *réseau de distribution* (11, 0, 3, 0) ; *réseau sémantique* (11, 0, 8, 0) ; *structure arborescente* (10, 2, 3, 1) ; *structure conceptuelle* (10, 0, 3, 0) ; *système coopératif* (18, 0, 4, 0) ; *travail coopératif* (18, 1, 4, 1)

L'analyse ne peut pas s'appuyer sur une simple observation de liste de termes extraits par mesures statistiques. Il convient d'analyser ce que représentent les chiffres pour construire une interprétation fiable. Chaque donnée de la comparaison doit être scrutée via une analyse des contextes d'occurrences des termes. L'examen des contextes révèle dans certains cas les limites de la méthode, et explique la difficulté à interpréter a priori certaines différences détectées par l'analyse automatique. Par exemple, la présence de *méthode de résolution* dans le tableau 7 n'est qu'un artefact de la méthode. Elle est due au fait que SYNTEX a analysé de façon légèrement différente dans les deux corpus le terme *méthode de résolution de problème*, qui y est également présent, en extrayant le sous-terme *méthode de résolution* dans le premier et pas dans le second car il n'y est pas utilisé seul. De façon générale, le retour systématique aux contextes est nécessaire pour vérifier que des phénomènes de polysémies ne viennent pas biaiser les résultats statistiques (cf. infra).

3.3.2 Une analyse des résultats : catégorisation

Nous nous intéressons ici aux termes (noms et syntagmes nominaux) qui sont répartis à la fois dans les deux corpus, c'est-à-dire présents dans au moins 3 articles dans chacun des deux corpus. Ils sont au nombre de 830. Par un parcours rapide mais

systématique de la liste de ces termes, on peut voir apparaître des lignes de force ou tendances. La pertinence, au sens de Sperber et Wilson (1989), lors de cette première lecture, sera générée par l'association d'une évolution et d'un regroupement de termes pouvant constituer un sous-groupe identifiable du domaine. Il s'agit d'abord de « faire parler » le corpus dans sa globalité, en acceptant une approche intuitive. Le propre de cette première phase est d'aller vite et sans a priori, ce qui peut sembler contradictoire avec une démarche d'ingénierie, en effet : « l'effort de traitement est un facteur négatif : toutes choses étant égales par ailleurs, plus l'effort de traitement est grand, plus la pertinence est faible » (Sperber et Wilson, 89).

On constitue ainsi des classes qui se définissent mieux les unes par rapport aux autres au fur et à mesure de la progression dans la liste. Elles « apparaissent » comme apparaissent les nuages de points dans l'analyse factorielle en composantes principales. On prend soin de définir une classe « réserve » de termes pouvant être utiles mais qu'on ne sait pas affecter dans l'immédiat. Cette classe évite ensuite de parcourir à nouveau la liste globale qui comporte du bruit. Il est probable que ces classes se constituent aussi en navigant autour d'éléments saillants ou prototypiques (Rosch, 1978). En constituant ces classes, on cherche à faire apparaître, outre les lignes de forces ou évolutions qui traversent le domaine, les disjonctions ou les recouvrements des corpus (dans notre cas, la stabilité puisqu'il s'agit de corpus qui se succèdent dans le temps). Cette étude fait aussi ressortir des points remarquables qui feront l'objet d'une analyse plus approfondie à l'aide des contextes Tête et Expansion.

Un terme polysémique et appartenant au langage courant n'apparaîtra intéressant qu'à cause de fréquences très différenciées dans les deux corpus, ce qui laisserait entendre qu'il a un sens différent de son sens le plus courant. Par exemple, le terme *contenu* a 51 occurrences dans le corpus 1 et 116 dans le corpus 2. Après lecture et validation de ses occurrences, nous en retenons 24 dans le corpus 1 et 86 dans le corpus 2 pour lesquelles il a un sens propre au domaine. La rectification de ces chiffres n'a actuellement pas d'impact sur les divers coefficients statistiques calculés, alors qu'ils seraient plus pertinents à prendre en compte.

Les différentes étapes de la première phase du travail sont donc les suivantes : 1) constituer les classes, 2) positionner les classes les unes par rapport aux autres, en créer de nouvelles, 3) aller dans les classes, enlever certains éléments, regrouper ou éclater les classes trop disproportionnées, 4) rectifier le nombre d'occurrences de chaque terme en allant voir les phrases.

D'après nos observations, une classe regroupant moins de 10 termes a peu de sens pour pouvoir exprimer une tendance. Inversement, il est intéressant de faire apparaître le plus de classes possibles, dans la limite de ce qui est cognitivement manipulable, et il ne semble pas utile que celles-ci soient composées de trop nombreux éléments : nous n'avons pas pu relever de pertinence supplémentaire au-dessus de 40 termes. Les classes que nous avons identifiées sont reportées dans le tableau 8, et le tableau en présente un exemple, la classe 1. Les classes de 1 à 33 expriment une tendance : elles sont disjointes et regroupent 408 termes classés sur les 830 termes proposés par l'interface (avec les paramètres préalablement décrits). Les classes qui se réfèrent à un thème d'intérêt, constituées en faisant abstraction de toute

tendance d'évolution, et choisies ici à titre d'exemple, sont les classe A à C. L'interface permet de parcourir la liste des termes et de les affecter à une ou plusieurs classes.

Tableau 8 : Les classes regroupant des termes dont la tendance peut être lue conjointement

classe	Intitulé du regroupement opéré sous la classe	Noms ou Syntagmes
1	Recueil , acquisition des connaissances en baisse	19
2	Vocabulaire de l'informatique en baisse	36
3	Développement de SBC en baisse	39
4	Terminologie, ontologies en baisse	24
5	Ingénierie des connaissances, plutôt stable	25
7	Terminologie, ontologies en hausse	25
8	Vocabulaire de l'informatique plutôt stable	34
9	IC autre que terminologie en hausse	62
33	Réserve, termes non affectés	144
total	Total de termes classés sur 830 termes candidats	408
A	Usage, outils, utilisateurs	7
B	Activité, activité coopérative	10
C	Gestion des connaissances, KM	4

Tableau 9 : La classe 1 regroupe les termes qui se réfèrent au recueil et à l'acquisition des connaissances et qui sont en baisse (termes classés par ordre décroissant de fréq1)

Classe 1 : recueil, acquisition des connaissances	Freq1	Freq2	Ndoc1	Ndoc2	f12	f21
expertise	177	31	14	9	0,85	0,14
expert	160	42	24	10	0,79	0,20
raisonnement	123	58	21	12	0,67	0,32
acquisition	90	39	18	12	0,69	0,30
savoir	84	47	13	8	0,64	0,35
acquisition de connaissance	63	13	12	10	0,82	0,17
connaissance de domaine	63	20	14	11	0,75	0,24
diagnostic	57	14	10	5	0,80	0,19
cogniticien	51	11	10	3	0,82	0,17
savoir-faire	43	25	12	6	0,63	0,36
constitution	38	18	12	8	0,67	0,32
planification	34	10	8	5	0,77	0,22
paradigme	25	9	4	5	0,73	0,26
expert de domaine	21	7	10	5	0,75	0,25
entretien	20	12	9	3	0,62	0,37
extrait	14	5	8	3	0,73	0,26
protocole	12	5	5	5	0,70	0,29

heuristique	8	5	6	3	0,61	0,38
analyse de activité	7	4	5	3	0,63	0,36

Ce sont la juxtaposition de ces termes et leurs fréquences en baisse qui sont parlantes et qui confirment ce qu'un observateur averti du domaine peut dire sans une analyse aussi fondée et quantitative que celle-ci. La baisse des termes se référant à l'expertise (*expert, expertise, expert de domaine)*doit être lue en même temps que la baisse de *acquisition de connaissance*, de *cogniticien* et de *entretien*. Moins d'auteurs traitent, dans le corpus 2 mais aussi généralement dans le domaine, du recueil de connaissances d'experts. Par contre, la baisse de *connaissance du domaine* doit être interprétée différemment : c'est justement parce qu'on ne s'intéresse plus qu'aux connaissances du domaine, qu'on ne précise plus que ce sont les connaissances du domaine, ou qu'on parle de préférence *d'ontologie* (en hausse dans la classe 7) ; cette baisse est à rapprocher de celles de *diagnostic*, *raisonnement* et *savoir-faire* représentant l'autre catégorie de connaissances qui a connu un moindre intérêt.

Tableau 10 : Classe 7 : les termes se référant à la terminologie et en hausse, classés selon freq2 décroissante

Classe 7 : terminologie, ontologies en hausse	Freq1	Freq2	ndoc1	ndoc2	f12	f21
ontologie	200	405	11	13	0,330	0,66
contenu	51	116	18	13	0,305	0,69
groupe	76	107	19	14	0,415	0,58
catégorie	61	94	15	11	0,393	0,60
nom	43	81	15	12	0,346	0,65
index	17	81	6	4	0,173	0,82
annotation	11	72	6	6	0,132	0,86
indexation	50	66	7	6	0,431	0,56
syntagme	21	29	4	4	0,42	0,58
construction de ontologie	9	28	5	7	0,243	0,75
définition de concept	9	19	5	3	0,321	0,67
descripteur	8	19	4	3	0,296	0,70
ontologie de domaine	12	16	5	8	0,428	0,57
assertion	3	16	3	3	0,157	0,84
dictionnaire	3	15	3	6	0,166	0,83
langue naturel	6	12	4	3	0,333	0,66
concept de base	4	12	3	3	0,25	0,75
extracteur	3	12	3	3	0,2	0,8
concept de ontologie	8	11	5	3	0,421	0,57
analyseur	4	11	4	3	0,266	0,73
analyse syntaxique	5	9	4	3	0,357	0,64
traitement de langue	5	8	3	6	0,384	0,61
analyse linguistique	8	7	3	3	0,533	0,46
hypertextuelle	5	6	3	3	0,454	0,54

polysémie	3	5	3	4	0,375	0,62

Au sein de la classe 7 (tableau 10), il faut interpréter ensemble *ontologie, construction d'ontologie, concept de ontologie* et *ontologie du domaine*. Globalement le thème des ontologies se développe, ce qui ne fait que corroborer une observation intuitive du domaine. A partir du terme *ontologie de domaine*, les noms et syntagmes apparaissent dans trop peu de textes et avec trop peu d'occurrences pour justifier la lecture de leurs occurrences isolément. Un terme qu'on aurait attendu a priori dans cette classe serait *corpus,* absent car utilisé par trop peu d'auteurs.

Tableau 11 : Classe C : les termes se référant à usage et utilisateur

Classe C : usage, utilisateur	Freq1	Freq2	ndoc1	ndoc2	f12	f21	classe
utilisateur	188	143	21	18	0,56	0,43	8,A
usage	26	42	5	10	0,38	0,61	9,A
système de aide	21	11	13	3	0,65	0,34	2,A
type de outil	5	5	3	4	0,5	0,5	8,A
besoin de utilisateur	4	6	3	4	0,4	0,6	9,A
homme-machine	4	5	3	4	0,44	0,55	8,A
utilisation de outil	4	4	3	3	0,5	0,5	5,A

Le terme de *contenu*, déjà cité, est intéressant. Dans le corpus 1, sur 24 occurrences validées, 7 proviennent du même auteur, lequel génère 31 occurrences sur les 84 validées du second corpus, plus celles d'auteurs de son laboratoire. Un autre auteur majeur sur ce sujet n'est présent que dans le second corpus et génère 39 occurrences. Le traitement des contenus est donc une évolution marquée entre les 2 corpus, alors que cette activité est présente dans la communauté depuis longtemps. L'étude des termes précise qu'elle s'est beaucoup affirmée sur les 3 dernières années, mais elle est portée sous ce nom là essentiellement par deux auteurs majeurs sur ce thème (70 occurrences sur 84 à eux deux). Les expressions dans lesquelles on trouve le mot *contenu* sont les suivantes : le *contenu des textes, des pages Web, audiovisuels, d'enseignement, la numérisation des contenus, la recherche par les contenus, le raisonnement sur les contenus, indexer par le contenu, le routage par le contenu*. Dans le corpus 1, on trouve les *contenus audiovisuels, sémantique, linguistique, des documents, du dossier, textuel, d'information, des images, l'organisation, l'enrichissement, le classement des contenus, l'indexation par le contenu*. Cet exemple fait aussi ressortir la nécessité d'étudier la stabilité ou l'évolution des contextes Tête et Expansion des termes au delà de l'évolution de leur fréquence.

La classe C (tableau 11) illustre la possibilité que donne l'interface de classer chaque terme dans deux classes simultanément. Ici, la dernière colonne mentionne les deux classes auxquelles chaque terme appartient. Ce peut être la classe 33 dite de réserve parce qu'un terme, intéressant du point de vue d'une thématique, n'était pas particulièrement parlant du point de vue des tendances. Nous avons encore peu travaillé ces thématiques. Cependant, on peut remarquer que les termes concernés par les regroupements en thématiques très précises sont moins nombreux que ceux qui apparaissent dans les grandes tendances. En fait, le classement par grande tendance

est presque exhaustif alors que le classement par thématique correspond à un centrage très particulier. Cela signifie aussi que ces thématiques ne sont pas « majeures » dans le domaine, lesquelles sont apparues naturellement au premier balayage et ont été associées à des tendances à la hausse ou à la baisse.

3.4 Comparaison des contextes Tête et des contextes Expansion

Le deuxième mode de comparaison vise à relever les termes qui ont des fréquences analogues dans les deux corpus (freq1 et freq2), mais dont le sens pourrait avoir évolué entre les deux corpus. Nous présentons dans les tableaux 12 et 13 les termes dont les contextes en Tête ou en Expansion ont le plus varié entre les deux corpus, c'est-à-dire ceux pour lesquels les coefficients proxT et proxE sont les plus faibles. Un coefficient nul signifie que le terme n'a aucun contexte identique dans au moins 3 articles et dans les deux corpus.

Tableau 12 : Termes dont les contextes Tête sont très différents entre les deux corpus (n_1, n_2 et a sont le nombre de contextes Tête présents dans au moins 3 articles dans corpus 1, corpus 2 et communs).

Terme	**n_1**	**n_2**	**a**	**proxT**	**freq1**	**freq2**
activité	1	7	1	0,14	204	324
corpus	1	3	0	0	118	138
document	3	6	1	0,12	266	266
modèle	12	5	2	0,13	613	437
niveau	5	0	0	0	297	146
organisation	0	6	0	0	109	205
processus	0	4	0	0	260	252
question	4	1	0	0	72	55
recherche	4	2	0	0	155	117
relation	7	0	0	0	404	247
travail	3	1	0	0	243	253
type	8	0	0	0	375	166

L'analyse de ces données est plus délicate que celles résultant de la comparaison des fréquences, mais elle est aussi plus enrichissante car elle dévoile des phénomènes qui sont difficiles à saisir « à l'œil nu ». Prenons quelques exemples.

- *organisation*. Ce terme est fréquent dans les deux corpus. Mais il ne possède aucun contexte Tête *stable* dans le corpus 1, alors qu'il en présente 6 dans le corpus 2. Les syntagmes correspondants à ces contextes sont les suivants : *type d'organisation*, *modèle de l'organisation*, *connaissance de l'organisation*,

connaissances dans l'organisation[2], *membres de l'organisation*, *acteur d'une organisation*. Cela signifie que, dans ce corpus, le terme *organisation* présente une stabilité syntaxique, et donc probablement sémantique, qu'il n'a pas dans le premier corpus où ses contextes sont certainement plus disparates. Ce phénomène correspond bien à une évolution du domaine et de la conférence IC, qui ont vu croître le nombre de projets en gestion des connaissances (où l'organisation de l'entreprise est cruciale) d'une part, et qui ont bénéficié de plusieurs contributions de chercheurs en sciences de gestion d'autre part.

Tableau 13 : Les termes qui ont des contextes Expansion très différents entre les deux corpus (n_1, n_2 et a sont le nombre de contextes Expansion dans corpus 1, corpus 2 et communs)

Terme	**n_1**	**n_2**	**a**	**proxE**	**freq1**	**freq2**
application	3	1	0	0	181	119
approche	3	5	0	0	167	142
cas	5	0	0	0	367	240
connaissance	13	18	3	0,11	902	1017
description	4	4	1	0,14	205	203
document	2	2	0	0	266	266
élément	4	1	0	0	133	71
étape	3	5	1	0,14	104	170
état	3	1	0	0	65	29
information	3	2	0	0	278	262
notion	3	1	0	0	123	102
partie	6	1	0	0	142	85
structure	10	1	0	0	274	98
utilisation	5	0	0	0	120	113
valeur	5	0	0	0	118	86

- *activité*. De même les contextes Tête du terme *activité* dans le corpus 2 correspondent aux syntagmes non repérés dans corpus 1 : *type d'activité, modèle de l'activité, modélisation de l'activité, traces de l'activité, théorie de l'activité*. Là encore, la mention explicite de l'étude de l'activité correspond bien à une évolution de la production scientifique du domaine, qui ne peut pas être interprétée en bloc comme une évolution du domaine lui-même. En effet, si on observe plus finement le corpus 2, on constate que ces syntagmes ne sont utilisés que par les auteurs venant de l'ergonomie ou des sciences de gestion. Au contraire, la plupart des autres auteurs ne mentionnent pas en tant que telle l'étude de l'activité au sein de laquelle vont être utilisés leurs systèmes ou leurs modèles.

[2] Ce syntagme est le résultat des analyses syntaxiques erronées des syntagmes *gestions des connaissances dans l'organisation*, *structuration des connaissances dans l'organisation*, *intervention sur les connaissances dans l'organisation*.

- *document*. Les phénomènes autour du terme *document* sont également très instructifs. Ce terme a un seul contexte Tête *type de document* et un seul contexte Expansion *documents textuels* communs aux deux corpus. Mais si on rapproche *document* de *documentation*, on trouve aussi *document technique* et *documentation technique* communs aux deux corpus, l'IC s'intéressant de manière constante aux textes techniques en priorité. L'évolution des contextes Tête (*ensemble de ~*, *partie de ~* dans le corpus 1 ; *base de ~*, *contenu de ~*, *indexation de ~*, *retrouver ~*, *annotation de ~* dans le corpus 2) correspond à l'arrivée de termes relatifs à la recherche d'information et à la gestion documentaire, domaines de plus en plus proches de l'IC, en particulier autour de l'étude du Web Sémantique.

- *tâche* : l'étude du terme *tâche* nous intéresse car nous savons a priori qu'il présente une polysémie : *tâche* peut désigner une structure de représentation des connaissances (sens1, à rapprocher de *méthode*) ou désigne la tâche prescrite et réalisée par un individu (sens2). On s'attend à retrouver des indices de ces deux points de vue à l'aide des contextes Tête ou Expansion. Or, tels que présentés dans le tableau 14, ces contextes renvoient plutôt au sens2 avec une évolution qui va du niveau individuel (de l'expert) à celui, global, de l'organisation. En constante, *tâche du domaine* renvoie à la tâche modélisée dans le système. Or si on regarde de plus près les articles parlant de tâche, on retrouve aussi le sens1 comme étant sujet à une évolution (plus présent dans le corpus1 que dans le corpus2), mais seulement chez quelques auteurs. L'analyse distributionnelle n'est pas assez productive au second ordre, une analyse manuelle s'impose donc pour compléter les interprétation.

Tableau 14 : Les contextes du terme *tâche* dans les deux corpus

contexte commun aux 2 corpus	*~ du domaine*, *~ utilisée*, *~ nécessaire*
contextes propres au corpus 1	*~ générale*, *~ médicale*, *~ experte*, *~ de l'expert*
contextes propres au corpus 2	*~ de l'organisation, ~ de l'entreprise, ~ métier, ~ organisationnelle, ~ générique, ~ spécifique, ~ implicite, ~ tacite , ~ terminologique*

4 Discussion

La méthode basée sur les contextes syntaxiques, utilisée dans cette étude, est analogue à celle adoptée par B. Habert (Habert *et al.*, 2002) pour analyser les évolutions de sens de certains mots dans un corpus de textes syndicaux. Ces textes sont les résolutions générales adoptées par les congrès de la CFTC et de la CDFT entre 1945 et 1990. Sur le plan technique, la principale différence tient au type de contextes syntaxiques exploités. Habert utilise les résultats de l'extracteur de syntagmes nominaux LEXTER, et les contextes disponibles sont donc uniquement de type nominal. Grâce à l'usage de SYNTEX, nous disposons ici de contextes nominaux et de contextes verbaux. La caractérisation contextuelle des mots est donc plus large, et les chances de mesurer des évolutions sont donc plus grandes.

L'analyse contrastive est ici particulière, puisqu'elle juge deux périodes rapprochées et une même communauté scientifique sur un seul pays. On peut considérer donc qu'il s'agit de corpus assez homogènes. Parmi les facteurs d'homogénéité, on trouve certains auteurs dans les deux corpus. Sur 75 auteurs du corpus 1, 18 sont présents dans le corpus 2 dont 5 d'entre eux au titre de 3 articles sur l'ensemble des deux corpus. Ceci doit être cependant être relativisé par le caractère pluridisciplinaire de la communauté IC. Ainsi le terme *conception* est utilisé dans le premier corpus par des informaticiens de l'IC qui parlent de la conception de leurs propres systèmes alors que dans le corpus 2, il est utilisé par des ergonomes et mécaniciens qui évoquent l'activité des utilisateurs qu'ils souhaitent assister.

Ces corpus présentent d'autres caractéristiques qui font ressortir au contraire une certaine hétérogénéité propre à l'IC. Ainsi, peu volumineux, ils reflètent la diversité des sujets abordés par les auteurs. Le domaine de l'IC est fortement pluridisciplinaire et les auteurs abordent donc des sujets variés avec des préoccupations et manières de rédiger influencées par d'autres disciplines. L'IC est un domaine relativement jeune, dont les concepts ne pas encore complètement définis et font l'objet de points de vue variés. C'est un domaine actif et en évolution, dont les objets évoluent. De plus, le style « article scientifique » n'impose aucune contrainte rédactionnelle. Notre méthode, appliquée sur ce type de corpus, présente alors des limites. Cette relative disparité syntaxique et sémantique doit être prise en compte pour pondérer l'analyse des résultats, essentiellement statistiques :

- Les contextes tête et expansion sur ces corpus ont permis de produire peu de contextes verbaux : on retrouve essentiellement des syntagmes nominaux ; ceci est normal car les syntagmes nominaux sont bien plus fréquents que les syntagmes verbaux dans les corpus de type scientifique ou technique ;
- Les tendances sont faites en interprétant freq1/(freq1 + freq2) ; or peu de termes présents dans plusieurs documents ont beaucoup d'occurrences ; ce ratio est alors très sensible et doit être interprété avec prudence ; il faudrait le moduler d'un facteur correctif qui rendrait compte de la diversité des fréquences dans l'échantillon étudié. Nos propositions sont donc à prendre comme un exemple de lecture contrastive de corpus permise par les outils présentés, et non comme des résultats absolus.
- On peut se demander comment déterminer la fin du processus : le début est passionnant, motivant pour un expert mais ensuite, le retour aux occurrences une à une est lourd et requière une curiosité sinon une compétence en linguistique ou en terminologie. De plus, le degré de finesse de l'analyse étant laissé à la décision de l'analyste, il semble difficile de fixer des critères de validation objectifs.

La lecture des évolutions du domaine ne doit bien sûr pas se limiter à l'utilisation de la ressource que fournissent ces outils. Une évolution importante du domaine, le web sémantique (Euzénat, 2001) n'apparaît ici que marginalement (mots site Webn page Web et Web sémantique sur le tableau 5), en partie parce que les textes les plus récents du corpus datent de 2001, année à partir de laquelle ces travaux se sont

vraiment développés en France. Pour consolider nos conclusions sur l'IC et poursuivre l'évaluation de l'approche, nous sommes en train de conduire plusieurs études complémentaires. Nous avons prévu d'utiliser d'autres logiciels d'analyse de texte ayant une composante statistique : Tétralogie (Mothe et al., 2001) et Sémio. Ensuite, nous souhaitons confronter les résultats obtenus sur le premier corpus à l'expérience TH(IC)2 menée sur ce même corpus (Bourigault, 2000). Enfin, une perspective plus lointaine serait de comparer l'approche à d'autres outils statistiques comme ALCESTE ou PROSPERO (Chateaureynaud, 2003).

5 Conclusion et perspective

Cette étude nous a permis à la fois de confirmer l'intérêt et les limites d'une étude constrastive de deux corpus à partir de l'analyse des résultats des logiciels d'analyse de texte Syntex et Upéry, et de mettre en forme des tendances relatives au domaine de l'ingénierie des connaissances entre 1990 et 2000. Nous avons ainsi retrouvé des évolutions bien connues du domaine (moins de recherches sur la représentation des connaissances, sur le recueil d'expertise, émergence significative de travaux sur ontologies et terminologies, etc.) mais aussi retrouvé des caractéristiques plus complexes : une disparité disciplinaire qui rend difficile une vraie unité de point de vue, une grande variété de sujets de recherche qui renvoient à autant de centres d'intérêt, une stabilité autour d'un noyau de concepts qui est parfois cachée derrière un renouvellement continu des objets de recherche liés à de nouvelles applications.

Nous avons souligné le lien entre les points forts ou faibles de notre approche et les caractéristiques d'un corpus. Elle nous laisse envisager le spectre des contextes dans lesquels ce type d'analyse semble la plus pertinente : les textes étudiés dans le temps doivent être regroupés en deux corpus, chacun étant relativement homogènes dans la forme, comparables entre eux (même genre textuel par exemple) et suffisamment volumineux et représentatifs du domaine.

Nous avons mis en évidence des limites à l'approche, comme la sensibilité de certains coefficients (comme le ratio de fréquences en cas de fréquence faible), comme le faible nombre de contextes commun à plusieurs documents d'un corpus qui ne permet d'identifier que quelques termes dont les contextes tête ou Expansion peuvent être comparés dans les deux corpus, ou encore le caractère très empirique, très marqué par les connaissances de l'analyste, des résultats. Ces remarques confirment que les résultats des logiciels d'analyse des textes doivent être vus comme des supports possibles parmi d'autres pour mener des études sur des corpus, confirmer des hypothèses ou identifier de nouveaux phénomènes. Il est clair que les interprétations pourraient être complétées par les résultats d'autres logiciels ou par la confrontation à des analyses des domaines étudiés.

Un des intérêts des logiciels utilisés est de permettre de tenir compte de la spécificité de chaque document au sein des corpus, et donc d'analyser finement les comportements des termes. L'interface de consultation des résultats des outils de TAL et d'enregistrement des résultats de l'analyse rend compte de ces différents paramètres. Elle s'avère déjà très intéressante pour organiser les termes en classes

thématiques renvoyant à de grandes tendances d'évolution du domaine. Bien sûr, elle doit être expérimentée sur d'autres domaines et encore améliorée. Ainsi, il semble souhaitable de pourvoir revenir sur les traitements statistiques après un premier dépouillement, et donc de reproduire des sessions de calcul en prenant en compte les premiers regroupements de mots. On pourrait donc imaginer un outil plus interactif, suffisamment simple pour être utilisé directement par celui qui dépouille. Il nous semble en effet primordial de bien comprendre le fonctionnement de l'outil, la nature des différents seuils et coefficients pris en compte, pour mieux interpréter les résultats.

Références

AUSSENAC N., KRIVINE J.-P. et SALLANTIN J. (1992), Editorial *Revue d'intelligence artificielle,* Numéro spécial sur l'acquisition des connaissance, Ed. : Paris : Hermès. Vol 6, N°1/2. p. 7-18.

AUSSENAC N., LAUBLET P., REYNAUD C., (1996). L'acquisition des connaissances, composante à part entière de l'informatique du futur. *Acquisition et ingénierie des connaissances : tendances actuelles.* Toulouse : Cépaduès-Editions, pp 3-25

BOURIGAULT D. & FABRE C. (2000). Approche linguistique pour l'analyse syntaxique de corpus, *Cahiers de Grammaires, n° 25*, 2000, Université Toulouse - Le Mirail, p. 131-151

BOURIGAULT D. (2002) Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus, *Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues* (TALN 2002), Nancy, p. 75-84

CHARLET, J., BACHIMONT B., (1998). De l'acquisition à l'ingénierie des connaissances : applications et perspectives. *Actes des assises nationales 1998 du PRC-I3*. Lyon. 81-84.

CHARLET J, ZACKLAD M., KASSEL G. & BOURIGAULT D. (2000a) Ingénierie des connaissances : recherches et perspectives. *Ingénierie des connaissances. Tendances actuelles et nouveaux défis*. CHARLET J, ZACKLAD M., KASSEL G. & BOURIGAULT D. (eds). Editions Eyrolles/France Telecom, Paris. Chapitre 1, p. 1-22.

CHARLET J, ZACKLAD M., KASSEL G. & BOURIGAULT D. (eds) (2000b) *Ingénierie des connaissances. Tendances actuelles et nouveaux défis*. Editions Eyrolles/France Telecom, Paris.

CHARLET J., (2001) Ingénierie des connaissances : un domaine scientifique, un enseignement ? *Actes des 5° Journées d'Ingénierie des Connaissances IC 2001*, Grenoble (F) : PUG. 233-252.

CHATEAURAYNAUD F. (2003) Prospero - une technologie littéraire pour les Sciences Humaines. Cnrs Editions, Paris, 406 pages.

EUZENAT J. (Ed.), Research Challenges and perspectives of the sémantic Web. Report of the EU-NFS strategic Workshop, Sophia-Antipolis, October 2001.

HABERT B., FOLCH H. & ILLOUZ G. (2002). Sortir des sens uniques : repérer les mots mouvants dans le domaine social. *Sémiotiques*.

J. MOTHE, C. CHRISMENT, D. DKAKI, B. DOUSSET, D. EGRET. (2001) Information mining: use of the document dimensions to analyse interactively a document, *European Colloquium on IR Research: ECIR 2001*. p. 66-77.

ROSCH E., LLOYD B.B. (1978) Cognition and categorization. Lawrence Erlbaum, Hillsdale, New Jersey.

SPERBER D.,WILSON D. (1989) *La pertinence*. Les éditions de minuit, Paris.

TEULIER R., CHARLET J., TCHOUNIKINE P. (2003) *Ingénierie des connaissances*. A paraître.