Elisabeth Kolmayer
ENSSIB, CERSI

BASES DE DONNÉES GRAND PUBLIC ET ORGANISATION DES CONNAISSANCES

LES SYSTÈMES d'interrogation se développent, touchant des domaines très divers et des publics non spécialisés : catalogues de bibliothèques avec les OPAC[1], répertoire des professions sur minitel (avec le 11), offres d'emploi de l'ANPE[2], etc.
L'ouverture au grand public a rendu plus aiguës les difficultés d'interrogation, conduisant à l'évitement, parfois même au refus de l'outil : 15 % des possesseurs de minitel interrogent le 11 par rubriques, rares sont les usagers d'une bibliothèque de quartier qui se risquent à une interrogation par sujets, et la comparaison d'interrogations à l'aide d'un catalogue-papier et d'un catalogue informatisé n'est pas à l'avantage de ce dernier, ni en termes d'efficacité, ni en termes de satisfaction (5). L'observation des conduites, au terminal et dans les rayons, montre qu'un OPAC ne répond pas réellement aux besoins des usagers d'une bibliothèque, et les réponses à des questionnaires passés avant et après l'interrogation confirment cette indication (7, 8).

L'organisation des connaissances sur le domaine interrogé

Les systèmes d'interrogation destinés à des spécialistes d'un domaine avaient mis l'accent sur les difficultés liées aux langages d'interrogation et aux stratégies de recherche. L'utilisation par des non-spécialistes ajoute une nouvelle cause de difficultés : l'organisation des connaissances sur le domaine interrogé.

Lorsqu'un spécialiste interroge une base consacrée à son domaine, tous deux ont en commun une représentation de ce domaine : son vocabulaire, ses concepts et les relations entre ceux-ci. Elaborée par la communauté scientifique, cette représentation est transmise à ses membres et reflétée par la base (quelle que soit la façon dont, informatiquement, ces relations sont mises en œuvre).
En revanche, lorsqu'un utilisateur grand public interroge une base, il ne connaît pas forcément le vocabulaire propre au domaine, l'organisation de celui-ci, ni, bien sûr, l'organisation que la base lui impose :
– un usager de la Bibliothèque publique d'information doit « voir » Mai 68 comme un événement politique concernant la France en 1968 ; mais s'il s'agissait, pour lui, d'un phénomène sociologique plutôt que politique ? (10)
– un utilisateur du 11, qui a besoin d'une lavallière pour un mariage ne

* ENSSIB : Ecole nationale supérieure des sciences de l'information et des bibliothèques ; CERSI : Centre d'études et de recherche en sciences de l'information.
1. OPAC : Online public access catalog.
2. ANPE : Agence nationale pour l'emploi.

doit pas considérer qu'il s'agit d'un accessoire, sous peine de voir apparaître à l'écran la liste des revendeurs d'accessoires-automobile (13).

Alors qu'il y avait consensus sur l'organisation du domaine interrogé, avec les bases destinées à des spécialistes, il risque fort d'y avoir incompréhension mutuelle avec les bases grand public ; cette incompréhension va être renforcée par le fait que les modes d'organisation des connaissances dont disposent les bases et ceux de l'opérateur humain ne sont pas forcément les mêmes ; un petit détour par la psychologie cognitive va nous montrer que ceux-ci sont divers et pas toujours logiques.

Des façons d'organiser les connaissances

La nécessité d'organiser ses connaissances est imposée dès l'enfance par la complexité de l'environnement. La psychologie génétique avait, avec Piaget et son école, étudié la capacité de catégorisation en classes logiques. Plus récemment, d'autres modes d'organisation, moins logiques mais plus liés à l'environnement, ont été mis en évidence.

Catégorisation en classes logiques hiérarchisées

Dans ce mode d'organisation, une catégorie est définie par un ensemble de propriétés nécessaires et suffisantes. L'appartenance d'un élément à la classe est donc toujours décidable et tous les éléments d'une classe le sont au même titre.

Il existe plusieurs niveaux de classification. Les classes se situent les unes par rapport aux autres selon une relation d'emboîtement sans qu'aucun niveau ne soit privilégié. Le niveau le plus élevé est défini par un petit nombre de propriétés mais possède une grande extension. Au contraire, aux plus bas niveaux, les classes sont définies de façon beaucoup plus riche mais leur extension est limitée. A chaque niveau, les éléments d'une classe héritent des propriétés de la classe « emboîtante ».

Ce type de catégorisation est lent à acquérir. Selon Piaget et Inhelder, le signe de sa maîtrise est la quantification de l'inclusion : tous les éléments de la sous classe A1 sont quelques éléments de la classe inclusive A. Des études plus récentes (2 et 3) se montrent plus exigentes sur le critère de maîtrise et situent vers 11-12 ans l'acquisition de l'inclusion.

Des catégorisations écologiques

A ce premier mode d'organisation, fondé sur une « *capture logique des propriétés des objets* » (3), viennent s'adjoindre ou s'opposer d'autres modes, génétiquement antérieurs à la catégorisation en classes logiques, mais qui subsistent chez l'adulte. On en distingue deux types, selon les connaissances qu'ils mettent en forme : les catégories conceptuelles à base de typicalité ; les schémas situationnels et événementiels.

Les catégories conceptuelles à base de typicalité

Dites aussi « catégories naturelles », elles ont été mises en évidence par Rosch (15). Elles présentent deux différences avec les classes logiques, l'une dans la définition des catégories, l'autre dans les niveaux de catégorisation.

En effet, l'appartenance d'un élément à une catégorie n'est pas déterminée par un ensemble de conditions nécessaires et suffisantes mais par des ressemblances, un « air de famille » avec un centre appelé prototype[3]. Celui-ci est l'élément qui possède le plus de traits caractéristiques de la catégorie. Les différents membres de celle-ci se distribuent autour de lui selon un gradient de typicalité, du plus typique vers l'atypique. Si prototype et atypique appartiennent tous deux à la catégorie, ils n'en sont pas des représentants au même titre : ainsi, la pomme, prototype, est un meilleur représentant de la catégorie des fruits que l'olive, atypique.

L'autre différence concerne les niveaux de catégorisation : il existe un niveau privilégié, le niveau de base, compromis entre l'information apportée et la charge en mémoire. Les catégories du niveau de base sont celles du niveau le plus abstrait pour lesquelles les éléments ont encore de nombreux traits communs, en particulier des traits perceptifs, et suscitent des comportements semblables. Il en est ainsi de la catégorie chien, par opposition à la catégorie plus abstraite et moins informative des animaux, comme à celles des bassets, épagneuls, etc., qui apportent peu d'information supplémentaire, et au prix d'une charge cognitive plus lourde. Il semble que, génétiquement, les enfants catégorisent d'abord au niveau de base.

La plupart des travaux menés autour de la typicalité ont porté sur des objets naturels ou fabriqués comme les oiseaux, les fruits, les meubles, les vêtements, etc. Certaines études semblent indiquer que la typicalité concerne aussi des objets plus abstraits comme des algorithmes (1), voire même, avec certaines réserves, des objets sociaux comme le travail, les rôles professionnels (9).

Les schémas situationnels et événementiels

Un schéma situationnel est un objet complexe comme une gare, un parking. Un schéma événementiel est un ensemble stéréotypé d'événements, d'actions, de rôles, comme aller chez

Lorsqu'un utilisateur grand public interroge une base, il ne connaît pas forcément le vocabulaire propre au domaine

3. La notion de prototype recouvre plusieurs définitions dont on trouvera une présentation et une discussion dans Dubois (6).

le médecin, prendre son petit déjeuner, etc. On peut dire que les premiers sont analogues aux *« frames »* de Minsky, tandis que les seconds s'apparentent aux *« scripts »*[4] de Schank.

Dans ce type d'organisation, il n'y a pas appartenance inclusive, mais seulement partitive de l'élément à l'ensemble. Par ailleurs, les relations entre éléments sont fondées non sur la possession de propriétés communes (comme dans les classes logiques) ni sur des ressemblances (comme dans les catégories à base de typicalité), mais sur des contiguïtés spatiales, temporelles ou fonctionnelles : ainsi, dans le schéma « la gare », les relations entre les rails, les locomotives, les voyageurs ; dans le script « aller chez le médecin », entre la personne qui vient vous ouvrir, la salle d'attente et la feuille de sécurité sociale.

Ces deux modes d'organisation – catégories à typicalité et schémas – sont fondés sur des régularités de l'environnement. Ceci est évident pour les schémas : c'est à partir d'un objet réel ou d'une situation existante que ceux-ci se constituent alors que tous les animaux que l'on catégorise comme chiens ne se trouveront jamais rassemblés au même endroit. Quoique moins apparent, ce rôle de l'environnement existe aussi dans la catégorisation naturelle : Rosch insiste sur le fait que le monde environnant est structuré et que les divers attributs des objets qui nous entourent ne sont pas aléatoirement distribués : par exemple, les animaux à plumes ont toutes les chances d'avoir aussi des ailes, et ceux à fourrure, en général possèdent quatre pattes. L'individu a intérêt à tenir compte, dans sa façon d'organiser ses connaissances, de ces corrélats d'attributs. Ainsi les catégories naturelles reflètent-elles les regroupements et les discontinuités de l'environnement. C'est le rôle de l'environnement qui a fait donner à ces modes d'organisation le qualificatif d'écologique.

Si les chercheurs s'interrogent sur les relations génétiques entre catégorisations logiques et écologiques et, à l'intérieur de celles-ci, entre scripts et catégories naturelles (11), tous s'accordent pour considérer que les trois modes existent chez l'adulte.

L'interrogation

Puisque nous possédons plusieurs modes d'organisation de nos connaissances, comment ceux-ci interviennent-ils dans le déroulement d'une interrogation de base de données ? Pour aborder cette question, on a cher-

4. Ces structures ont été utilisées en intelligence artificielle pour la reconnaissance de formes et la compréhension de récits. Ce sont des cadres conceptuels préétablis, auxquels sont comparés les éléments de la figure particulière que l'on cherche à reconnaître ou de l'histoire que l'on veut interpréter.

Les questions

Catégories à prototypes

a. Une classe de CM2 doit aller visiter une coopérative de fruits dans la région ; les enfants aimeraient bien trouver un reportage sur les fruits, avant leur visite.

id. avec : pommes / noix

b. Avec « Les ateliers du mercredi », les 8-10 ans vont aller visiter un centre d'élevage pour animaux près de Roanne ; ils recherchent des documentaires sur les animaux.

id. avec : chiens-Roanne / dauphins-Cap d'Agde

c. Les grands de la maternelle étudient l'arbre ; ils sont allés en voir, ils en ont rapporté et planté des tout petits dans la classe ; la maîtresse voudrait une cassette qui présente la vie de l'arbre.

id. avec : chêne

d. En prévision des mercredis pluvieux, l'Animation-Jeunes recherche des cassettes sur les sports de ballon, pour les 7-10 ans qui sont déjà des mordus.

id. avec : foot

Catégories sans prototype

e. Une classe de CM1 fait un travail sur les oiseaux avant un départ en classe verte ; les enfants sont venus demander si on avait des cassettes sur les oiseaux à leur prêter.

id. avec : cigognes / pingouins

f. La Maison des Loisirs et de la Culture recherche une vidéo sur les véhicules de maintenant et d'autrefois.

id. avec : voitures

g. Un groupe d'étrangers vient d'arriver pour un stage dans le quartier ; ils viennent de la campagne et n'ont pas l'habitude d'une grande ville ; on voudrait une cassette qui leur montre les principaux bâtiments d'une ville.

id. avec : les mairies, les hôtels de ville, ce genre de choses que l'on trouve dans une ville.

h. Le centre social de Lyon-Champvert demande si on peut lui prêter une vidéo sur les logements à Lyon 9e.

id. avec : les appartements en location

Abstrait / concret

i. Le groupe « Habitat » du quartier voudrait quelque chose sur le problème de la pollution dans les grandes villes.

id. avec : le problème des déchets industriels

j. L'association « Retravailler » recherche des émissions sur l'emploi des femmes.

id. avec : sur les petits boulots des femmes

k. A l'Atelier pédagogique personnalisé, plusieurs stagiaires préparent des formations paramédicales. On cherche une vidéo qui présente ce qu'est le travail paramédical.

id. avec : formations d'aide-soignante

l. Un groupe d'élèves de 3° du collège voudrait une vidéo sur les diverses sources d'énergie que l'on sait utiliser actuellement.

id. sur le gaz naturel, comment on l'utilise

m. Des jeunes du quartier aimeraient trouver une vidéo sur le rock.

id. avec : le groupe INXS

Scripts

n) L'association « Lyon-Accueil » recherche pour un groupe de migrants une cassette sur le thème « Aller chez le médecin ».

id. avec : une cassette qui montre ce que c'est que se faire examiner quand on est malade.

o) Dans un stage d'apprentissage du français, on recherche une vidéo sur le thème "Aller au restaurant".

id. avec : comment commander son menu quand on va manger quelque part.

p) Pour un groupe de femmes étrangères, on recherche une cassette qui montre ce qui se passe quand on va à un mariage en France.

q) Pour les petits du cours préparatoire, on recherche une cassette qui montre comment on fait sa toilette le matin.

ché à observer comment variaient les descripteurs proposés, lors d'une recherche, selon que le domaine interrogé faisait appel à des scripts, à des catégories où la typicalité a été mise en évidence, ou à d'autres où elle ne s'exprime pas[5].

Expérimentation

Avant de décrire la situation expérimentale, ses variables et ses hypothèses, on présentera la population étudiée et le contexte d'expérimentation choisi.

Il s'agit d'une population de bas niveau de qualification : demandeurs d'emploi en cours de stage de formation. Cette population est peu étudiée, beaucoup moins que celle, traditionnelle, des étudiants en psychologie. Elle présente cependant un « effet de loupe » utile lorsqu'on se préoccupe des difficultés cognitives que suscitent ou révèlent les nouvelles technologies. Le contexte d'expérimentation est la recherche de cassettes dans un fichier de vidéothèque (uniquement des documentaires), activité peut-être perçue comme moins aride que la recherche de livres, et qui permet cependant d'aborder des types de connaissances très variés, en particulier des scripts.

On demande à deux (parfois trois) groupes de vingt personnes de trouver des vidéos sur seize sujets. Chaque sujet est abordé à deux niveaux de généralité : par exemple, si un groupe recherche des cassettes sur le travail paramédical, dans l'autre groupe, la question portera sur le travail de l'aide-soignante. Parmi ces seize sujets, quatre font appel à des scripts : comme aller au restaurant / commander son menu. Quatre autres portent sur des domaines où la typicalité s'exprime fortement : les questions portent alors sur la catégorie, son prototype, un atypique (par exemple : fruit / pomme-noix). Un troisième groupe de questions concerne des catégories où il n'existe pas de prototype reconnu : par exemple, logement / appartement. Enfin, dans un quatrième groupe, on fait varier le niveau d'abstraction : l'objet de la question est abstrait au niveau plus générique, concret au niveau plus spécifique ; par exemple : énergie / gaz naturel.

Pour chaque question, les participants doivent proposer quatre descripteurs avec, à chaque fois, la consigne : *« Et si celui-là ne "marche" pas, que pourrait-on essayer d'autre ? »*[6]. Les descripteurs sont comparés au mot inducteur contenu dans la question et classés en termes plus génériques (G), équivalents (E), plus spécifiques (S) ou termes associés (TA)[7]. Il est bien évident que ces termes recueillis sont des mots et non les représentations conceptuelles que l'individu possède en mémoire. On peut cependant penser qu'ils présentent une certaine corrélation avec ces représentations conceptuelles : utiliser des descripteurs génériques ou spécifiques est lié à l'organisation hiérarchique des représentations en mémoire alors que les termes associés renvoient plutôt aux propriétés de celles-ci.

On s'attendait à ce que les scripts suscitent plus de termes associés que les autres questions. Parmi celles-ci, on a fait l'hypothèse que la typicalité favorisait la catégorisation, les catégories où la typicalité s'exprime fortement ayant plus de réponses inclusives que celles où elle s'exprime peu et, d'autre part, les prototypes suscitant plus de réponses génériques que les atypiques.

Les résultats observés ne confirment pas entièrement ces hypothèses.

Résultats

Les questions sur les scripts suscitent plus de termes associés que les autres. Ceci est cohérent avec l'organisation du script, basée, non sur l'inclusion mais sur la contiguïté – spatiale, temporelle ou fonctionnelle de ses éléments.

Les catégories à prototypes n'ont pas plus de réponses hiérarchiques que les autres. Les différences observées (scripts exclus) dans les nombres de réponses G + E + S ne sont pas significatives.

La typicalité n'aide pas à « monter ». Les prototypes ou les éléments les

5. Cette étude a été subventionnée par le ministère de la Recherche et de la Technologie dans le cadre du programme *Hommes, travail, technologie.*

6. Il s'agit d'une simulation d'interrogation pour neutraliser la variable parasite que serait la réponse de la base.

7. On a utilisé pour cela le thésaurus MOTBIS du Centre national de la recherche pédagogique.

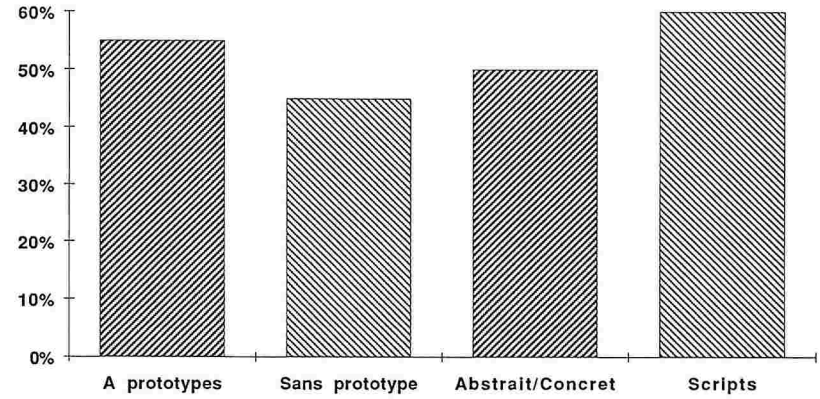
plus typiques ne suscitent pas plus de descripteurs génériques que les atypiques. Par exemple, une question sur les chiens ou les pommes ne suscite pas plus de descripteurs de type G qu'une question sur les dauphins ou les noix. Cependant la typicalité montre sa présence : prototypes et atypiques ne suscitent pas les mêmes termes catégoriels. Le niveau de base est important : l'existence d'un niveau privilégié de catégorisation est le second trait caractéristique des organisations à base de typicalité. Il semble jouer plus fortement que le fait d'être ou non prototype.

Ainsi, lorsqu'une question est située au-dessous du niveau de base, le nombre de descripteurs génériques augmente : chêne fait massivement passer à arbre alors que pomme n'appelle que moyennement fruit. Pourtant tous deux sont prototypes de leur catégorie. INXS fait encore plus massivement passer à rock, beaucoup plus que foot ne fait passer à sport dont il est pourtant le prototype.

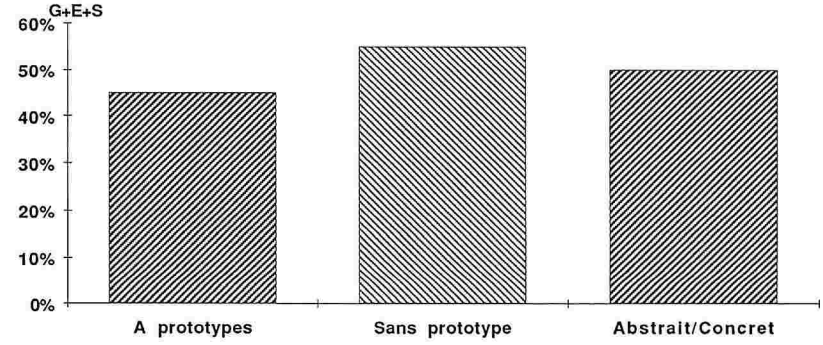
Il semble que la notion de niveau de base ait été moins étudiée que celle de prototype. Le rôle qu'elle joue dans l'interrogation fait espérer qu'elle retiendra l'attention des chercheurs en psychologie[8].

A côté de ces résultats, un autre s'est imposé, non prévu, et qui laisse perplexe : c'est la faiblesse de la relation générique-spécifique. On demandait aux utilisateurs quatre descripteurs pour chaque question : le mot inducteur contenu dans la question, puis trois autres de leur choix. On pensait que les réponses E + G + S (le mot inducteur, ses synonymes, génériques et spécifiques) représenteraient environ 75 % des réponses ; or on atteint à peine 50 %. Plus curieux encore, ce pourcentage baisse si on ne prend en compte que le premier descripteur proposé par les utilisateurs (à l'exception, bien évidemment du mot inducteur). Ceci élimine le rôle possible des différences de disponibilité lexicale des termes associés et des génériques-spécifiques.

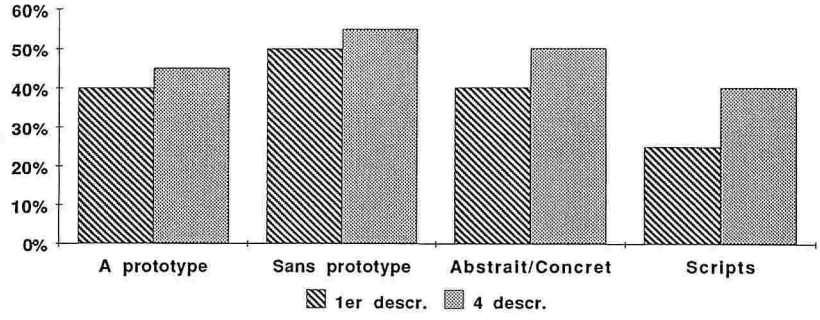
Graphique 1
Pourcentage de descripteurs de type Terme Associé selon le type de connaissances mises en jeu par les questions



Graphique 2
Pourcentage de descripteurs sur la relation hiérarchique selon le type de catégorie conceptuelle



Graphique 3
Pourcentage de descripteurs sur la relation hiérarchique*



* - 1er descripteur différent du mot inducteur
- ensemble des 4 descripteurs proposés.

8. Ceci semble en train de se réaliser ; on se référera à l'ouvrage de Cordier (à paraître).

La faiblesse d'utilisation de la relation hiérarchique risque de rendre très difficiles les interrogations lorsque la question que se pose l'utilisateur ne se situe pas d'emblée au niveau de généralité / spécificité adapté à l'indexation des documents de la base.

Les résultats présentés ici montrent qu'interviennent, dans une interrogation de base de données, des modes d'organisation des connaissances « écologiques » comme les scripts ou la catégorisation avec effet de typicalité. Cette intervention se manifeste par le type de descripteurs que choisit un utilisateur pour indexer sa question. Alors que les scripts renforcent l'importance des termes associés, la typicalité agit, moins par les prototypes que par le niveau de base : l'utilisateur a tendance à « monter » pour indexer sa question au niveau de base lorsque celle-ci est posée à un niveau plus spécifique. Les résultats montrent également la difficulté des usagers à utiliser la relation hiérarchique. Certes, l'étude porte sur une population particulière (celle dite des bas niveaux de qualification) et n'aborde qu'un petit nombre de domaines interrogés. Ses résultats sont donc plus à considérer comme des questions ouvertes que comme des données sûres.

Mais les résultats montrent de façon nette à quel point il est difficile, pour un utilisateur, d'effectuer l'ajustement hiérarchique de sa question. Ils plaident ainsi pour une nouvelle répartition de cette fonction au sein de l'ensemble utilisateur-système d'interrogation. Celle-ci est entièrement à la charge de l'utilisateur. On peut envisager que le système d'interrogation l'assume ou qu'il fournisse des aides cognitives à ses usagers.

Les deux attitudes ne sont nullement antagonistes mais complémentaires. La première vise à comprendre la question de l'utilisateur et à la « faire comprendre » à la base. Ce rôle, en général celui du traitement du langage naturel, semble particulièrement nécessaire dans le cas d'interrogations effectuées par des publics faiblement qualifiés. Mais comprendre sa question et la traduire dans une forme canonique adaptée à la base restent transparents à l'utilisateur. Il n'en reçoit aucune aide pour se construire une représentation du domaine qu'il interroge. Or cette représentation, avec les relations logiques, à base d'inclusion, qui la sous-tendent, lui manque bien souvent. D'autres aides lui sont donc nécessaires : lesquelles, présentées sous quelles formes ? Répondre à ces questions nécessitera bien des recherches, faisant appel à de multiples compétences.

Mai 1992

BIBLIOGRAPHIE

1. Adelson, B., « Comparing natural and abstract categories : a case study of computer science », *Cognitive Science,* 1985, 9, p. 417-430.

2. Bideaud, Jacqueline, *Logique et bricolage chez l'enfant,* Presses universitaires de Lille, 1988, 434 p.

3. Bideaud Jacqueline, Houde Olivier, « Le développement des catégorisations : « capture » logique ou « capture » écologique des propriétés des objets ? », *L'Année psychologique,* 1989, p. 87-123.

4. Cordier, Françoise, (1992), *Les représentations cognitives privilégiées. Typicalité et niveau de base,* Presses universitaires de Lille (à paraître).

5. Dalrymple, Prudence W., « Retrieval by reformulation in two library catalogs : toward a cognitive model of searching behavior », *JASIS,* 1990, t. 41, n° 4, p. 272-281.

6. Dubois, Danièle, *Sémantique et cognition. Catégories, prototypes, typicalité,* Paris, CNRS, 1991, Sciences du langage, 342 p.

7. Hancock-Beaulieu, Micheline, « Evaluating the impact of an online library catalogue on subject searching behaviour at the catalogue and at the shelve », *Journal of Documentation,* 1990, t. 46, n° 4, p. 318-338.

8. Hancock-Beaulieu, Micheline, « Evaluations of online catalogues : eliciting information from the user », *Information Processing and Management,* 1991, t. 27, n° 5, p. 523-532.

9. Huteau, Michel, « Organisation catégorielle des objets sociaux : portée et limite des catégorisations de E. Rosch », *Sémantique et cognition : catégories, prototypes, typicalité,* ss la dir. de D. Dubois, Paris, Editions du CNRS, 1991, p. 71-88.

10. Le Marec, Joëlle, *Dialogue ou labyrinthe ? La consultation des catalogues informatisés par les usagers,* Paris, Centre Georges Pompidou BPI, 1989.

11. Nelson, Katherine, « Where do taxonomic categories come from ? », *Hum. Dev.,* 1988, n° 31, p. 3-10.

12. Piaget, Jean ; Inhelder, Bärbel, *De la logique de l'enfant à la logique de l'adolescent,* Paris, PUF, 1970, 314 p.

13. Riondet, Odile, *Interrogation de l'annuaire électronique par des populations de bas niveau de qualification : rapport d'expérimentation,* CLEFI, 1991, 85 p.

14. Rosch, Eleonor, « Classification d'objets du monde réel : origines et représentations dans la cognition », *in* Ehrlich, E Tulving (éd.), *Bulletin de Psychologie,* 1976, numéro spécial « La mémoire sémantique », p. 242-250.

15. Rosch, Eleonor, « Principles of categorization » in E. Rosch B.B. Lloyd (éd.), *Cognition and Categorization,* Hillsdale (N.J.), L. Erlbaum , 1978, p. 27-47.